Investigating Machine Learning Approaches for Alzheimer's Disease Detection

Maya Ravichandran

University of Oxford

September 2022

Abstract

Alzheimer's disease (AD) is a debilitating and ultimately fatal neurodegenerative disorder, with no known treatment or prevention method, and is widespread, with 1.6% of the US population estimated to have AD. AD is estimated to be 70% heritable, but the underlying genetic causes of AD are not fully known. To better understand the genetic basis of AD, this work aimed to create a machine learning model that can predict if an individual has AD based on their DNA sequence. This will help us identify the genes relevant for AD. Utilizing the Alzheimer's Disease Neuroimaging Initiative dataset (ADNI) (518 cases of AD or mild cognitive impairment, 276 controls), transformer and support vector machine (SVM) models were applied to whole genome sequencing data of various genes previously associated with AD. The transformer model was unable to find a signal on the ADNI dataset, but its prediction worked when an artificial signal was introduced to the DNA sequencing data. The SVM model was able to find a signal on many SNPs within the genes tested. These results suggest that SVM models are more effective when applied to small datasets with limited signal, and that transformers require a larger dataset or greater amount of signal to be effective. To our knowledge, this is the first study applying neural networks to unbroken stretches of DNA sequencing data to attempt the prediction of a phenotypic trait.

Contents

List of Figures

List of Tables

List of Abbreviations

AD . . . . . Alzheimer's disease

ADNI . . . . . Alzheimer's Disease Neuroimaging Initiative

APOE . . . . . Apolipoprotein E

BERT . . . . . Bidirectional Encoder Representations from Transformers

bp . . . . . base pair

CNN . . . . . convolutional neural network

GWAS . . . . . Genome-wide association study

MCC . . . . . Matthews correlation coefficient

MCI . . . . . mild cognitive impairment

NN . . . . . neural network

ROC AUC . . . Area Under the Receiver Operating Characteristic Curve

RNN . . . . . recurrent neural network

SNP . . . . . single nucleotide polymorphism

SVM . . . . . support vector machine

WES . . . . . whole exome sequencing

WGS . . . . . whole genome sequencing

1 Introduction

1.1 Rationale

Alzheimer's disease (AD) is a neurodegenerative disorder that affects memory and brain function. In its late stages, it can severely affect an afflicted person's ability to perform basic functions and can ultimately result in death [Breijyeh and Karaman, 2020]. The available treatments for Alzheimer's disease treat only symptoms, and do not cure the disease or slow its progression [Breijyeh and Karaman, 2020]. Moreover, Alzheimer's disease is widespread, with 1.6% of the US population in 2014 having Alzheimer's disease and related dementias [Matthews *et al.*, 2019]. This makes treatment of Alzheimer's disease a large unmet need.

Alzheimer's disease is highly heritable [Andrews *et al.*, 2020]. AD's estimated heritability of 70% makes it one of the human multifactorial diseases with the greatest heritability [Bellenguez *et al.*, 2020]. There are currently over 40 gene loci known to be associated with Alzheimer's disease [Novikova *et al.*, 2021]. However, we still do not understand the causal relationship and functional variants underlying the statistical relationship between these genes

and Alzheimer's disease [Novikova *et al.*, 2021]. A better understanding of the genetic basis of Alzheimer's disease could help us discover prevention and treatment methods and identify drug targets. It could also help identify people at risk for developing the disease, who may benefit from preventative measures and early treatment once these methods are discovered.

To better understand the genetic basis of Alzheimer's disease, it may be helpful to use powerful modeling approaches. One such approach is deep learning. Deep learning has achieved impressive results in areas such as image recognition (e.g. AlexNet [Krizhevsky *et al.*, 2012]), natural language (e.g GPT-3 [Brown *et al.*, 2020]), and biological problems such as protein folding (e.g. AlphaFold [Jumper *et al.*, 2021]). These successes are encouraging for the applications of deep learning in other domains, such as in biomedical problems.

Deep learning approaches have benefited from large amounts of labeled training data. The advent of high-throughput genomic sequencing, other omic techniques, and electronic health records have made the collection of large amounts of biomedical data possible. Various initiatives have emerged to collate this data and make it available for research. The UK Biobank plans to collect whole exome sequencing (WES) data from 500,000 participants, with 200,000 samples already amassed [Szustakowski *et al.*, 2021]. Other data collection initiatives exist as well, such as The Cancer Genome Atlas, which has collected WES data from 20,000 tumor and matched normal samples [Institute, n.d.]. Despite this massive proliferation of data, there is currently little research on applying neural networks directly to raw DNA sequence data. It may be worth exploring the potential of deep learning in this area, and the results of such studies could be applicable to a variety of human genetic diseases. Among the range of human genetic diseases, Alzheimer's disease is a strong choice for investigation using deep learning approaches due to its high heritability and large unmet medical need.

1.2 Objectives

This work aims to create a machine learning (ML) model that can predict if an individual has Alzheimer's disease based on their DNA sequence. Such a model can be used to better understand the genetic basis of AD by identifying genes and variants that are associated with AD. Further, this work aims to provide insight on the utility of deep learning and other machine learning methods when applied to genome sequencing data.

1.3 Contributions

The work of this dissertation has contributed the following:

- **SVM positive results:** The support vector machine (SVM) model was able to find a signal in prediction of Alzheimer's disease in participants on many of the single nucleotide polymorphisms (SNPs) within the genes examined from the Alzheimer's Disease Neuroimaging Initiative (ADNI) whole genome sequencing (WGS) dataset. Some of the SNPs have been described in the literature as being related to AD, while other SNPs' relation to AD have not been described.

- **Transformer negative results:** The transformer model, based on the DNABERT pre-trained model, was unable to find a signal in predicting AD on any of the 100 base pair-long windows of the *APOE* gene from the ADNI dataset or on any of the other three gene sequences examined.

- **Validation of transformer model on DNA sequence data:** The transformer model, based on the DNABERT pre-trained model, was able to predict the presence of AD with high accuracy when an artificial signal was introduced into the ADNI whole genome sequencing data.

- **Novelty of running NNs on DNA sequence data to predict phenotype:** There has been little research on applying neural networks to raw DNA sequence data in general. To my knowledge, this is the first study applying NNs to attempt to predict the phenotype (presence or absence of AD) of an individual based on their DNA sequence. Previous studies applying NNs to genome sequence data have attempted to predict information about the DNA itself, such as whether a region is a promoter, rather than predicting phenotype.

1.4 Dissertation Structure

Chapter 2 contains a literature review of previous research in the fields of Alzheimer's disease genetics and applying neural networks to genomic data. It discusses known genetic factors of AD, along with methods of discovering these genetic factors, including wet lab experiments, genome-wide association studies, and machine learning techniques. It also discusses a variety of prior studies that have applied NNs to sequencing data, whether on single nucleotide polymorphisms, amino acid sequences, or whole genome sequencing data.

Chapter 3 provides the background on the techniques used in this study. Explanations of support vector machines and transformers are provided, along with explanations of the metrics used to evaluate the performance of these models when used for binary classification. In addition, an explanation of the computational pipeline used to analyze next-generation whole genome sequencing data is provided. A sample DNA sequence file is also shown.

Chapter 4 explains the datasets used in this project and the preprocessing applied to the data, and describes the models and experiments conducted. It describes the ADNI dataset, results of exploratory data analysis on this dataset, the SVM and transformer models and architectures used, and the experiments run with these models.

Chapter 5 contains the results of all the experiments run in this study. Experiments were run using SVM and transformer models on SNP and DNA sequencing data. Data tables with the results of the prediction metrics, figures of the training curves, and narrative explanations of the results are provided.

Chapter 6 provides an interpretation of the results. It links the results of this study back to previous studies on AD genetics. It also attempts to explain trends in the results, including why certain models performed better than others and departures from previous results. It also compares the results of this model with previous results on the prediction of AD using genetic data and on applying NNs to genetic data. It explains challenges and limitations of the study, which notably include issues with the size of the dataset that may have adversely affected the prediction accuracy. It also describes directions for future study.

Chapter 7 discusses the conclusions of this study and broader implications of this work, connecting this study to the broader field of machine learning research and the potential impacts of applying machine learning to genomic data.

2 Related Work

2.1 Previous research on predicting Alzheimer's disease

2.1.1 Known genetic factors of AD

Alzheimer's disease is a strongly genetic disease, with 70% of AD cases being related to genetic factors [Breijyeh and Karaman, 2020]. Previous work has identified over 40 genes or loci related to AD [Bellenguez *et al.*, 2020]. Bellenguez *et al.*, 2020's paper contains a table of genes shown to be associated with AD and their corresponding odds ratios for the development of AD.

There are two types of Alzheimer's disease: late onset, in which in which the disease begins at age 65 or later, and early onset, in which the disease begins before age 65. The *APOE* gene, located on chromosome 19, is the gene that is the highest known risk factor for late onset AD [Potkin *et al.*, 2009; C.-C. Liu *et al.*, 2013]. The $\varepsilon 4$, $\varepsilon 3$, and $\varepsilon 2$ alleles of the APOE gene are associated with a higher, neutral, and lower risk of AD, respectively [Breijyeh and Karaman, 2020]. Having two copies of the deleterious $\varepsilon 4$ allele results in an odds ratio of 14.49 for developing AD [Bellenguez *et al.*, 2020]. The genes *APP*, *PSEN1*, and

PSEN2 are the genes with the greatest risk factor for early onset AD [Cacace *et al.*, 2016]. *APOE* is also associated with early onset AD [Cacace *et al.*, 2016].

Studies on AD genetics have also been conducted on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, which was used for this dissertation project. (More information on the ADNI dataset is in Section 4.1.1.) Potkin *et al.*, 2009 conducted a GWAS study on the ADNI dataset, using hippocampal atrophy determined through MRI scans as a measure of mild AD. They confirmed that *APOE* was a risk factor for AD and also identified the novel gene *TOMM40* (translocase of outer mitochondrial membrane 40), on chromosome 19, as a risk factor for AD, with a p-value of $\leq 10^{-6}$.

2.1.2 Methods for discovering genetic factors of AD

Lab-based and clinical approaches

Prior to the advent of next-generation sequencing technology, and to supplement gnomic analyses with validated laboratory experiments, a variety of clinical and lab-based approaches have been used to discover and validate genes associated with Alzheimer's disease. Several of the genes with the greatest implication in AD have been investigated in wet lab studies. For example, the study by Weyer *et al.*, 2011 investigated APP (amyloid precursor protein) knockout mice, which have had their *APP* gene deactivated. They found that these mice had diminished body weight, brain weight, and grip strength, and that aged mice had learning and memory issues. The *APOE* gene's role in AD has also been validated through lab studies. Namba *et al.*, 1991 showed that Apolipoprotein-E is deposited in the senile plaques located in the brains of AD patients.

Clinical studies have also demonstrated the association between genes and AD. For example, Smith *et al.*, 1998 evaluated patients with AD and MCI and normal control subjects on a series of cognitive tests, finding that patients

with AD and MCI who also had the *APOE* ε4 allele performed worse on cognitive tests than those without the allele, but this effect was not present in healthy patients without AD or MCI.

Genome-wide association studies

The advent of next-generation, high-throughput genome sequencing has enabled the proliferation of whole genome and whole exome sequencing datasets. These datasets can be used for genome-wide association studies (GWAS). GWAS search for genetic variants which have alleles that are more frequent in people with a certain phenotype, controlling for ancestry [Uffelmann *et al.*, 2021]. The genetic variant that is most frequently used are single nucleotide polymorphisms (SNPs), although copy-number variants or sequence variants can also be considered [Uffelmann *et al.*, 2021]. Since GWAS usually consider SNPs, they do not consider the effects of longer stretches of DNA sequences on the formation of a phenotypic trait. However, GWAS have the strength of being able to find associations across the genome rather than only focusing on a single locus at a time, which can be useful in complex, polygenic traits such as height, schizophrenia, and AD. To test for associations, GWAS typically use linear regressions or logistic regressions [Uffelmann *et al.*, 2021]. The individuals' inputted genotypes are usually found with microarrays, whole exome sequencing, or whole genome sequencing [Uffelmann *et al.*, 2021]. Many software tools exist for conducting GWAS, such as PLINK [Purcell *et al.*, 2007], which is widely used for finding associations. As of 2021, more than 5,700 GWAS have been carried out [Uffelmann *et al.*, 2021].

A study by Potkin *et al.*, 2009 conducted a GWAS of 381 participants from the ADNI dataset and found the *APOE* gene as a risk gene, a new risk gene, *TOMM40*, and identified five new candidate risk genes for AD.

Genome-wide association studies by proxy (GWAX) use parental history of a disease or a trait as a proxy for an individual possessing a trait. This helps

increase the sample size of the number of positive cases of a disease, which is especially important for AD, since its late onset means that younger participants may not be able to have their cases of AD confirmed, excluding them from the study if only confirmed individual cases of AD were considered. Using parental history as a proxy decreases the statistical power compared to using only confirmed individual cases, but can increase the statistical power overall due to the increase in sample size [Andrews *et al.*, 2020]. Marioni *et al.*, 2018 conducted a GWAX of the UK Biobank dataset, using data from 314,278 participants, and identified 27 susceptibility loci for AD, including three novel loci.

Other GWAS for AD have been conducted. Jansen *et al.*, 2019 performed a GWAS on 455,258 participants (71,880 cases, 383,378 controls) using data from a variety of AD consortia, using AD and proxy AD cases, resulting in the finding of 29 risk loci. Moreover, Kunkle *et al.*, 2019 performed a GWAS that included 94,437 individuals: 35,274 diagnosed cases of late-onset AD, and 59,163 controls, identifying 24 susceptibility loci.

As discussed, GWAS have been successful in identifying many loci associated with AD. However, there are limitations to GWAS. One major limitation is that GWAS identify many susceptibility loci, but do not distinguish between causal variants and variants that are not causal but are highly correlated to the causal variants [Nicholls *et al.*, 2020]. GWAS also do not provide information on the biological mechanisms underlying a gene's association with AD, and more work is needed to discover this [Andrews *et al.*, 2020]. In addition, GWAS struggle with determining the genetic basis of complex traits, such as those that depend on gene-gene or gene-environment interactions [Tam *et al.*, 2019]. Moreover, GWAS struggles to recognize epistasis, the phenomenon in which the expression of certain genetic variants depend on the alleles present in other genes [Tam *et al.*, 2019]. Given these limitations of GWAS, it may be useful to investigate alternative approaches to determining the genetic factors underlying diseases.

One such approach is machine learning (ML). According to Nicholls *et al.*, 2020, machine learning can be an important tool for prioritizing loci identified by GWAS. More studies using ML for prediction of AD and other genetic diseases are described below.

Machine learning

Several studies have used machine learning for classification of AD based on SNP data. The study by Ghose *et al.*, 2022 developed a novel approach called Genome wide association neural networks (GWANN), which they used to identify nonlinear and SNP-SNP interactions in the UK Biobank dataset of family history of AD. They were able to identify previously known AD genes, target nominations, and novel genes. They used SNPs and other covariates as input and a convolutional neural network-based architecture for their model.

Another study by Venugopalan *et al.*, 2021 utilized stacked denoising autoencoders and 3D-convolutional neural networks, to predict the stage of AD based on magnetic resonance imaging, SNP, and clinical test data. They found that the deep learning approaches outperformed support vector machine, decision tree, and k-nearest neighbor techniques [Venugopalan *et al.*, 2021].

The study by Jo *et al.*, 2022 used convolutional neural network (CNN), random forest, and XGBoost approaches on windows of multiple SNPs from the ADNI dataset, finding that the results were comparable between the models. They also confirmed that the *APOE* gene was the most significant within their dataset.

Other research studies have used machine learning for predictive tasks related to AD, but that do not directly use genome sequencing data. A study by Huang *et al.*, 2018 utilized a support vector machine to classify AD candidate genes using gene expression data and gene network data that was specific to the human brain. In addition, a machine learning model, using a logistic regression, was used to predict candidate drugs for repurposing to treat AD

using data on lists of gene names, the levels of gene expression associated with these genes, and the stage of AD as identified by the Braak staging method [Rodriguez *et al.*, 2021].

2.2 Previous research on applying NNs to genome sequencing data

While there is much research on the genetic basis of AD, and many of these studies use computational or machine learning techniques, to my knowledge, there have been no studies that have worked with windows greater than 1 base pair (bp) long of raw whole genome sequencing data for prediction of AD. Studies that use multiple base pairs at a time, such as those by Ghose *et al.*, 2022 and Jo *et al.*, 2022, use multiple SNPs at a time, rather than unaltered strings of raw DNA sequences. While there are no studies operating on strings of DNA sequences specific to Alzheimer's disease, there have been studies that applied machine learning and deep learning to raw whole genome sequencing data to predict other characteristics. Some of these studies are described below.

BERT-based models

The DNABERT model, created by Ji *et al.*, 2021, applies deep learning to raw WGS data. The DNABERT paper applies principles used for training natural language processing models such as BERT (Bidirectional Encoder Representations from Transformers) [Devlin *et al.*, 2018]. Previous findings have suggested that DNA shares similarities to natural language (for example, Brendel and Busse, 1984), motivating the use of language models for making predictions about DNA.

The BERT model has been a highly successful language model, and it achieved state-of-the-art performance on many natural language processing tasks and benchmarks at the time of its publication in 2018 [Devlin *et al.*, 2018].

One of its key contributions was the prediction of values in a bidirectional manner, rather than in a traditional left-to-right or right-to-left manner used by other models. To train the BERT model in a bidirectional manner, the authors used a training procedure of masking certain words (tokens) from the inputted text and training the BERT model to predict these masked words. This resulted in a pre-trained model that was later able to be fine-tuned on a variety of tasks, achieving high performance in many of them.

The DNABERT model applies many of the principles used in the BERT natural language model to DNA. It is also a bidirectional encoder representation transformer, and has a similar pre-training procedure to BERT, as it was pre-trained on the task of completing missing pieces of DNA sequence that were removed from the training data of actual DNA sequences. Unlike other SNP-based models, the DNABERT model actually operated directly on unbroken stretches of whole genome sequencing data. The inputted DNA sequences were split into k-mers with k between 3 and 6. These k-mer tokens contained overlapping stretches of DNA sequence to give the model more information on the local context surrounding each nucleotide [Ji *et al.*, 2021]. Their model took 25 days to pre-train on 8 NVIDIA 20280Ti GPUs, and the authors have released the pre-trained model on GitHub. The DNABERT pre-trained model can then be fine-tuned on a variety of tasks, similarly to BERT. The DNABERT authors fine-tuned their pre-trained model on the tasks of predicting proximal and core promoter regions, transcription factor binding sites, and splice sites, achieving state-of-the-art performance on these tasks [Ji *et al.*, 2021]. The authors hope that the DNABERT model is generalizable, stating that they 'anticipate that the pre-trained DNABERT model can be [fine] tuned to many other sequence analyses tasks' [Ji *et al.*, 2021].

Other models

One study that has used deep learning on raw sequencing data is that of Sundaram *et al.*, 2018. Rather than using DNA sequencing data as input, they used amino acid sequences. The inputs of their deep residual convolutional neural network model were the amino acid sequences on either side of the variant being investigated, along with the orthologous sequences from other primate species [Sundaram *et al.*, 2018]. The output they tried to predict was the pathogenicity of the variant in question, and they trained the model using a dataset of variants labeled as either pathogenic or benign by humans [Sundaram *et al.*, 2018]. They achieved 91% accuracy on predicting variants as benign, an improvement over previous methods. This study is an interesting in that it predicts a phenotypic trait of clinical consequence, namely the pathogenicity of a variant. This study also takes the interesting approach of bringing an evolutionary biology perspective to human disease, as it compares variants in humans to variants in primates, including chimpanzees, bonobos, and macaques, to determine if a variant in humans is pathogenic, asserting that a missense variant present in another primate species is usually benign in humans [Sundaram *et al.*, 2018]. The article tends to focus on whether variants are benign from an evolutionary standpoint, rather than if they are directly implicated in a human disease such as AD.

The SpliceAI model, proposed by Jaganathan *et al.*, 2019, used a ResNet CNN-based architecture to predict if a pre-mRNA transcript was a splice site, based on a training dataset of pre-mRNA transcripts and labels of whether a transcript is a splice site. They achieved 95% accuracy as defined by their top-k metric [Jaganathan *et al.*, 2019]. While this study did not use DNA sequencing data, it did use strings of RNA sequencing data, and this data was preprocessed using one-hot encoding.

Avsec *et al.*, 2021 utilized a transformer and convolution-based model, called the Enformer model, for the prediction of gene expression. They took in very large stretches of DNA as their input, taking in 196,608 bp at a time. They used a dataset consisting of 38,000 DNA sequences [Avsec *et al.*, 2021]. They achieved a mean correlation of 0.85, while the experimental accuracy is estimated to be 0.94 [Avsec *et al.*, 2021]. Gene expression is also a phenotypic rather than genotypic trait.

NNs have also been applied to the genomes of non-human species. Ranawana and Palade, 2005 used neural networks to identify promoter regions within DNA sequence strings from *Escherichia coli*. This work was published in 2005, indicating that the application of neural networks to DNA sequences has been a topic of study for many years.

3 Technical Background

3.1 Support Vector Machines

Support vector machines (SVMs) are models used for supervised machine learning problems. The idea behind support vector machines is illustrated in Figure 3.1. An SVM aims to find a hyperplane to separate two classes, such that all the data points of each class fall on one side of the hyperplane [Bishop, 2006]. The SVM attempts to choose a hyperplane that falls between the two data classes. The SVM aims to maximize the margin, which is the distance between the hyperplane and the closest data point [Bishop, 2006]. The closest data points in each of the two classes are called the support vectors.

The SVM maximizes the margin by writing the problem as a constrained optimization problem. Lagrange multipliers can then be applied to the problem to frame it as a quadratic optimization problem, which can be solved with computational techniques [Bishop, 2006].

In order to use an SVM on non-linearly separable data, we can use kernels. This involves replacing dot products with a non-linear kernel function, which results in the data being projected to a higher-dimensional feature space.

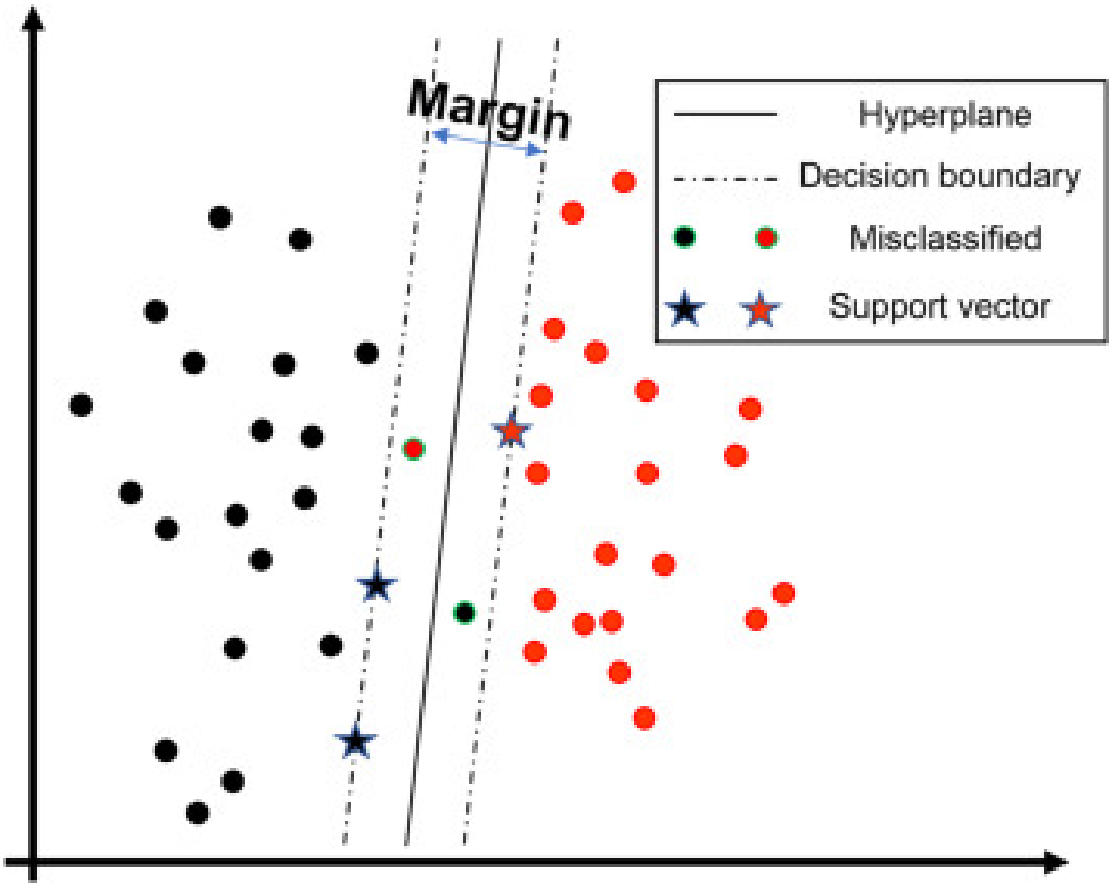

Figure 3.1: Illustration of an SVM hyperplane, support vectors, margin, and classification of data points on either side of the hyperplane. Figure from Misra *et al.*, 2020.

In this higher-dimensional space, it may be possible to separate the data with a hyperplane.

Some commonly-used kernels, that were used in this project, are as follows. The formulas are from Bishop, 2006 and Scikit-learn Developers, n.d.[a].

1. Linear kernel: $k(\mathbf{x}, \mathbf{x}\prime) = \mathbf{x}^T\mathbf{x}\prime$ (also known as the identity kernel)

2. Polynomial kernel: $k(\mathbf{x}, \mathbf{x}\prime) = (\mathbf{x}^T\mathbf{x}\prime + c)^M$, where c is a constant > 0 and M is the desired degree

3. Radial basis function (RBF) kernel: $k(\mathbf{x}, \mathbf{x}\prime) = \exp(-\gamma||\mathbf{x} - \mathbf{x}\prime||^2)$, where γ is a parameter > 0

SVMs are popular classifiers that can be used effectively in a wide variety of cases, including in high-dimensional datasets [Scikit-learn Developers, n.d.(a)]. They may struggle with overfitting in cases in which there are many more features than samples in the dataset [Scikit-learn Developers, n.d.(a)].

3.2 Transformers

Transfomrmers were first described by Vaswani *et al.*, 2017. Transformers were initially built for sequence transduction tasks, e.g. machine translation, as an innovation over the recurrent neural networks (RNNs) and convolutional neural networks commonly used at the time [Vaswani *et al.*, 2017]. Instead of using convolutional or recurrent layers, they use an attention mechanism.

The attention function involves 'mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors' [Vaswani *et al.*, 2017].

The attention computation returns a weighted sum, and is defined as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q is the query matrix, K is the keys matrix, V is the values matrix, and d_k is the dimension of the keys (formula from Vaswani *et al.*, 2017). $\sqrt{d_k}$ serves as a scaling factor.

They perform a multi-head attention mechanism, which involves projecting the keys, values, and queries into different subspaces, running the attention function on each of these, concatenating the results, and projecting the results again [Vaswani *et al.*, 2017]. They utilize multi-head attention for self-attention, which allows each encoder position to use attention on every position of the previous encoder layer, and for each decoder position to use attention on itself and every position in the decoder previous to that position [Vaswani *et al.*, 2017].

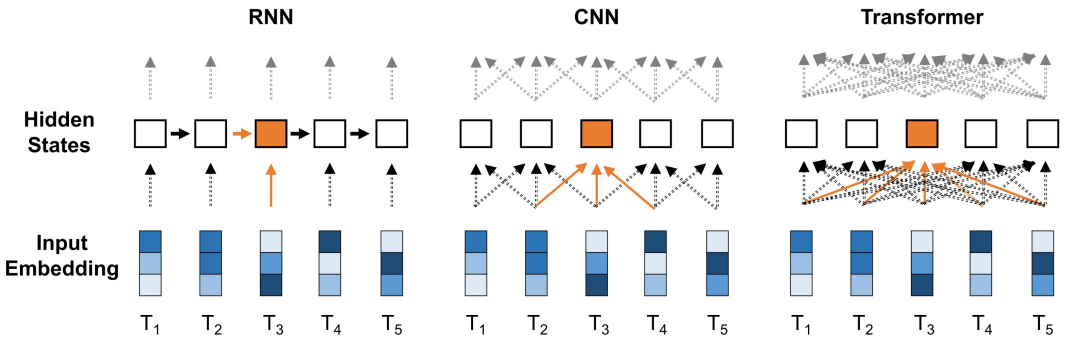

Figure 3.2: Illustration of the various approaches to local and global dependencies taken by RNNs, CNNs, and transformers, from Ji *et al.*, 2021. RNNs consider the data sequentially, CNNs consider the data along with local context, and transformers consider the data in a global context, taking long-range dependencies into account.

One of the key advantages of transformer models over RNNs and CNNs is that its self-attention mechanism allows for a shorter 'path length between long-range dependencies'. Transformers have a lower big-O upper bound for maximum path length compared to RNNs and CNNs [Vaswani *et al.*, 2017]. This makes long-range dependencies easier to learn [Vaswani *et al.*, 2017]. Figure 3.2 provides an illustration of the various approaches to local and global dependencies taken by RNNs, CNNs, and transformers. RNNs consider the data sequentially, CNNs consider the data along with local context, and transformers consider the data in a global context, taking long-range dependencies into account.

The architecture of the original transformer model is shown in Figure 3.3.

The transformer model achieved the state of the art on machine translation benchmarks when it was released, and it has since achieved great success in natural language processing in models such as BERT and GPT-3 [Devlin *et al.*, 2018; Brown *et al.*, 2020]. It has even been applied to DNA sequences in models such as DNABERT and the Enformer, described in Section 2.2.

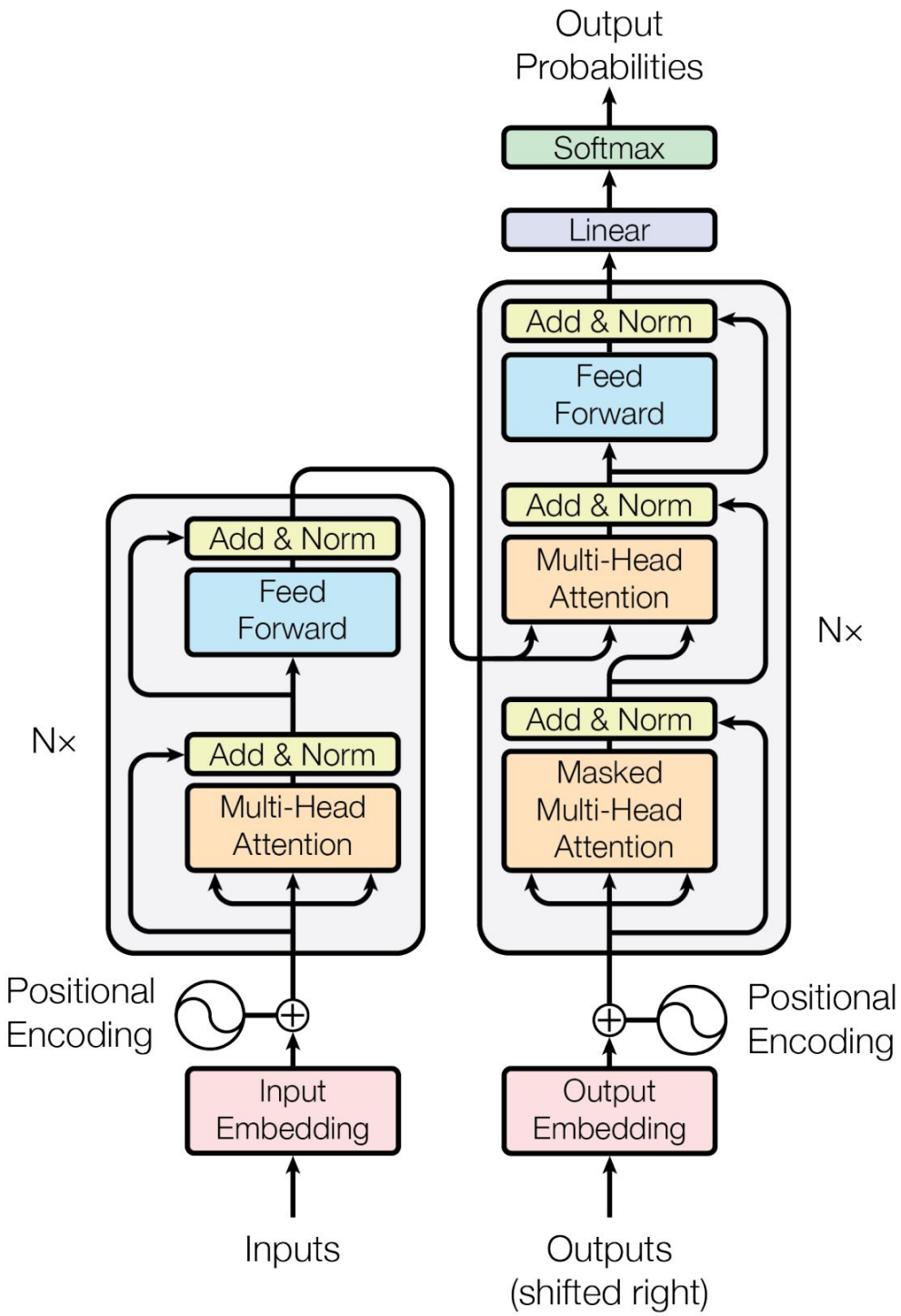

Figure 3.3: Illustration of the Transformer model architecture, from Vaswani *et al.*, 2017.

3.3 Evaluation of binary classifiers

For datasets with an imbalance between the number of data points in each of the two classes, evaluating a binary classifier's performance becomes more difficult, as one must pay special attention to the classifier's ability to distinguish between members of the two classes. For example, suppose we have a dataset for an uncommon disease in which 98% of the samples in the dataset do not have the disease, and only 2% of the samples do. A classifier that always predicted the mode of the dataset as its output, in this case always predicting that an individual did not have the disease, would achieve a 98% accuracy on the dateset overall, but a 0% accuracy on patients with the disease, and would fail to inform patients and their healthcare providers that they had a disease. This could have dangerous consequences if a disease that requires early treatment, such as cancer, is not caught. As such, we use a suite of metrics to evaluate a binary classifier's performance, taking the possibility of imbalanced datasets into account. These metrics are defined below.

In the metrics below, y denote the true values, and $\hat{y}$ denote the predicted values.

Accuracy

Accuracy is the simplest metric with which to interpret a binary classifier's performance. It can be thought of as the proportion of samples which were correctly placed into their class. The accuracy ranges from 0 to 1, with higher accuracies, greater than 0.5, being better. An accuracy of 0.5 indicates that the model has performed the same as random classification of the data points would perform.

The formula for calculating accuracy is

$$\text{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y)$$

where 1() is the indicator function (formula from [Scikit-learn Developers, n.d.(b)]).

As mentioned above, a high accuracy score may not reflect a model's actual performance in an imbalanced dataset.

Confusion Matrix

For the remaining metrics, it is helpful to discuss the confusion matrix. The confusion matrix contains four values:

1. true positives (TP): number of data points which are actually positive, and correctly predicted as positive by the classifier
2. false positives (FP): number of data points which are actually negative, but incorrectly predicted as positive by the classifier
3. true negatives (TN): number of data points which are actually negative, and correctly predicted as negative by the classifier
4. false negatives (FN): number of data points which are actually positive, but incorrectly predicted as negative by the classifier

The TP, FP, TN, and FN values are illustrated in the confusion matrix in Figure 3.4.

With the definitions of TP, FP, TN, and FN in hand, we can define the remaining evaluation metrics.

Balanced Accuracy

Blaanced accuracy can be thought of as a weighted average of the accuracies for each sample, where each sample is weighted by the inverse of the prevalence of its actual class [Scikit-learn Developers, n.d.(b)]. It is not as prone to the providing inflated results on imbalanced datasets as accuracy is [Scikit-learn Developers, n.d.(b)]. Balanced accuracy can be interpreted like an accuracy

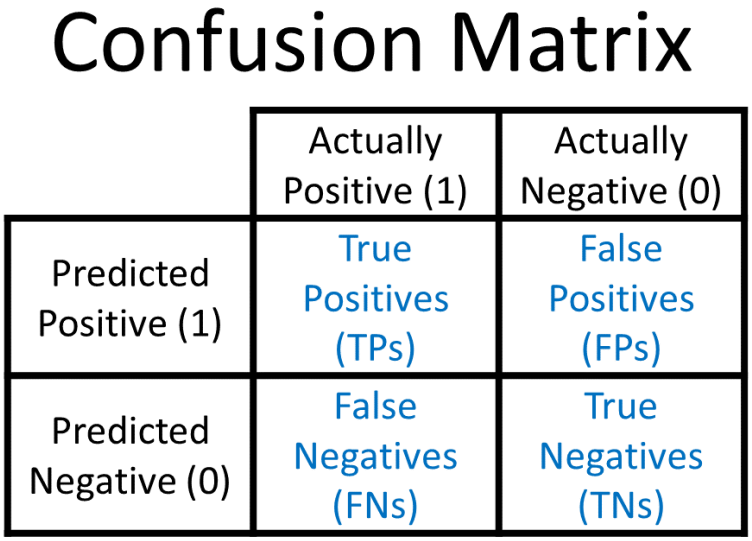

Figure 3.4: Illustration of the confusion matrix with true positives, false positives, false negatives, and true negatives, from Draelos, 2019.

score, and it ranges from 0 to 1, with higher scores being better. In datasets with equal class balance, accuracy and balanced accuracy are equal [Scikit-learn Developers, n.d.(b)].

The formula for calculating balanced accuracy is

$$\text{balanced\_accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

(formula from [Scikit-learn Developers, n.d.(b)]).

Precision

Precision can be thought of as a model's ability to only identify as positive those data points that are actually positive, avoiding false positives. It ranges from 0 to 1, with higher values being better.

The formula for calculating precision is

$$\text{precision} = \frac{TP}{TP+FP}$$

Recall

Recall can be thought of as a model's ability to not miss positive samples and incorrectly label them as negative. It ranges from 0 to 1, with higher values being better.

The formula for calculating recall is

$$\text{precision} = \frac{TP}{TP + FN}$$

F1 score

F1 score is a weighted average (harmonic mean) of precision and recall. It ranges from 0 to 1, with higher values being better.

The formula for calculating F1 score is

$$\text{F1\_score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

ROC AUC

The receiver operating characteristic (ROC) curve is a plot of the true positive rate on the y-axis and the false positive rate on the x-axis. The true positive (TP) rate is the same as recall. The false positive (FP) rate is the proportion of negative data points that are incorrectly labeled as positive out of all negative data points.

The formula for calculating the false positive rate is

$$\text{false\_positive\_rate} = \frac{FP}{FP + TN}$$

The area under the receiver operating characteristic curve (ROC AUC) is the two-dimensional area under the ROC curve. An example of the ROC curve and the area under the curve are shown in Figure 3.5.

Intuitively, as we vary the classification threshold of our model (the value above which the model needs to score a data point for it to be considered

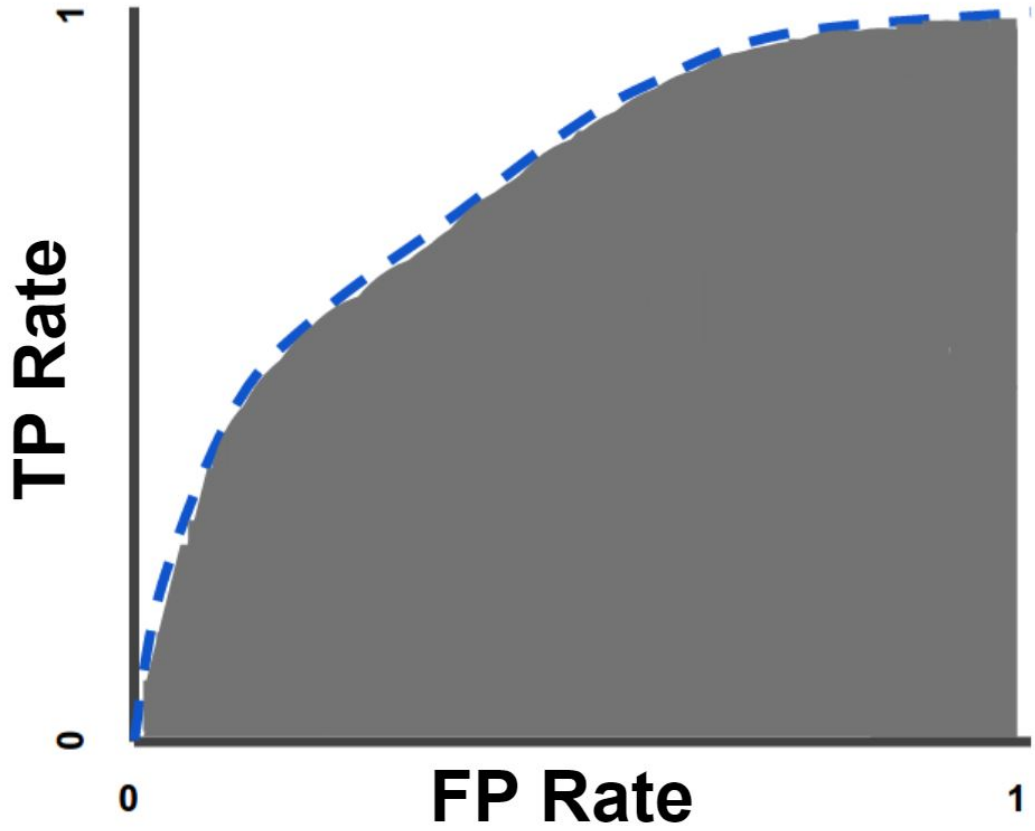

Figure 3.5: Illustration of the receiver operating characteristic (ROC) curve and the area under the curve (AUC). The y-axis is the true positive rate, and the x-axis is the false positive rate. The area under the curve (AUC) is shaded in. Image from Google Developers, 2022.

positive), we achieve different TP and FP rates. These different TP and FP rates are shown in the ROC curve. At all classification thresholds, we would like to maximize the true positive rate. As such, we would like to maximize the area under the ROC curve, as this would achieve a maximization of true positives across all values of the false positive rate.

The value of ROC AUC can be thought of as the classifier's ability to distinguish between the classes at all classification thresholds. The ROC AUC ranges from 0 to 1, with values closer to 1 indicating better performance. An ROC AUC of 0.5 indicates an inability of the classifier to distinguish

between the classes.

Matthews Correlation Coefficient (MCC)

Matthews Correlation Coefficient (MCC) is a value useful for indicating a model's predictive performance, even on unbalanced classes. It ranges from 1 to -1 and can be interpreted as a correlation coefficient, with a value of 1 indicating a perfect positive correlation, a value of 0 indicating no correlation, and a value of -1 indicating a perfect negative correlation [Scikit-learn Developers, n.d.(b)].

The formula for calculating MCC is

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

(formula from [Scikit-learn Developers, n.d.(b)]).

3.4 Genome sequencing

Next-generation sequencing, which is massively parallel and high-throughput, has enabled the collection of vast amounts of genomic data. One popular platform for next-generation sequencing is the set of products by the company Illumina. Illumina sequencing has been used by a variety of medical data projects, including ADNI.

The process for converting a file outputted by the Illumina sequencing technology to a usable DNA sequence file is multi-step. First, raw sequence files in FASTQ format are generated by the sequencing machine. FASTQ format is very similar to the FASTA format listed below. These files are of small snippets of the genome, since the chromosomes are cut into pieces during sequencing to allow for sequencing to be conducted in a massively parallel manner. Next, these FASTQ raw sequence files are aligned to the reference genome using an aligner such as the Burrows-Wheeler Aligner. The reference genome is a consensus sequence of the human genome that research bodies have agreed

upon to be the standard human genome, without mutations. An example of a reference genome is GRCh37 (Genome Reference Consortium Human Build 37). Next, the aligned genome is converted to a BAM file. A BAM is a binary version of a file containing sequence alignment data. Then, variant calling is conducted on the BAM file to obtain the differences between the individual's genome and the reference genome, known as variants. The results of this are stored in a VCF file. More information on the next-generation sequencing data processing pipeline is contained in a document by University of Warwick [Nho, n.d.].

Analysis can be conducted on VCF files, but I chose to work with FASTA files because they provide the DNA sequence in an easy-to-read format. Another processing step needs to be completed to convert the VCF files to FASTA files. If desired, we can extract only the data from the chromosome or region we are examining into the FASTA file. The FASTA files I had access to only contained the DNA sequences from chromosome 19, so I only studied genes located on chromosome 19.

The ultimate file format I worked with looked like this:

```
>19
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
GATCACAGAGGCTGGGCTGCTCCCCACCCTCTGCACACCTCCTGCTTCTAACAGCAGAGC
TGCCAGGCCAGGCCCTCAGGCAAGGGCTCTGAAGTCAGGGTCACCTACTTGCCAGGGCCG
ATCTTGGTGCCATCCAGGGGGCCTCTACAAGGATAATCTGACCTGCAGGGTCGAGGAGTT
GACGGTGCTGAGTTCCCTGCACTCTCAGTAGGGACAGGCCCTATGCTGCCACCTGTACAT
GCTATCTGAAGGACAGCCTCCAGGGCACACAGAGGATGGTATTTACACATGCACACATGG
CTACTGATGGGGCAAGCACTTCACAACCCCTCATGATCACGTGCAGCAGACAATGTGGCC
TCTGCAGAGGGGGAACGGAGACCGGAGGCTGAGACTGGCAAGGCTGGACCTGAGTGTCGT
CACCTAAATTCAGACGGGGAACTGCCCCTGCACATAGTGAACGGCTCACTGAGCAAACCC
```

FASTA files consist of labels and sequences. The labels are on the line starting with '>'. In this case, the label is 19 to indicate this sequence came from chromosome 19. The line after the label contains the sequence. The sequence is the sequence of DNA present in this individual's genome. A value of N indicates ther was no data present at that location. To my knowledge, the sequences I used in my study were not missing nucleotides.

The end results of completing the genomic data processing pipeline for one individual's sequence are two FASTA files, one for each set of alleles belonging to a person. For my analysis, I extracted the DNA sequences for each individual from each FASTA file, referencing the sequence I wanted by its location within the chromosome.

4 Methods

4.1 Dataset

4.1.1 ADNI

Description

This project used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). As per the ADNI website, 'Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD)' [Alzheimer's Disease Neuroimaging Initiative, 2016].

For this project, I used ADNI's whole genome sequencing (WGS) dataset. The WGS data was collected using Illumina technology during 2012-2013 from ADNI participants [Alzheimer's Disease Neuroimaging Initiative, 2015]. The

dataset has 518 individuals labeled as cases: those with confirmed cases of AD or mild cognitive impairment (MCI) and 276 patients as controls. MCI is a form of cognitive impairment that does not affect a person's ability to function in their daily activities, while dementia can affect a person's daily activities. MCI is regarded as a transitional stage between normal functioning and dementia. It may progress to dementia, but does not always [Campbell *et al.*, 2013]. In keeping with the labels provided by the ADNI dataset as to whether each patient is a case or control, I labeled the patients with AD or MCI as positive (1) and those without either as negative (0).

From the whole genome sequence BAM files, the data were aligned to the Genome Reference Consortium Human Build 37 (GRCh37) reference genome, and the consensus sequence was saved to FASTA files. The preprocessing to these FASTA files was completed by other lab members.

From these FASTA files, four genes were investigated: *APOE* (chr19: 45,409,038 - 45,412,650), *TOMM40* (chr19: 45,394,476 - 45,406,946), *SIGLEC11* (chr19: 50,452,249 - 50,464,429) and *EXOC3L2* (chr19: 45,715,878 - 45,737,469). *SIGLEC11* stands for Sialic Acid Binding Ig Like Lectin 11, and *EXOC3L2* stands for Exocyst Complex Component 3 Like 2. The locations of these genes were given by the UCSC Genome Browser [Kent *et al.*, 2002]. These genes were chosen because they were shown to be related to AD by previous analyses published in journal articles (*APOE*, *TOMM40*) or conducted by lab members (*SIGLEC11*, *EXOC3L2*) [Potkin *et al.*, 2009; Saykin *et al.*, 2010; C.-C. Liu *et al.*, 2013]. The DNA sequences of these four genes for each patient were extracted from these FASTA files using the pysam library. Unless otherwise indicated, the chromosome positions reported refer to the GRCh37 reference genome.

Because humans have two copies of each chromosome, each patient has two alleles for each location on the chromosome. This results in two FASTA files being available for each patient, with each FASTA file containing one of the two alleles present at each location in the chromosome. Due to time constraints,

Table 4.1: Number of SNPs found in each gene of ADNI's WGS data.

Gene	Gene length	Number of SNPs	Percent of nucleotides with differences
APOE	3,612	14	0.388%
TOMM40	12,471	106	0.850%
SIGLEC11	12,181	86	0.706%
EXOC3L2	21,592	199	0.922%

only one set of alleles for each patient was investigated for all the experiments described below. The pre-trained DNABERT model was also set up to only take in one set of alleles at a time, which is why I began with the approach of inputting one allele at a time rather than both.

Exploratory Data Analysis

To determine how much of a signal was present in the dataset, and how much the DNA sequences actually differed between the case and control patients, I ran an analysis to determine the positions of the nucleotides which had any differences at all between the patients. My motivation for performing this analysis was that the transformer models, which were the first I tried, were not picking up a signal from the data, so I wanted to check how much variation was present in the data and be sure that the patients' DNA sequences were not all identical or near-identical.

The results of this analysis are shown in Table 4.1. A SNP was defined as any nucleotide that did not have an identical value for all patients in the dataset. All other nucleotides not counted in the number of SNPs were identical between all the patients. On average, each of the genes examined only had 0.716% of their nucleotides having any differences at all between patients. This is a very small amount of signal to work with, especially when considering that the transformer would be operating on the entire gene or windows of the gene at once. The results of this exploratory data analysis show that there is a low signal to noise ratio in the ADNI WGS dataset.

4.2 SVM model

For each of the SNPs found in each of the four genes investigated, the nucleotides were one-hot encoded (as a four-bit vector, with each bit representing the count of the nucleotides A, T, C, and G), and then an SVM model was run on this encoding along with its positive or negative label to attempt to classify the patient as a case of AD or as a control. A second set of experiments was run, in which all SNPs found for each gene were one-hot encoded and inputted into the SVM at once for classification of AD. Three different SVM kernels were tested: linear, sigmoid, and RBF. The linear kernel experiment used the scikit-learn implementation of the stochastic gradient descent classifier with the hinge loss function [Scikit-learn Developers, n.d.(c)]. The sigmoid and RBF kernels used the corresponding settings from the scikit-learn C-Support Vector Classification function [Scikit-learn Developers, n.d.(d)]. All hyperparameters were set to the scikit-learn defaults. The numpy random seed was set to 42 for all SVM experiments.

Different methods of balancing the dataset were investigated as well. In addition to running the SVM models on the full dataset, I also ran random downsampling to create a balanced dataset. This is to avoid issues that come with there being far more data points in one class than the other, as there were almost twice as many case as control samples in the dataset. Random downsampling involves randomly choosing a subset of the data points from the larger class to keep in the dataset, discarding the rest of the data points from this class. The random downsampling approach resulted in 276 data points remaining in each of the case and control classes. When the train-test split was taken, the same dataset balance, whether that of the original dataset or the downsampled dataset, was maintained in each of the training and testing splits. A 90%/10% training-testing split of the was used for the models. A split with such a high proportion of data in the training split was chosen because

of the small size of the dataset, to give the model as much data as possible to train on. The prediction results for each gene on both the original and downsampled datasets are shown in Section 5.1.

The SVM's performance on the metrics of accuracy, balanced accuracy, precision, recall, F1 score, ROC AUC, and MCC were calculated. The scikit-learn implementation of these metrics was used [Scikit-learn Developers, n.d.(b)]. Along with reporting the location of each SNP in the chromosome and its prediction accuracy results, the rsID (reference SNP cluster ID) of each SNP was also reported. The rsID, also referred to as an rs number, is a unique identifier for each SNP that has been assigned by researchers. The rsIDs were included so that the results of this study can easily compared to the many other studies which also identify their SNPs using rsIDs, and to make the results of this study more relevant in a biological context to other researchers studying the genetic basis of AD and other diseases. The rsID for each SNP was programmatically retrieved from the NCBI SNP database [National Library of Medicine, n.d.] with the Biopython Entrez package [The Biopython Contributors, n.d.] using its chromosome number and position within the chromosome as the search term.

The results of the SVM experiments are described in Section 5.1.

4.3 Transformer model

A transformer architecture was chosen as the deep learning approach for this project. The success of the DNABERT and Enformer models [Ji *et al.*, 2021; Avsec *et al.*, 2021] in predicting traits related to DNA from unbroken stretches of whole genome sequencing data seemed promising for the use of transformers to predict the presence or absence of AD from DNA sequencing data. In addition, the nature of a transformer may be well-suited to tasks relating to prediction from DNA. As described in Section 2.2, DNA shares

similarities to natural language, and transformers have achieved great success on natural language processing tasks. Furthermore, as discussed by Ji *et al.*, 2021, transformers have the advantage of taking global contexts into account via their self-attention mechanisms, while CNNs only take local contexts into account. Taking global contexts into account could be useful for DNA-related tasks, since DNA regions which are far apart may impact each others' expression. As such, because of the strength and flexibility of transformers, and the promising previous results achieved with them, transformers were chosen for this project. I used the pre-trained DNABERT implementation because I thought the additional information about DNA's structure and patterns that the model had learned during its 25-day pre-training process might come in useful for the AD predictions. The DNABERT model had already been shown to be generalizable to a variety of tasks relating to predictions on DNA [Ji *et al.*, 2021], and I thought it might work on AD prediction as well.

The pre-trained DNABERT model, as described by Ji *et al.*, 2021, was used. The DNABERT model is a bidirectional encoder representation transformer, pre-trained on the task of completing missing pieces of DNA sequence that were removed from training data of actual DNA sequences. The architecture of the DNABERT model is shown in Figure 4.1. The DNABERT model contains 3 embedding layers, 12 transformer blocks which then feed to an additional hidden layer, and a classification layer.

I used the pre-trained DNABERT model that was set up to perform binary classification tasks, specifically the dnaprom task of predicting whether a region of DNA is a promoter. I used the pre-trained model that was trained to fill in missing sections of DNA, rather than the model fine-tuned on predicting whether a region is a promoter. I ran the DNABERT model on two tasks, described in the sections below. For both of these tasks, I used the hyperparameters described in Table 4.2, which were the defaults provided by Ji *et al.*, 2021 in their model code. I also balanced the ADNI dataset to ensure the

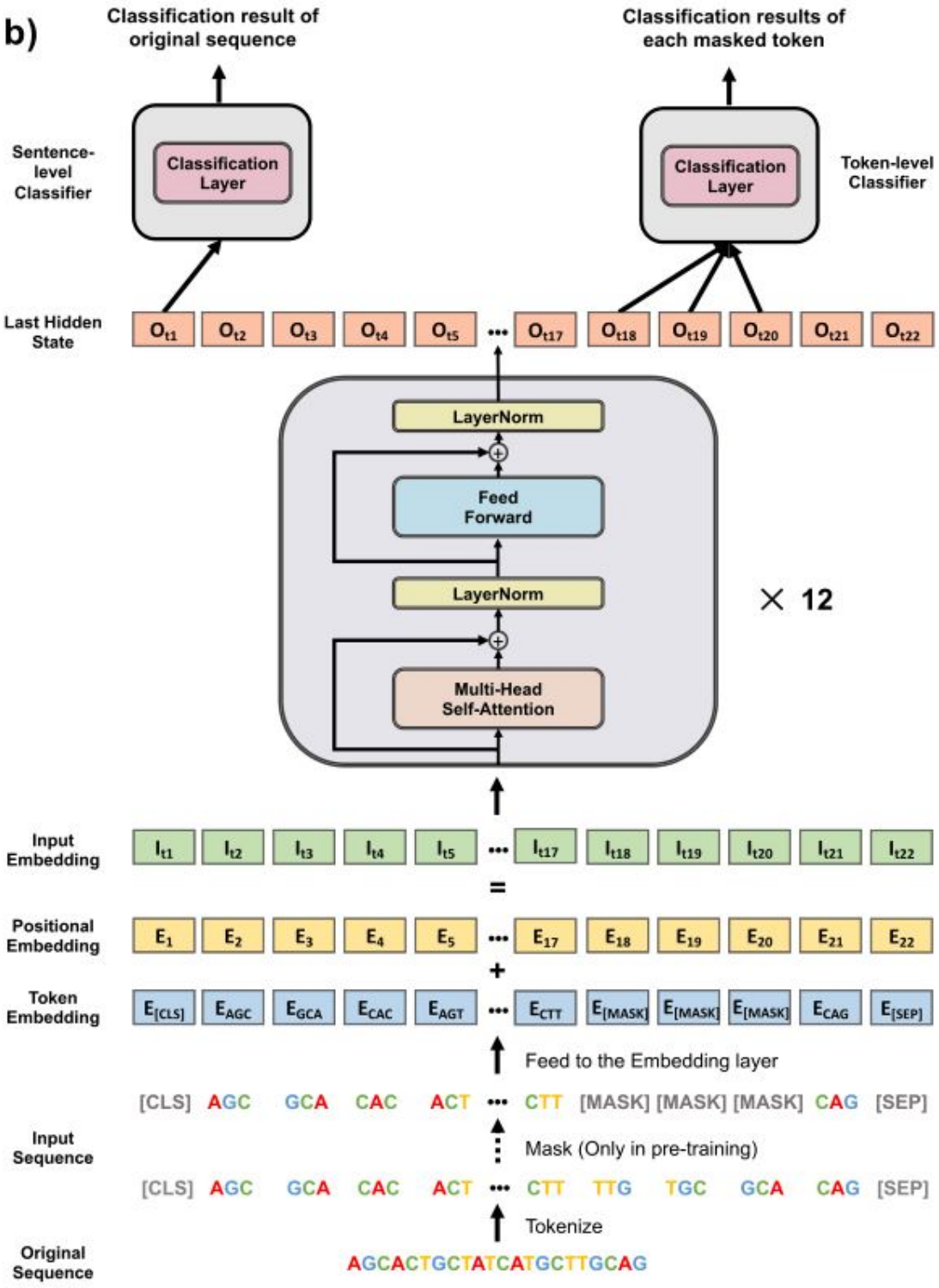

Figure 4.1: DNABERT transformer model's architecture, from Ji *et al.*, 2021's paper.

Table 4.2: Transformer model's hyperparameters.

Hyperparameter	Value
Batch size	32
Learning rate	2.00E-04
Training epochs	5
Warmup percent	0.1
Hidden dropout probability	0.1
Weight decay	0.01
Max sequence length	100

number of cases and controls was the same by using random downsampling, resulting in 276 each of case and control samples.

4.3.1 On each window of *APOE* gene

I fine-tuned this pre-trained DNABERT model on the task of predicting if an individual has a case of AD (labeled 1) or is a control (labeled 0) based on a 100 bp-long window of their *APOE* gene sequence. The *APOE* gene sequence was split into windows of size 100, except for the window at the very end of the sequence, which was made only as long as the remaining number of nucleotides left in the *APOE* gene after all the previous 100 bp windows had been allotted. Between each subsequent window, there was a 50-bp overlap. A separate transformer model was trained for each window of the *APOE* gene. Each window's DNA sequence was tokenized into k-mers with $k = 6$ using the methodology described by Ji *et al.*, 2021, as this was the preprocessing methodology they used on the raw DNA sequence data inputted to DNABERT. A random seed of 24, the default used by the DNABERT authors, was used for all DNABERT experiments. The results of these experiments are described in Section 5.2.1.

4.3.2 On entire gene sequences

In addition to running the models on the smaller windows of the *APOE* gene, I also ran the models on the task of predicting if an individual has AD based on the full gene sequences of each of the four genes, *APOE*, *TOMM40*, *SIGLEC11*, and *EXOC3L2*. These results are also described in Section 5.2.1.

4.3.3 Introduction of artificial signal

I found that my transformer models were achieving poor results on the ADNI dataset, and I wanted to rule out the possibility of an error in the set up and training. I also wanted to validate that the transformer could indeed pick up on signals within the data. As such, I introduced an artificial signal into the DNA sequence data from the ADNI dataset. The artificial signal was as follows: a certain percentage of the DNA sequences that belonged to patients with cases of AD were changed to sequences of all 'A' nucleotides in both the train and test data. The DNA sequences were also preprocessed into 6-mers, as described in Section 4.3.1. The DNABERT transformer was then trained and tested on this altered dataset. This experiment was conducted setting the percentage of positive data points altered to 10%, 70%, and 90%. The 70% level was chosen because AD is considered to be approximately 70% heritable, and I wanted to see the models' prediction performance when 70% of the data points had a strong genetic signal. The 90% and 10% levels were chosen to see the results when almost all, or almost none, of the positive data points, respectively, had a signal. The results of these experiments are described in Section 5.2.3.

The results on the metrics of accuracy, F1 score, and ROC AUC during training and testing are provided in Section 5.2. These metrics were calculated by functions within Ji *et al.*, 2021's code and utilized the scikit-learn functions for these metrics [Scikit-learn Developers, n.d.(b)]. I also modified the training and testing code to enable logging of metrics for the visualization of training curves.

5 Results

5.1 SVM results

The results of the SVM experiments on the SNPs within the *APOE*, *TOMM40*, *SIGLEC11*, and *EXOC3L2* genes in the ADNI dataset, using a linear kernel and without any random downsampling, are shown in Tables 5.1, 5.3, 5.5, and 5.7. The results that the linear kernel SVM was able to achieve on prediction of AD using just the value of this single nucleotide on the metrics of accuracy, balanced accuracy, precision, recall, F1 score, ROC AUC, and MCC are shown in the tables. Each SNP is identified by its location within chromosome 19 and its rsID, which is a unique identifier for each SNP. The genes listed in these tables had more than ten SNPs present in the ADNI dataset. Only the top ten SNPs for which the highest metric of balanced accuracy was achieved are shown in the tables, in order to highlight only the most promising results. The full results for all SNPs found in each of the four genes are shown in the Supplementary Materials. Out of all the metrics that take imbalanced dataset classes into account, the metric of balanced accuracy was chosen for ranking the results because it is easy to interpret, as it can be thought of as an accuracy score.

In addition, balanced accuracy correlates closely with the ROC AUC score, which is another commonly used metric for binary classification performance. The rest of the various metrics calculated do not necessarily correlate with each other in the results obtained, so it was essential to pick only one metric on which to base the prioritization of results. (Note: For any results in which the MCC calculation result was undefined due to the denominator being 0, the MCC value was reported as 0, which is the default scikit-learn behavior.)

In addition to reporting the SVM results for the full, unbalanced dataset, the results for the datasets for each gene balanced with random downsampling to achieve an equal number of case and control samples are shown in Tables 5.2, 5.4, 5.6, and 5.8. The effects of random downsampling were mixed, but it seemed to improve the results for *SIGLEC11* and worsen the results for *APOE*.

The SNP with the highest SVM prediction performance was the SNP at position 45,409,167 within the *APOE* gene of chromosome 19, with a corresponding rsID of rs11542028. This SNP achieved a balanced accuracy of 0.65, an F1 score of 0.77, an ROC AUC of 0.65, and an MCC of 0.28 when investigated without random downsampling. These values are fairly high across all the metrics. The next highest score for balanced accuracy was achieved at the SNP at position 50,454,375 within the *SIGLEC11* gene of chromosome 19, with a corresponding rsID of rs117180821. This SNP achieved a balanced accuracy of 0.61, an F1 score of 0.36, an ROC AUC of 0.61, and an MCC of 0.36 when investigated with random downsampling.

The highest balanced accuracy performance within the *TOMM40* gene was achieved at the SNP at position 45,406,673 within chromosome 19, with a corresponding rsID of rs58185379. This SNP achieved a balanced accuracy of 0.59, an F1 score of 0.59, an ROC AUC of 0.59, and an MCC of 0.15 when investigated without random downsampling. Furthermore, the highest balanced accuracy performance within the *EXOC3L2* gene was achieved at the SNP at position 45,722,517 within chromosome 19, with a corresponding

rsID of rs60528995. This SNP achieved a balanced accuracy of 0.56, an F1 score of 0.2, an ROC AUC of 0.56, and an MCC of 0.24 when investigated with random downsampling.

The effects of varying the kernels and sampling techniques are shown in Table 5.9. The SNP with the highest score across the evaluated metrics, rs11542028 located at 45,409,167 in chromosome 19 within the *APOE* gene, was used for this analysis because this SNP provides a larger amount of signal to work with, better illustrating the positive or negative effects of varying kernel and sampling techniques, as opposed to a SNP without any signal, which might show the same results of there being no signal regardless of which techniques are tried on it. Linear, polynomial, and RBF kernels were used for the SVM, and the results of each are shown. The kernel achieving the highest balanced accuracy was the linear kernel. As such, this kernel was used for the other experiments testing the SVM on the various genes' SNPS. The effects of varying the sampling techniques are also shown in Table 5.9. Using random downsampling decreased the ROC AUC compared to the peak value achieved without downsampling using the linear kernel for this particular SNP. However, random downsampling increased the balanced accuracy performance for other genes and SNPS, and increased the balanced accuracy for the polynomial and RBF kernels, as shown in Table 5.9.

Table 5.10 shows the results of running the SVM linear kernel model on all SNPs of each of the four genes at once, with the SNP data one-hot encoded. Table 5.11 shows the results of the same experiment but with the datasets balanced using random downsampling. The results were generally worse for the SVM model operating on all SNPs in each gene compared to the results of the top SNP for each gene, for both of the experiments with and without downsampling. For all four genes, running the SVM on all SNPs decreased the balanced accuracy, precision, ROC AUC, and MCC, but increased the

Table 5.1: AD prediction accuracy results when running SVM on ADNI SNPs from *APOE* gene with linear kernel.

rsID	Location in chr19	Accuracy	Balanced accuracy	Precision	Recall	F1 score	ROC AUC	MCC
rs11542028	45,409,167	0.69	0.65	0.83	0.73	0.77	0.65	0.28
rs1486677963	45,409,283	0.75	0.52	0.75	1	0.86	0.52	0.19
rs769451	45,410,911	0.74	0.52	0.74	0.98	0.85	0.52	0.09
rs9282609	45,409,113	0.74	0.5	0.74	1	0.85	0.5	0
rs769448	45,409,579	0.74	0.5	0.74	1	0.85	0.52	0
rs769449	45,410,002	0.74	0.5	0.74	1	0.85	0.64	0
rs61357706	45,410,273	0.74	0.5	0.74	1	0.85	0.51	0
rs74253333	45,410,444	0.74	0.5	0.74	1	0.85	0.51	0
rs115299243	45,410,548	0.74	0.5	0.74	1	0.85	0.51	0
rs201672011	45,411,064	0.74	0.5	0.74	1	0.85	0.51	0

Table 5.2: AD prediction accuracy results when running SVM on ADNI SNPs from *APOE* gene with linear kernel, with random downsampling.

rsID	Location in chr19	Accuracy	Balanced accuracy	Precision	Recall	F1 score	ROC AUC	MCC
rs11542028	45,409,167	0.58	0.58	0.56	0.74	0.63	0.58	0.18
rs769449	45,410,002	0.58	0.58	0.67	0.3	0.41	0.58	0.19
rs9282609	45,409,113	0.49	0.5	0.49	1	0.66	0.5	0
rs1486677963	45,409,283	0.49	0.5	0.49	1	0.66	0.5	0
rs769448	45,409,579	0.49	0.5	0.49	1	0.66	0.46	0
rs61357706	45,410,273	0.49	0.5	0.49	1	0.66	0.5	0
rs74253333	45,410,444	0.49	0.5	0.49	1	0.66	0.52	0
rs115299243	45,410,548	0.49	0.5	0.49	1	0.66	0.5	0
rs769451	45,410,911	0.51	0.5	0	0	0	0.5	0
rs201672011	45,411,064	0.49	0.5	0.49	1	0.66	0.5	0

recall and F1 score (in the case of *EXOC3L2* without downsampling, the MCC remained the same).

5.2 Transformer results

5.2.1 On each window of *APOE* gene

The results of the transformer experiments on the windows of the *APOE* gene within the ADNI dataset's WGS data are shown in Table 5.12. Only the results

Table 5.3: AD prediction accuracy results when running SVM on ADNI SNPs from *TOMM40* gene with linear kernel.

rsID	Location in chr19	Accuracy	Balanced accuracy	Precision	Recall	F1 score	ROC AUC	MCC
rs58185379	45,406,673	0.52	0.59	0.82	0.46	0.59	0.59	0.15
rs34095326	45,395,844	0.4	0.58	0.92	0.2	0.33	0.58	0.19
rs11668327	45,398,633	0.7	0.57	0.77	0.85	0.81	0.57	0.15
rs59841965	45,406,798	0.76	0.55	0.76	1	0.86	0.55	0.27
rs71337246	45,397,512	0.56	0.52	0.75	0.61	0.67	0.52	0.03
rs2238680	45,398,264	0.56	0.52	0.75	0.61	0.67	0.52	0.03
rs386539078	45,398,716	0.56	0.52	0.75	0.61	0.67	0.52	0.03
rs140684051	45,399,456	0.74	0.52	0.74	0.98	0.85	0.52	0.09
rs183743534	45,404,866	0.75	0.52	0.75	1	0.86	0.52	0.19
rs59915866	45,397,229	0.72	0.51	0.74	0.97	0.84	0.51	0.03

Table 5.4: AD prediction accuracy results when running SVM on ADNI SNPs from *TOMM40* gene with linear kernel, with random downsampling.

rsID	Location in chr19	Accuracy	Balanced accuracy	Precision	Recall	F1 score	ROC AUC	MCC
rs16979513	45,396,144	0.56	0.56	0.64	0.26	0.37	0.56	0.15
rs58185379	45,406,673	0.56	0.56	0.58	0.41	0.48	0.56	0.13
rs74253332	45,404,691	0.55	0.55	0.53	0.74	0.62	0.55	0.11
rs76841546	45,402,589	0.53	0.52	1	0.04	0.07	0.52	0.14
rs183743534	45,404,866	0.51	0.52	0.5	1	0.67	0.52	0.13
rs117843462	45,405,634	0.53	0.52	1	0.04	0.07	0.52	0.14
rs59841965	45,406,798	0.51	0.52	0.5	1	0.67	0.52	0.13
rs1160985	45,403,412	0.51	0.51	0.5	0.67	0.57	0.51	0.03
rs56951511	45,403,858	0.51	0.51	0.5	0.67	0.57	0.51	0.03
rs59019406	45,404,431	0.51	0.51	0.5	0.67	0.57	0.51	0.03

for the first 10 windows are shown for brevity, while the full results for the all the windows are shown in the Supplementary Materials. 'AUC' in the DNABERT results tables refers to ROC AUC.

As can be seen, none of the windows had a signal for either the testing or training tasks. The accuracy of approximately 50% on the dataset balanced with random downsampling is similar to what would have been achieved with predicting the outcome randomly. The relatively low F1 score and the AUC

Table 5.5: AD prediction accuracy results when running SVM on ADNI SNPs from *SIGLEC11* gene with linear kernel.

rsID	Location in chr19	Accuracy	Balanced accuracy	Precision	Recall	F1 score	ROC AUC	MCC
rs61467868	50,456,770	0.49	0.53	0.76	0.44	0.56	0.53	0.05
rs57860877	50,457,346	0.49	0.53	0.76	0.44	0.56	0.53	0.05
rs73932071	50,459,806	0.28	0.51	1	0.02	0.03	0.51	0.07
rs4802641	50,452,341	0.74	0.5	0.74	1	0.85	0.5	0
rs114819375	50,452,606	0.74	0.5	0.74	1	0.85	0.5	0
rs143688215	50,452,641	0.74	0.5	0.74	1	0.85	0.5	0
rs138574928	50,452,960	0.74	0.5	0.74	1	0.85	0.5	0
rs200448773	50,453,203	0.74	0.5	0.74	1	0.85	0.5	0
rs56579996	50,453,317	0.74	0.5	0.74	1	0.85	0.43	0
rs201942673	50,453,351	0.74	0.5	0.74	1	0.85	0.5	0

Table 5.6: AD prediction accuracy results when running SVM on ADNI SNPs from *SIGLEC11* gene with linear kernel, with random downsampling.

rsID	Location in chr19	Accuracy	Balanced accuracy	Precision	Recall	F1 score	ROC AUC	MCC
rs117180821	50,454,375	0.62	0.61	1	0.22	0.36	0.61	0.36
rs2076155786	50,454,383	0.62	0.61	1	0.22	0.36	0.61	0.36
rs117126572	50,457,876	0.62	0.61	1	0.22	0.36	0.61	0.36
rs76691680	50,457,915	0.62	0.61	1	0.22	0.36	0.61	0.36
rs117971487	50,457,927	0.62	0.61	1	0.22	0.36	0.61	0.36
rs117428283	50,458,411	0.62	0.61	1	0.22	0.36	0.61	0.36
rs79972908	50,458,488	0.62	0.61	1	0.22	0.36	0.61	0.36
rs56579996	50,453,317	0.56	0.57	0.53	0.89	0.67	0.57	0.18
rs62126307	50,454,086	0.56	0.57	0.53	0.89	0.67	0.57	0.18
rs10405621	50,455,351	0.56	0.57	0.53	0.89	0.67	0.57	0.18

around 0.5 also indicate a lack of signal.

In addition to conducting the experiments described in section 4.3, I also tried the DNABERT transformer model on the tasks of predicting longer windows of 200 bp and predicting the presence of AD using a window surrounding a known significant AD SNP within the *APOE* gene from the GWAS Catalog [EMBL-EBI, n.d.]. I also tried using a lower learning rate of $1.00 \cdot 10^{-5}$ as compared to the original learning rate of $2.00 \cdot 10^{-4}$. None of

Table 5.7: AD prediction accuracy results when running SVM on ADNI SNPs from *EXOC3L2* gene with linear kernel.

rsID	Location in chr19	Accuracy	Balanced accuracy	Precision	Recall	F1 score	ROC AUC	MCC
rs59647713	45,736,003	0.62	0.55	0.76	0.71	0.74	0.55	0.09
rs112759099	45,726,845	0.32	0.54	1	0.08	0.16	0.54	0.15
rs28645301	45,724,692	0.31	0.53	1	0.07	0.13	0.53	0.14
rs28564302	45,724,868	0.31	0.53	1	0.07	0.13	0.53	0.14
rs386809738	45,724,961	0.31	0.53	1	0.07	0.13	0.53	0.14
rs60269219	45,724,963	0.31	0.53	1	0.07	0.13	0.53	0.14
rs59356929	45,725,127	0.31	0.53	1	0.07	0.13	0.53	0.14
rs58213824	45,725,185	0.31	0.53	1	0.07	0.13	0.53	0.14
rs57294488	45,725,481	0.31	0.53	1	0.07	0.13	0.53	0.14
rs73568222	45,725,975	0.31	0.53	1	0.07	0.13	0.53	0.14

Table 5.8: AD prediction accuracy results when running SVM on ADNI SNPs from *EXOC3L2* gene with linear kernel, with random downsampling.

rsID	Location in chr19	Accuracy	Balanced accuracy	Precision	Recall	F1 score	ROC AUC	MCC
rs60528995	45,722,517	0.56	0.56	1	0.11	0.2	0.56	0.24
rs57354345	45,717,615	0.51	0.52	0.5	0.85	0.63	0.52	0.04
rs1969894901	45,724,561	0.51	0.52	0.5	0.89	0.64	0.52	0.05
rs1426173634	45,724,633	0.51	0.52	0.5	0.89	0.64	0.52	0.05
rs12978617	45,724,658	0.51	0.52	0.5	0.89	0.64	0.52	0.05
rs73568222	45,725,975	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs57399322	45,726,106	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs58715307	45,726,458	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs10423753	45,726,563	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs113728460	45,726,654	0.53	0.52	0.6	0.11	0.19	0.52	0.07

Table 5.9: Effect of altering SVM kernel and dataset balancing method on SVM accuracy of predicting AD for SNP with ID rs11542028 in *APOE* gene in ADNI dataset.

Kernel	Dataset balancing method	Accuracy	Balanced accuracy	Precision	Recall	F1 score	ROC AUC	MCC
Linear	None	0.69	0.65	0.83	0.73	0.77	0.65	0.28
Polynomial	None	0.74	0.5	0.74	1	0.85	0.5	0
RBF	None	0.74	0.5	0.74	1	0.85	0.65	0
Linear	Random downsampling	0.58	0.58	0.56	0.74	0.63	0.58	0.18
Polynomial	Random downsampling	0.56	0.57	0.54	0.7	0.61	0.57	0.14
RBF	Random downsampling	0.56	0.57	0.54	0.7	0.61	0.57	0.14

Table 5.10: AD prediction accuracy results when running SVM on all SNPs in each gene tested from ADNI dataset.

Gene	Accuracy	Balanced accuracy	Precision	Recall	F1 score	ROC AUC	MCC
APOE	0.74	0.5	0.74	1	0.85	0.6	0
EXOC3L2	0.74	0.52	0.74	0.98	0.85	0.53	0.09
SIGLEC11	0.7	0.47	0.73	0.95	0.82	0.52	-0.12
TOMM40	0.74	0.5	0.74	1	0.85	0.48	0

Table 5.11: AD prediction accuracy results when running SVM on all SNPs in each gene tested from ADNI dataset, with random downsampling.

Gene	Accuracy	Balanced accuracy	Precision	Recall	F1 score	ROC AUC	MCC
APOE	0.49	0.5	0.49	1	0.66	0.5	0
EXOC3L2	0.45	0.46	0.47	0.93	0.62	0.47	-0.2
SIGLEC11	0.45	0.46	0.47	0.74	0.57	0.46	-0.1
TOMM40	0.53	0.53	0.52	0.63	0.57	0.54	0.06

these methods resulted in a signal being found.

5.2.2 On entire gene sequences

In addition to running the DNABERT model on smaller windows of the *APOE* gene, I also ran it on the entire DNA sequence of each of the four genes at once. The results of these experiments are shown in Table 5.13. As can be seen by accuracy and AUC scores on the balanced dataset being around 0.5, a signal was not found during these experiments either.

Table 5.12: Results of running DNABERT transformer model on each window of *APOE* gene from ADNI dataset, using random downsampling for dataset balance.

Window start pos.	Window end pos.	Train accuracy	Train F1 score	Train AUC	Test accuracy	Test F1 score	Test AUC
45,409,038	45,409,138	0.501	0.334	0.482	0.491	0.329	0.482
45,409,088	45,409,188	0.499	0.333	0.388	0.509	0.337	0.388
45,409,138	45,409,238	0.499	0.333	0.321	0.509	0.337	0.321
45,409,188	45,409,288	0.499	0.333	0.519	0.509	0.337	0.519
45,409,238	45,409,338	0.501	0.334	0.482	0.491	0.329	0.482
45,409,288	45,409,388	0.501	0.334	0.500	0.491	0.329	0.500
45,409,338	45,409,438	0.499	0.333	0.482	0.509	0.337	0.482
45,409,388	45,409,488	0.501	0.334	0.481	0.491	0.329	0.481
45,409,438	45,409,538	0.499	0.333	0.500	0.509	0.337	0.500

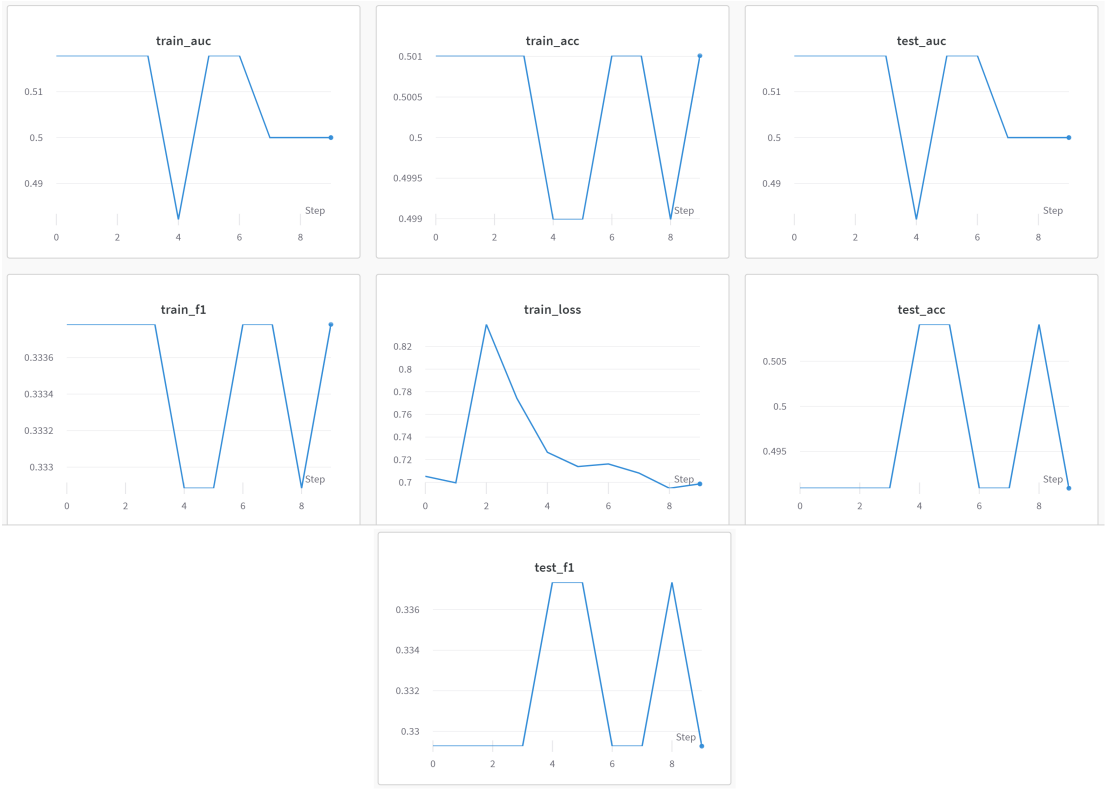

Figure 5.1: Training loss curve and train and test accuracy metric curves over the steps of running the DNABERT transformer model on predicting presence of AD based on the *APOE* gene sequence from the ADNI WGS dataset.

Figure 5.1 shows the training and testing curves for the accuracy, F1 score, and AUC metrics, as well as the training loss curves, over the training steps of the model. The accuracy, F1 score, and AUC do not trend upwards over time, and merely fluctuate, indicating the model struggling to find a signal and converge. The loss function also does not have a clear downward trend, as it ends at a similar place to where it started, also indicating the model having difficulty in training and converging.

5.2.3 With introduction of artificial signal

The results of the transformer experiments on the entire APOE gene within the ADNI dataset's WGS data, with an artificial signal introduced into 10%, 70%, and 90% of the positive samples, are shown in Table 5.14.

Table 5.13: Results of running DNABERT transformer model on entire gene sequences from ADNI dataset, using random downsampling for dataset balance.

Gene	Train accuracy	Train F1 score	Train AUC	Test accuracy	Test F1 score	Test AUC
APOE	0.501	0.334	0.519	0.491	0.329	0.519
EXOC3L2	0.499	0.333	0.446	0.509	0.337	0.446
SIGLEC11	0.501	0.334	0.482	0.491	0.329	0.482
TOMM40	0.499	0.3329	0.5185	0.5091	0.3373	0.5185

The results from when 10% of the positive data points had an artificial signal introduced are similar to the results from the experiments in which no artificial signal was introduced. The train and test accuracy and AUC are approximately 0.5, and the train and test F1 scores do not surpass 0.5. This suggests that the model was not able to find a signal with only 10% of the data points having an artificial signal either.

The results for the experiments in which 70% and 90% of the data points had an artificial signal introduced look much better. The test accuracy, F1 score, and AUC results are slightly higher than the percentage of positive data points for which an artificial signal was introduced, thereby surpassing either 70% or 90% in each case, which can be interpreted as a high accuracy. This indicates that the DNABERT model was able to distinguish between the positive and negative data points when the artificial signal was introduced in these cases.

In addition, the training and testing accuracy metric curves and the training loss curve over the training steps for the experiment in which an artificial signal was introduced in 90% of data points are shown in Figure 5.2. As can be seen, the loss curve has a clear downward trend over time. The training and testing curves for accuracy, F1 score, and AUC do not have a clear upward trend. Rather, they fluctuate slightly or remain the same, but all started at a high value in the first iteration, indicating a generally high performance by the model.

Table 5.14: Results of running DNABERT transformer model on the entire APOE gene sequence from the ADNI dataset, using random downsampling for dataset balance and introducing artificial signal into a percentage of positive-labeled data points.

Artificial signal percentage	Train accuracy	Train F1 score	Train AUC	Test accuracy	Test F1 score	Test AUC
10%	0.539	0.414	0.519	0.527	0.404	0.519
70%	0.847	0.843	0.750	0.764	0.752	0.750
90%	0.952	0.951	0.931	0.927	0.927	0.931

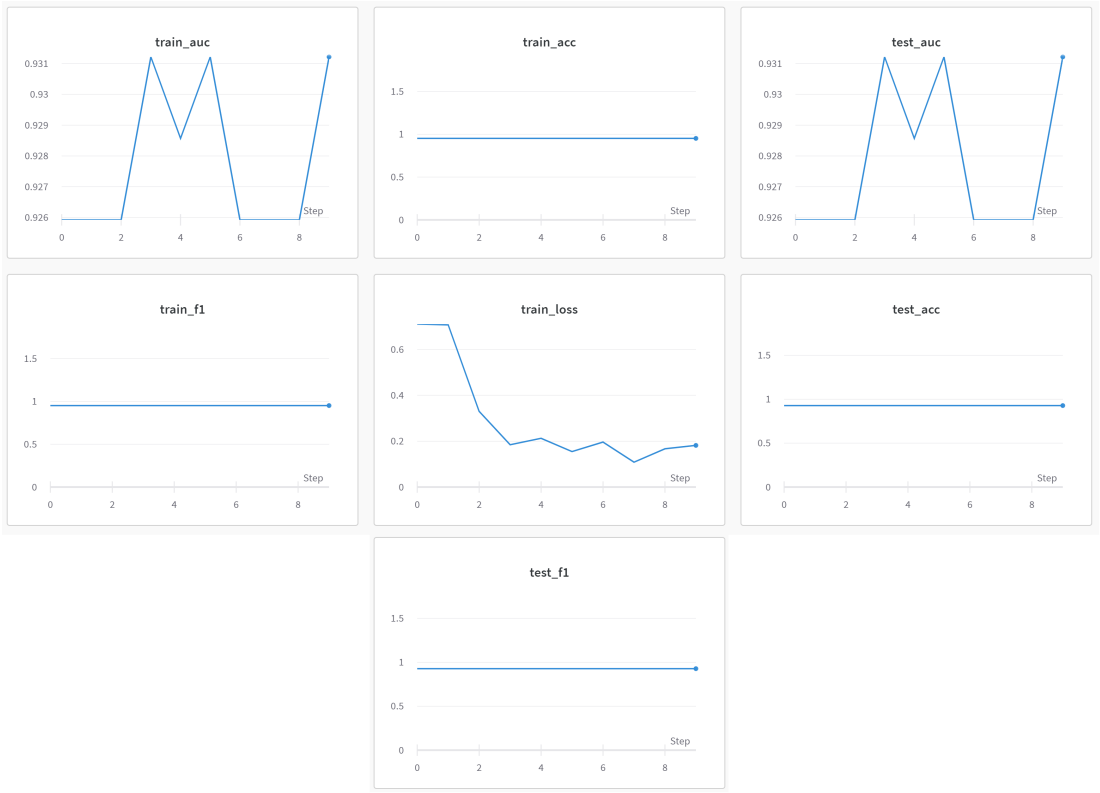

Figure 5.2: Training loss curve and train and test accuracy metric curves over the steps of running the DNABERT transformer model on predicting presence of AD based on the *APOE* gene sequence from the ADNI WGS dataset, having introduced an artificial signal into 90% of positive-labeled data points.

6 Discussion

6.1 Discussion

6.1.1 Interpretation of AD prediction results

The SNP with the single highest overall balanced accuracy and results on other metrics was rs11542028 in the *APOE* gene. The fact that this SNP with highest accuracy metrics came from the *APOE* gene supports previous results, as *APOE* is the gene with the highest known relationship to AD [Potkin *et al.*, 2009; C.-C. Liu *et al.*, 2013]. Previous analyses run on the ADNI dataset have also found the *APOE* gene to be the most significant gene for AD within this dataset [Potkin *et al.*, 2009].

The *SIGLEC11* gene had the greatest number of SNPs with a balanced accuracy above 0.60. While this gene is not commonly cited in studies or review articles about the genetic basis of AD, there have been studies linking this gene to AD. For example, the review article by Salminen and Kaarniranta, 2009 discusses studies linking Siglec receptors and their corresponding proteins to accumulation of amyloid plaques, which are implicated in AD. The high

accuracy results of my study on the *SIGLEC11* gene may suggest that this gene could be a good target for further research on studies related to AD.

One departure that my study has from previous results is that the *APOE* SNP that my study identified as most significant was also not the same as the two SNPs reported by Bellenguez *et al.*, 2020's review article to be most significant: rs429358 and rs7412. A second departure from previous results is that while the *APOE* gene had the single SNP with the highest signal for prediction of AD, and had a couple of other SNPs that demonstrated a signal, the *APOE* gene did not have as many SNPs with a high balanced accuracy as *SIGLEC11* did. This is surprising, since the *APOE* gene is generally regarded to be the gene with the greatest known implication in Alzheimer's disease.

There are several factors we could potentially attribute these departures from previous results to. One factor is the small sample size of this dataset. With such a small dataset used, especially compared to the larger GWAS studies, associations found will have less statistical power and may be more likely to be due to chance. My models may have also labeled false negatives due to the small dataset, and the previously known significant ADNI SNPs could be some of these false negatives. Another factor is the difference in methodology used. While previous studies used a GWAS, including, for example, Potkin *et al.*, 2009's study that found *APOE* to be the most significant gene within the ADNI dataset, my study used SVMs. Using a different model may achieve different results. In addition, the higher number of SNPs with signal found in the other genes as compared to *APOE* may be due in part to the size of the genes; *APOE* is approximately four to seven times smaller in length than the other three genes studied, and has far fewer SNPs present in the dataset than the other genes do, as shown in Table 4.1. As the total number of SNPs in the gene increases, the likelihood of finding SNPs with a signal present may also increase.

Moreover, my study only used one of the two alleles present in its input data. Other studies that used both alleles may have achieved different results

for the findings of which SNPs within *APOE* are most related to AD. The use of only one of the two alleles may also have an effect on which genes are found to be most significant. *APOE* is known to be dosage-dependent: for example, a patient with two alleles of the deleterious *APOE* ε4 allele is more likely to have AD than a patient with only one of these alleles [Bellenguez *et al.*, 2020]. Using only one of the two alleles as input data to the SVM may result in cases which both alleles are present and which would thereby have a stronger association with AD being missed. In addition, the other genes investigated might have dominant effects, meaning that only one copy of the allele needs to be present to result in the full effect of the allele being expressed in the phenotype. *APOE* is known to be dosage-dependent and to not have dominant effects, so other genes that may have dominant effects may show a stronger association with AD than *APOE* does when only one allele is used for analysis.

Relation of SNPs found to previous studies

I attempted to connect the results of this study to previous biomedical studies related to AD. I did so by investigating previous studies that have been conducted on the SNPs my study found to be most predictive of AD. Table 6.1 shows the results of this investigation. I examined each of the top two SNPs with the highest balanced accuracy for each of the four genes across the experiments with both the unbalanced and balanced datasets. In cases which there is a tie in the balanced accuracy, I examined the SNP that appeared first in the data tables in Section 5.1, which are listed in order of lowest to highest position within chromosome 19. I searched each SNP's rsID in the NCBI dbSNP SNP database [National Library of Medicine, n.d.], and investigated the database's results for this SNP. For those SNPs with an rsID that has been merged into another as indicated by NCBI dbSNP, the results from the new merged rsID are shown. In Table 6.1, I have listed the number of studies referencing each SNP as listed in dbSNP and the number of studies that

reference this SNP that relate to AD or dementia. Studies were defined as being related to AD or dementia if their titles contained the words 'Alzheimer's', 'dementia', or 'cognitive impairment' or something similar; no reference to the main text of the articles was made. With further time, all of the SNPs found to have a signal could be examined in this way, rather than only the top two SNPs for each gene.

The results of this search for previous studies relating to my study's most significant SNPs provides some validation of my results, as they are consistent with previous results. The SNP with the best performance on the accuracy metrics, rs11542028 within *APOE*, was referenced 33 times in papers, and 6 times in papers concerning AD or dementia. In addition, SNP rs58185379 within *TOMM40*, a gene which has been shown to be associated with AD [Potkin *et al.*, 2009], was referenced in 8 studies relating to AD or dementia. These results being replicated across multiple studies is supportive of the idea that these SNPs are related to AD. On the other hand, the most significant SNPs found in the *SIGLEC11* and *EXOC3L2* genes were not referenced in any studies at all, as per dbSNP. I would suggest that further research is done into these genes to see if they are associated with AD, given the results of my and my labmate's experiments, both of which suggest that these genes are indeed related to AD.

6.1.2 Comparison of prediction accuracy to previous results

Prediction of AD using ML

Jo *et al.*, 2022 utilized a convolutional neural network (CNN) on AD SNPs from the ADNI dataset, achieving a highest mean accuracy from cross-validation of 75.02% and area under the curve (AUC) of 0.8157. The results from their CNN were comparable with the results of the other machine learning approaches they tried, random forest and XGBoost [Jo *et al.*, 2022]. My transformer and SVM results were not as high as these. Jo *et al.*, 2022's model utilized SNPs

Table 6.1: The number of studies on any topic and on topics relating to AD or dementia that reference each of the SNPs with highest balanced accuracy present, as per the NCBI dbSNP database.

rsID	Gene	Highest balanced accuracy achieved	Downsampled?	# citations	# citations related to AD/dementia
rs11542028	*APOE*	0.65	No	33	6
rs769449	*APOE*	0.58	Yes	24	4
rs58185379	*TOMM40*	0.59	No	14	8
rs34095326	*TOMM40*	0.58	No	1	0
rs117180821	*SIGLEC11*	0.61	Yes	0	0
rs2076155786	*SIGLEC12*	0.61	Yes	0	0
rs60528995	*EXOC3L2*	0.56	Yes	0	0
rs57354345	*EXOC3L3*	0.52	Yes	0	0

from across the genome, testing windows of between 100 to 10,000 SNPs at a time, with the highest accuracy and AUC results being achieved when using a 4000-SNP window. My model used fewer SNPs at a time, from only one gene at a time, which may explain the lower accuracy results reported. I also only looked at four genes on chromosome 19, while Jo *et al.*, 2022's study looked at all genes across the genome, which may also negatively impact my models' performance compared to theirs. In addition, my model only considered one of the two alleles present, while theirs may have considered both, resulting in more signal being present.

Ghose *et al.*, 2022 used a CNN to predict AD using SNP data, but did not report accuracy metrics for the predictive performance of their model, and only reported *p*-values. This makes it difficult to compare my models' performance with theirs, as I did not report *p*-values. However, I did directly correspond with the author to inquire about their accuracy results. They estimated their accuracy of predicting AD to be around 65-66%, which was achieved with the *APOE* gene. My results on the highest *APOE* SNP are comparable to theirs, as my model achieved a balanced accuracy of 65% on the highest-performing *APOE* SNP.

There were some differences between my study and Ghose *et al.*, 2022's, though. Ghose *et al.*, 2022 used the UK Biobank dataset, which is larger than the ADNI dataset that I used. They also used additional variables, such as age, education level, and gender for their classification, while my model only used genetic data. Age, education level, and gender are all environmental variables, whose information is not contained in genetic data, that are correlated with AD. AD onsets later in life, making AD correlated with age. Higher education has been shown to be related to a faster cognitive decline from AD [Bruandet *et al.*, 2008]. Furthermore, according to Andrew and Tierney, 2018, two thirds of AD cases occur in women. While the genetic basis of sex is contained in the X and Y chromosomes (though I did not use these chromosomes in my analysis and only used chromosome 19), purely genetic data lacks information on the environmental impacts of gender, which could include, for example, differential access to higher education based on gender. These additional environmental variables provide more signal to the data that may improve others' accuracy results compared to mine.

Prediction using NNs on DNA sequencing data

Prior results have shown the ability of NN-based models to achieve high accuracy on predictive tasks related to properties of DNA. The DNABERT model achieved accuracy, F1, and MCC scores greater than 0.9 on all tasks related to prediction of proximal and core promoter regions and transcription factor bindings sites within DNA sequences [Ji *et al.*, 2021]. SpliceAI achieved a 95% top-k accuracy in the prediction of splice junctions from pre-mRNA transcripts [Jaganathan *et al.*, 2019]. These results demonstrate the promise of NN approaches in predictions on DNA sequences. The DNABERT transformer model did not achieve as high accuracy results when I applied it to the prediction of AD from WGS data from ADNI. However, the task is different and more difficult that the predictive tasks described in Ji *et al.*, 2021's paper,

since it involves predicting a phenotypic trait that is not entirely determined through genetics, as it is only 70% heritable.

6.1.3 Commentary on models' performance

The best of the SVM results are fairly close to the baseline results, though still not quite as good. On the other hand, the results from the transformers were poor and were worse than my SVM results and results from previous studies. Potential reasons for why Jo *et al.*, 2022's results on the ADNI dataset outperformed my SVM results were discussed in Section 6.1.2.

Within the SVM experiments, the SVMs operating on a single SNP at a time performed better on the best SNPs found than the SVMs operating on multiple SNPs at a time on all metrics except recall and F1 score. The decrease in performance on metrics besides recall and F1 score may be due to the SVMs that operated on multiple SNPs overfitting and not being able to find a separator between all the data points. In addition, the inclusion of additional SNPs to the model that might not have had a signal may have decreased the signal-to-noise ratio. The increase in performance on recall and F1 score on the SVM models that used multiple SNPs may indicate the model's increased ability to not miss positive samples when there is more information present from the multiple SNPs, which is reflected in the increase in recall score and which in turn affects the F1 score, as the F1 score is the harmonic mean of precision and recall. In real-world applications in which maximizing the recall is paramount, such as in cases where the cost of a false negative result is high, running the SVM model on all SNPs at once may be helpful.

For some of the genes, the random downsampling to balance the dataset improved the predictive performance of the SVMs, while for other genes, the random downsampling worsened the predictive performance. I attribute this to differences in the dataset, as the model used remained the same. Perhaps

for some of the genes, having a balanced dataset enabled to model to better distinguish between classes, while for other genes, the presence of more training data was of more use to the model.

Although transformer models are larger and more sophisticated than SVMs, the SVM models outperformed the transformers in this scenario. There are several reasons why this could be the case. One reason is the difference in the data inputted. The SVMs were fed one or more SNPs as input, while the transformers were fed sequences of 100 or more nucleotides at once. As shown in Table 4.1, there were very few variations of the genome sequences between the different patients in the ADNI dataset, with only 0.812% of the nucleotides across all four genes having any differences between patients. Considering these nucleotides as the signal, there was 0.812% of signal and 99.188% of noise in the dataset. The 0.812% of signal could be considered an upper bound on the amount of signal in the dataset, as not all the SNPs present might be related to AD. Some of the SNPs could be related to other traits, or may have no effect on gene expression or phenotype at all. From this viewpoint, the SVM models were fed data with a much higher signal-to-noise ratio than the transformers were, which could help explain the difference in accuracy between the two.

Despite all the extra noise in the datasets taking the whole DNA sequence windows as input, these DNA sequences were still a superset of the information contained in the SNPs, and all the SNPs were indeed present in these windows. One could hypothesize that the transformers, which had access to all the information that the SVMs did, could have matched or exceeded the performance of the SVMs. Even in stretches of DNA where there is no variation between the patients, the nearby or distal DNA context may have an effect on the expression of a particular SNP, so the additional information provided by the full DNA sequence data as opposed to just SNPs could be valuable. The superior performance of the SVMs compared to the transformers may support the claim that SVMs and other traditional machine learning methods

perform better on datasets with less of a signal and fewer data points, and that transformers are data-hungry and require large quantities of data with a high amount of signal. Natural language datasets, which transformers have shown exceptional results on, may have a large amount of signal, but this may not be the case for genomic datasets, in which there may be few variations between patients, lots of junk DNA present, and few available samples due to the relative difficulty of collecting genome sequencing samples from patients versus the massive amount of human-created text available in digital format. Large transformer-based natural language processing models, such as GPT-3, have used datasets many times bigger than what I used. While GPT-3 used the Common Crawl dataset which contains petabytes of data [Brown *et al.*, 2020], the balanced dataset I used contained data from only 576 patients.

Beyond the small size of the ADNI dataset, there are further complications in the quality and amount of signal in the data present. Not all the cases in the ADNI dataset actually had Alzheimer's disease, with many having mild cognitive impairment instead. MCI is a transitional stage between AD and dementia. MCI could progress to dementia, but it does not always do so. MCI could also progress to a variety of dementias besides AD, including vascular and frontotemporal dementia [Campbell *et al.*, 2013]. These other dementias may not have the same genetic basis and mechanisms as Alzheimer's-type dementia. The *APOE* gene is also known as a risk factor for MCI [Campbell *et al.*, 2013], so there is some evidence that MCI and AD are genetically related, but as explained above, MCI and AD are not always the same. As such, the presence of MCI patients within the sample may add additional noise to the task of trying to predict AD from genetic data.

Moreover, there may be issues with the correctness of the labels in the dataset. Those patients labeled as having AD may have been misdiagnosed. The only way to diagnose AD with certainty is to perform a postmortem autopsy of the brain [Gaugler *et al.*, 2013], which I would imagine that few

patients in the ADNI dataset have had, especially if they are still alive. Studies have reported that 12-23% of patients with AD are misdiagnosed, as their brain pathology during autopsy is not consistent with the presence of AD [Gaugler *et al.*, 2013]. As such, it is not certain that all the cases in the dataset that are labeled as AD actually have it, which may further negatively impact the performance of the models.

Another factor that could contribute to poor prediction of AD based on genetic factors is that AD is not purely a genetic disease. While 70% of AD cases are related to genetic factors [Breijyeh and Karaman, 2020], the remaining 30% are not related to genetic factors. If patients whose AD is not due to genetic factors are in the dataset, then the model will be unable to predict these patients' presence or absence of AD solely based on their genome sequence. Indeed, one could imagine that the highest accuracy achievable in predicting AD from genetic data is the heritability of the disease, at 70%.

The difficulty of predicting presence of AD is also discussed by Osipowicz *et al.*, 2021. They assert that the maximum performance of a model for classifying AD based on genotype is an AUC of 0.55-0.7, and that models with a higher AUC are likely to be overfitting. By this measure, my SVM models' performance is acceptable and in line with the expected accuracy.

6.2 Critical evaluation of work

6.2.1 Strengths

Novelty of my approach for NNs on DNA sequences

While other studies have used machine learning for prediction of AD-related traits, few studies have used ML or NNs for prediction of AD based on genome sequencing data. Jo *et al.*, 2022 and Ghose *et al.*, 2022's studies are two such studies that did so, but they used SNP data, rather than WGS or WES data. My approach used ML for prediction of AD using WGS data.

In addition, there have been few studies applying deep learning to genomic data. To my knowledge, my study is the first to attempt to predict a phenotypic trait (presence or absence of AD) from raw genome sequencing data, rather than a genotypic trait (e.g. whether a region of DNA is a promoter).

Finding signal with a simpler model

When my initial approach of transformers did not work on my dataset, I pivoted and tried an alternative, simpler approach of SVMs, which did find some signal. It was good that my SVM models were able to find some signal in the data and that some of the SNPs that the SVMs identified were consistent with previous results in the literature.

6.2.2 Limitations

Size of dataset

The main limitation of this study is the size of the dataset used. With such a small dataset containing only 518 cases and 276 controls, there is not much information available for machine learning models to learn how to predict AD. While some signal was still found with the SVM models, the transformer model was not able to find a signal. With more data, it is possible that the transformer would have achieved better results.

I realized the limitation of the size of the dataset and attempted to pivot from using the ADNI dataset to using the UK Biobank dataset, which has over 500,000 participants [Szustakowski *et al.*, 2021]. I ran the preprocessing steps of indexing the whole exome sequencing CRAM files with the reference genome, converting the CRAM files to binarized BAM files, converting the outputted BAM files to FASTA files, and extracting the appropriate region from the FASTA files. This preprocessing computation was very time-consuming, taking 16 days to run when parallelized across 100 out of the 104 cores present on the lab's computing cluster. Because the preprocessing computation took so long

to run, I did not have time to attempt to re-run my analyses on the UK Biobank data. The preprocessed data and the batch code to run the preprocessing in parallel is now stored on the lab's server, making it easy for other lab members to pick up this analysis if they choose.

Even in the UK Biobank dataset, there are still limitations. Whole genome sequencing data is available for many individuals within the UK Biobank, but my lab only had access to whole exome sequencing (WES) data from the UK Biobank due to the high cost of accessing the WGS data. WES data contains less infomration than WGS data, as many stretches of DNA that are not expressed and are absent from the exome are left blank in the WES data. Despite there being less data in the WES data, having a larger dataset could be helpful. However, there is still not that much data for patients with AD present in the UK Biobank. Within the WES data files from the UK Biobank that my lab has access to, there are only 95 confirmed cases of AD present. This is fewer than the number of WGS samples present in the ADNI dataset, so these files alone would likely not result in an improvement over the ADNI results.

To augment the amount of data for positive AD cases available, my lab and others have used the sequences from people who either have AD or have a family history of AD as cases and those with neither as controls. While this might reduce the amount of signal present since people with only a family history of AD might have less strong of a signal for AD, studies find that the increase in the amount of data available that comes from using family history generally outweighs this drawback (see discussion on GWAX in Section 2.1.2). I investigated the breakdown of cases and controls as per this methodology in the UK Biobank WES dataset. I counted patients with AD or with either a positive maternal or paternal history of AD (or both parents positive) as cases and those with both parents negative as controls, leaving out those with no data. This resulted in a total of 3,558 cases and 19,169 controls.

This amount of data is an improvement over the ADNI datset, but even if the UK Biobank data had been used, there still may not have been enough signal in the data. People with family history of AD may not actually develop the disease, as there is still an environmental component to AD and not everyone with a family history of AD will develop the disease. As such, the UK Biobank dataset still has limitations.

Even with such a large medical data collection initiative as the UK Biobank, there are still challenges in finding enough high-quality data to make predictions for this project. As such, I would argue that collecting sufficient medical data is a challenge for medical AI projects in general.

Validation techniques

If I had more time, I would have run k-fold cross-validation to determine the results. This would help ensure that the results reported hold up over all permutations of the data and are not due to randomness.

I also would have run the model over multiple random seeds and reported the average of the accuracy metrics. I ran the model on only one random seed, as reported, but the results may change over different random seeds, so it might be more thorough to incorporate this into the reporting of results.

It may also be worth investigating some additional genes which have not been shown in previous studies to be associated with AD to serve as controls. With the high number of nucleotides investigated (49,856 across all four genes), some of the signals found in the SNPs may have been due to chance. Experiments attempting the SVM prediction methods used in this study on other genes could help show that these genes are truly related to AD if the same experiments find no signal on genes not known to be related to AD. It may also be helpful to report p-values for the SNPs found to evaluate their significance levels.

6.3 Future work

6.3.1 Experiments related to this project

This study used only one of the two alleles present in the ADNI WGS data due to time constraints. To enhance the dataset fed to the SVM models, I would use both alleles present for the SVM prediction models. I would use a 'two-hot encoding' method to encode the data: each SNP would be represented as 4 variables that can take the values of 0, 1, or 2, depending on how many of each nucleotide A, T, C, and G are present across the two alleles. I would choose this encoding style because the quantity of alleles present could affect the prevalence of AD, as with *APOE* and its known dependence on dosage [Bellenguez *et al.*, 2020].

I would also complete the additional validation steps described in Section 6.2.2.

In addition, data on age and gender could have been added alongside the genomic dataset. AD is correlated with both of these variables, and previous studies have often used these variables in their predictions of AD [Ghose *et al.*, 2022]. The addition of these features may enhance the signal present. These features may not necessarily help in understanding the genetic basis of AD, but they could help in understanding the interactions between genes and the environment and may improve the classification accuracy.

I would also try a random forest model on the data from multiple SNPs (e.g. the experiments described in Table 5.10). A random forest might be able to better handle many different binary variables than SVMs can. It may also be helpful to try other machine learning models, such as XGBoost, CNNs, and regressions on the data used in this study to see if any of the other models yield better results. Jo *et al.*, 2022 achieved slightly better results than I did with random forests, XGBoost, and CNNs, so it might be worth trying to replicate and improve upon their results. I would also try the transformer model on

the multiple SNP experiments. Graph neural networks (GNNs) may also be a promising class of model to use with windows of multiple SNPs. GNNs can be thought of as a generalization of CNNs, and CNNs have already shown good results as per Jo *et al.*, 2022 and Ghose *et al.*, 2022. In addition, the GNNs may be able to represent more data on the relationship between SNPs, such as the distance between them or known gene-gene interactions, which may have bearing on the prediction results.

In addition to the random downsampling used to balance the dataset in this study, I would try random upsampling to see if adding data rather than deleting data results in better predictive performance.

To enhance the transformer models, I would train them on more epochs than just five to give them additional time to potentially converge. Each transformer took between one to two minutes to train over the five epochs I used, so it would be feasible to train the transformer for more epochs and still finish training within a reasonable amount of time. I would also add k-fold cross-validation to the transformer experiments. I would also conduct further hyperparameter tuning on the DNABERT model.

An additional method of validating that the DNABERT transformer is capable of making predictions from DNA sequences is to use a more realistic artificial signal than the one I used. The signal I used was changing the entire sequence to all 'A' nucleotides. While this signal is strong, it is not very realistic. A more realistic artificial signal might come from simulating mutations such as SNPs, insertions, deletions, and larger structural variants. A random selection of these types of known mutations may help with a proof of concept for DNABERT. To make the artificial signal experiments more relevant to AD, it may be helpful to do a literature review on which exact mutations are present in AD patients, and to simulate those within the dataset.

With additional time, I would repeat my analyses on the larger UK Biobank dataset. The bulk of the preprocessing has already been completed, so all

that is needed is to re-run the code from my project on this new dataset. Hopefully, this larger dataset would yield a better signal. I would also perform the extensions described above on the UK Biobank dataset.

6.3.2 Other directions

To increase the amount of high-quality data available for predictions, it may be helpful to try to predict other traits for which more data is available. For example, the trait of height is highly heritable, with a heritability estimated at 80% [Génin, 2020], and there are 499,956 patients whose height is available in the UK Biobank [*Data-Field 50* n.d.]. This is far more than the number of patients with AD present in the Biobank. As such, with a much larger quantity of data available, more of a signal might be able to be found. This could make height a better first trait than AD for NN models applied to unbroken DNA sequencing data to attempt to predict.

One of the contributions of this project was applying NNs on unbroken DNA sequencing data to attempt to predict a phenotypic trait. It seemed that the AD data from the ADNI dataset had very few mutations present, and little variation between patients. It is possible that other diseases which are the result of broader-scale mutations, rather than a small number of SNPs and point mutations, might be a better fit for transformer-based approaches, which take the global context of each input data point into account. For example, some forms of cancer result from chromothripsis, an event in which a chromosome shatters and is repaired with mechanisms that are error-prone [Forment *et al.*, 2012]. Such large-scale mutations may result in a large amount of signal being present in DNA sequencing data, and may be well-suited to a transformer model. The transformer model could potentially interpret the new mutations present in the genome as cancer-causing.

In addition, to address the general problem of the quality and quantity of medical data available, which I encountered in this project, I would suggest that

more efforts are undertaken to collect large amounts of high-quality medical data and make it available to researchers.

7 Conclusion

7.1 Conclusions from study

This study provides a first attempt at using neural networks, specifically the complex, natural-language inspired language models of transformers, to predict the presence of a disease based on unbroken whole genome sequencing data. However, the transformer models were unable to find a signal in predicting Alzheimer's disease from the experiments run on the ADNI dataset. The SVM models were able to find a signal in prediction of AD from many of the SNPs within the four genes examined from the WGS data in the ADNI dataset. Several of the SNPs found were also found to be related to AD in previous studies. One can conclude from this study that the amount of data used was not enough to elicit a signal from the transformer models, and that more data should be used when trying to train a transformer model for prediction of AD or other complex phenotypic traits. While the NN-based approaches may be more powerful, the simpler, SVM-based approaches displayed more effectiveness in predicting the presence or absence of AD in this study. This may support the idea that SVMs are more effective than large, complex models

such as transformers for problems on smaller datasets.

7.2 Broader context

The difficulties encountered in this study with finding large quantities of high-quality data reaffirm the commonly-held belief that one of the great challenges in machine learning today is obtaining appropriate data to use with our models, rather than merely improving the architectures of our models. The challenges faced in this study emphasize the need to collect large amounts of high-quality medical data and make it available to researchers in order to make strides in the field of applying ML to biomedical problems.

While there were no positive results from the deep learning models applied in this study when they were applied to the genomic dataset, the potential for applying deep learning and other machine learning techniques to genomic data remains vast. Natural language models are growing more advanced, especially with the advent of large language models such as GPT-3, PaLM, and DALL-E. Many in the field of AI believe that large language models are the closest approach we currently have to achieving artificial general intelligence. Just as natural language is the language of human knowledge, with high generalizability to a variety of domains, DNA is the language that underpins biological systems. Understanding DNA through the use of deep learning approaches and natural language approaches could unlock the answers to many unsolved problems in the biological domain. I believe the potential for this field is vast, and that more research applying deep learning approaches to genomic data is needed.

Bibliography

Alzheimer's Disease Neuroimaging Initiative (May 2015). *Whole genome sequencing (WGS) data*. URL: https://adni.loni.usc.edu/data-samples/data-types/genetic-data/wgs/.

Alzheimer's Disease Neuroimaging Initiative (Jan. 2016). *ADNI Manuscript Citations*. URL: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Manuscript_Citations.pdf.

Andrew, M. K. and M. C. Tierney (2018). 'The puzzle of sex, gender and Alzheimer's disease: Why are women more often affected than men?' In: *Women's Health* 14, p. 1745506518817995.

Andrews, S. J., B. Fulton-Howard and A. Goate (2020). 'Interpretation of risk loci from genome-wide association studies of Alzheimer's disease'. In: *The Lancet Neurology* 19.4, pp. 326–335.

Avsec, Ž., V. Agarwal, D. Visentin, J. R. Ledsam, A. Grabska-Barwinska, K. R. Taylor, Y. Assael, J. Jumper, P. Kohli and D. R. Kelley (2021). 'Effective gene expression prediction from sequence by integrating long-range interactions'. In: *Nature methods* 18.10, pp. 1196–1203.

Bellenguez, C., B. Grenier-Boley and J.-C. Lambert (2020). 'Genetics of Alzheimer's disease: where we are, and where we are going'. In: *Current Opinion in Neurobiology* 61, pp. 40–48.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Vol. 4. 4. Springer.

Breijyeh, Z. and R. Karaman (2020). 'Comprehensive review on Alzheimer's disease: causes and treatment'. In: *Molecules* 25.24, p. 5789.

Brendel, V. and H. Busse (1984). 'Genome structure described by formal languages'. eng. In: *Nucleic acids research* 12.5, pp. 2561–2568. ISSN: 0305-1048.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei (2020). 'Language Models are Few-Shot Learners'. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin. Vol. 33. Curran Associates, Inc.,

pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Bruandet, A., F. Richard, S. Bombois, C. Maurage, I. Masse, P. Amouyel and F. Pasquier (2008). 'Cognitive decline and survival in Alzheimer's disease according to education level'. In: *Dementia and geriatric cognitive disorders* 25.1, pp. 74–80.

Cacace, R., K. Sleegers and C. Van Broeckhoven (2016). 'Molecular genetics of early-onset Alzheimer's disease revisited'. In: *Alzheimer's & Dementia* 12.6, pp. 733–748. ISSN: 1552-5260. DOI: https://doi.org/10.1016/j.jalz.2016.01.012. URL: https://www.sciencedirect.com/science/article/pii/S1552526016000790.

Campbell, N. L., F. Unverzagt, M. A. LaMantia, B. A. Khan and M. A. Boustani (2013). 'Risk factors for the progression of mild cognitive impairment to dementia'. In: *Clinics in geriatric medicine* 29.4, pp. 873–893.

Data-Field 50 (n.d.). URL: https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=50.

Devlin, J., M. Chang, K. Lee and K. Toutanova (2018). 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. In: *CoRR* abs/1810.04805. arXiv: 1810.04805.

Draelos, R. (Feb. 2019). *Measuring Performance: The Confusion Matrix*. URL: https://scikit-learn.org/stable/modules/model_evaluation.html.

EMBL-EBI (n.d.). *Trait: Alzheimer disease*. URL: https://www.ebi.ac.uk/gwas/efotraits/MONDO_0004975.

Forment, J. V., A. Kaidi and S. P. Jackson (2012). 'Chromothripsis and cancer: causes and consequences of chromosome shattering'. In: *Nature Reviews Cancer* 12.10, pp. 663–670.

Gaugler, J. E., H. Ascher-Svanum, D. L. Roth, T. Fafowora, A. Siderowf and T. G. Beach (2013). 'Characteristics of patients misdiagnosed with Alzheimer's disease and their medication use: an analysis of the NACC-UDS database'. In: *BMC geriatrics* 13.1, pp. 1–10.

Génin, E. (2020). 'Missing heritability of complex diseases: case solved?' In: *Human Genetics* 139.1, pp. 103–113.

Ghose, U., W. Sproviero, L. Winchester, M. Fernandes, D. Newby, B. Ulm, L. Shi, Q. Liu, C. Adams, A. Albukhari, M. Almansouri, H. Choudhry, C. van Duijn and A. Nevado-Holgado (2022). 'Genome wide association neural networks (GWANN) identify novel genes linked to family history of Alzheimer's disease in the UK BioBank'. In: *medRxiv*. DOI: 10.1101/2022.06.10.22276251. eprint: https://www.medrxiv.org/content/early/2022/06/14/2022.06.10.22276251.full.pdf. URL: https://www.medrxiv.org/content/early/2022/06/14/2022.06.10.22276251.

Google Developers (July 2022). *Classification: ROC Curve and AUC*. URL: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

Huang, X., H. Liu, X. Li, L. Guan, J. Li, L. C. A. M. Tellier, H. Yang, J. Wang and J. Zhang (2018). 'Revealing Alzheimer's disease genes spectrum in the whole-genome by machine learning'. In: *BMC neurology* 18.1, pp. 1–8.

Institute, N. C. (n.d.). *The Cancer Genome Atlas Program*. URL: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

Jaganathan, K., S. K. Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz *et al.* (2019). 'Predicting splicing from primary sequence with deep learning'. In: *Cell* 176.3, pp. 535–548.

Jansen, I. E., J. E. Savage, K. Watanabe, J. Bryois, D. M. Williams, S. Steinberg, J. Sealock, I. K. Karlsson, S. Hägg, L. Athanasiu *et al.* (2019). 'Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk'. In: *Nature genetics* 51.3, pp. 404–413.

Ji, Y., Z. Zhou, H. Liu and R. V. Davuluri (2021). 'DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome'. In: *Bioinformatics* 37.15, pp. 2112–2120.

Jo, T., K. Nho, P. Bice, A. J. Saykin and F. T. A. D. N. Initiative (Feb. 2022). 'Deep learning-based identification of genetic variants: application to Alzheimer's disease classification'. In: *Briefings in Bioinformatics* 23.2. bbac022. ISSN: 1477-4054. DOI: 10.1093/bib/bbac022. eprint: https://academic.oup.com/bib/article-pdf/23/2/bbac022/42804951/bbac022.pdf. URL: https://doi.org/10.1093/bib/bbac022.

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žıdek, A. Potapenko *et al.* (2021). 'Highly accurate protein structure prediction with AlphaFold'. In: *Nature* 596.7873, pp. 583–589.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler and D. Haussler (2002). 'The human genome browser at UCSC'. In: *Genome research* 12.6, pp. 996–1006.

Krizhevsky, A., I. Sutskever and G. E. Hinton (2012). 'ImageNet Classification with Deep Convolutional Neural Networks'. In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C. Burges, L. Bottou and K. Weinberger. Vol. 25. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Kunkle, B. W., B. Grenier-Boley, R. Sims, J. C. Bis, V. Damotte, A. C. Naj, A. Boland, M. Vronskaya, S. J. Van Der Lee, A. Amlie-Wolf *et al.* (2019). 'Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing'. In: *Nature genetics* 51.3, pp. 414–430.

Liu, C.-C., T. Kanekiyo, H. Xu and G. Bu (2013). 'Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy'. In: *Nature Reviews Neurology* 9.2, pp. 106–118.

Marioni, R. E., S. E. Harris, Q. Zhang, A. F. McRae, S. P. Hagenaars, W. D. Hill, G. Davies, C. W. Ritchie, C. R. Gale, J. M. Starr *et al.* (2018). 'GWAS on family history of Alzheimer's disease'. In: *Translational psychiatry* 8.1, pp. 1–7.

Matthews, K. A., W. Xu, A. H. Gaglioti, J. B. Holt, J. B. Croft, D. Mack and L. C. McGuire (2019). 'Racial and ethnic estimates of Alzheimer's disease and related dementias in the United States (2015–2060) in adults aged $\geq$ 65 years'. In: *Alzheimer's & Dementia* 15.1, pp. 17–24.

Misra, S., H. Li and J. He (2020). 'Noninvasive fracture characterization based on the classification of sonic wave travel times'. In: *Machine Learning for Subsurface Characterization*, pp. 243–287.

Namba, Y., M. Tomonaga, H. Kawasaki, E. Otomo and K. Ikeda (1991). 'Apolipoprotein E immunoreactivity in cerebral amyloid deposits and neurofibrillary tangles in Alzheimer's disease and kuru plaque amyloid in Creutzfeldt-Jakob disease'. In: *Brain Research* 541.1, pp. 163–166. ISSN: 0006-8993. DOI: https://doi.org/10.1016/0006-8993(91)91092-F. URL: https://www.sciencedirect.com/science/article/pii/000689939191092F.

National Library of Medicine (n.d.). *dbSNP*. URL: https://www.ncbi.nlm.nih.gov/snp/.

Nho, K. (n.d.). *Practical Guideline for Whole Genome Sequencing*. URL: https://warwick.ac.uk/fac/sci/statistics/staff/academic-research/nichols/presentations/ohbm2014/imggen/Nho-ImgGen-WGSeqPractical.pdf.

Nicholls, H. L., C. R. John, D. S. Watson, P. B. Munroe, M. R. Barnes and C. P. Cabrera (2020). 'Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci'. In: *Frontiers in genetics* 11, p. 350.

Novikova, G., M. Kapoor, J. Tcw, E. M. Abud, A. G. Efthymiou, S. X. Chen, H. Cheng, J. F. Fullard, J. Bendl, Y. Liu *et al.* (2021). 'Integration of Alzheimer's disease genetics and myeloid genomics identifies disease risk regulatory elements and genes'. In: *Nature communications* 12.1, pp. 1–14.

Osipowicz, M., B. Wilczynski, M. A. Machnicka and A. D. N. Initiative (2021). 'Careful feature selection is key in classification of Alzheimer's disease patients based on whole-genome sequencing data'. In: *NAR Genomics and Bioinformatics* 3.3, lqab069.

Potkin, S. G., G. Guffanti, A. Lakatos, J. A. Turner, F. Kruggel, J. H. Fallon, A. J. Saykin, A. Orro, S. Lupoli, E. Salvi *et al.* (2009). 'Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease'. In: *PloS one* 4.8, e6501.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly *et al.* (2007). 'PLINK: a tool set for whole-genome

association and population-based linkage analyses'. In: *The American journal of human genetics* 81.3, pp. 559–575.

Ranawana, R. and V. Palade (2005). 'A neural network based multi-classifier system for gene identification in DNA sequences'. In: *Neural Computing & Applications* 14.2, pp. 122–131.

Rodriguez, S., C. Hug, P. Todorov, N. Moret, S. A. Boswell, K. Evans, G. Zhou, N. T. Johnson, B. T. Hyman, P. K. Sorger *et al.* (2021). 'Machine learning identifies candidates for drug repurposing in Alzheimer's disease'. In: *Nature communications* 12.1, pp. 1–13.

Salminen, A. and K. Kaarniranta (2009). 'Siglec receptors and hiding plaques in Alzheimer's disease'. In: *Journal of molecular medicine* 87.7, pp. 697–701.

Saykin, A. J., L. Shen, T. M. Foroud, S. G. Potkin, S. Swaminathan, S. Kim, S. L. Risacher, K. Nho, M. J. Huentelman, D. W. Craig *et al.* (2010). 'Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans'. In: *Alzheimer's & dementia* 6.3, pp. 265–273.

Scikit-learn Developers (n.d.[a]). *1.4. Support Vector Machines*. URL: https://scikit-learn.org/stable/modules/svm.html#.

Scikit-learn Developers (n.d.[b]). *3.3. Metrics and scoring: quantifying the quality of predictions*. URL: https://scikit-learn.org/stable/modules/model_evaluation.html.

Scikit-learn Developers (n.d.[c]). *sklearn.linear_model.SGDClassifier*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html.

Scikit-learn Developers (n.d.[d]). *sklearn.svm.SVC*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC.

Smith, G. E., D. L. Bohac, S. C. Waring, E. Kokmen, E. G. Tangalos, R. J. Ivnik and R. C. Petersen (1998). 'Apolipoprotein E genotype influences cognitive 'phenotype' in patients with Alzheimer's disease but not in healthy control subjects'. In: *Neurology* 50.2, pp. 355–362. ISSN: 0028-3878. DOI: 10.1212/WNL.50.2.355. eprint: https://n.neurology.org/content/50/2/355.full.pdf. URL: https://n.neurology.org/content/50/2/355.

Sundaram, L., H. Gao, S. R. Padigepati, J. F. McRae, Y. Li, J. A. Kosmicki, N. Fritzilas, J. Hakenberg, A. Dutta, J. Shon *et al.* (2018). 'Predicting the clinical impact of human mutation with deep neural networks'. In: *Nature genetics* 50.8, pp. 1161–1170.

Szustakowski, J. D., S. Balasubramanian, E. Kvikstad, S. Khalid, P. G. Bronson, A. Sasson, E. Wong, D. Liu, J. Wade Davis, C. Haefliger *et al.* (2021). 'Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank'. In: *Nature genetics* 53.7, pp. 942–948.

Tam, V., N. Patel, M. Turcotte, Y. Bossé, G. Paré and D. Meyre (2019). 'Benefits and limitations of genome-wide association studies'. In: *Nature Reviews Genetics* 20.8, pp. 467–484.

The Biopython Contributors (n.d.). *Bio.Entrez package*. URL: https://biopython.org/docs/1.75/api/Bio.Entrez.html.

Uffelmann, E., Q. Q. Huang, N. S. Munung, J. De Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen and D. Posthuma (2021). 'Genome-wide association studies'. In: *Nature Reviews Methods Primers* 1.1, pp. 1–21.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin (2017). 'Attention Is All You Need'. In: *CoRR* abs/1706.03762. arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

Venugopalan, J., L. Tong, H. R. Hassanzadeh and M. D. Wang (2021). 'Multimodal deep learning models for early detection of Alzheimer's disease stage'. In: *Scientific reports* 11.1, pp. 1–13.

Weyer, S. W., M. Klevanski, A. Delekate, V. Voikar, D. Aydin, M. Hick, M. Filippov, N. Drost, K. L. Schaller, M. Saar *et al.* (2011). 'APP and APLP2 are essential at PNS and CNS synapses for transmission, spatial learning and LTP'. In: *The EMBO journal* 30.11, pp. 2266–2280.

Supplementary Materials

The first set of tables in this section contains the full results of the SVM models run on all SNPs within the *APOE, TOMM40, SIGLEC14, and EXOC3L2* genes.

The last table contains the full results of the DNABERT experiment run on all 100-bp windows of the *APOE* gene.

AD prediction accuracy results when running SVM on ADNI SNPs from *APOE* gene with linear kernel.

rsID	Location in chr19	Accuracy	Balanced accuracy	Precision	Recall	F1 score	ROC AUC	MCC
rs11542028	45,409,167	0.69	0.65	0.83	0.73	0.77	0.65	0.28
rs1486677963	45,409,283	0.75	0.52	0.75	1	0.86	0.52	0.19
rs769451	45,410,911	0.74	0.52	0.74	0.98	0.85	0.52	0.09
rs9282609	45,409,113	0.74	0.5	0.74	1	0.85	0.5	0
rs769448	45,409,579	0.74	0.5	0.74	1	0.85	0.52	0
rs769449	45,410,002	0.74	0.5	0.74	1	0.85	0.64	0
rs61357706	45,410,273	0.74	0.5	0.74	1	0.85	0.51	0
rs74253333	45,410,444	0.74	0.5	0.74	1	0.85	0.51	0
rs115299243	45,410,548	0.74	0.5	0.74	1	0.85	0.51	0
rs201672011	45,411,064	0.74	0.5	0.74	1	0.85	0.51	0
rs769452	45,411,110	0.26	0.5	0	0	0	0.5	0
rs61228756	45,411,941	0.74	0.5	0.74	1	0.85	0.39	0
rs769455	45,412,040	0.74	0.5	0.74	1	0.85	0.51	0
rs3200542	45,412,079	0.68	0.47	0.73	0.9	0.8	0.47	-0.08

AD prediction accuracy results when running SVM on ADNI SNPs from *APOE* gene with linear kernel, with random downsampling.

rsID	Location in chr19	Accuracy	Balanced accuracy	Precision	Recall	F1 score	ROC AUC	MCC
rs11542028	45,409,167	0.58	0.58	0.56	0.74	0.63	0.58	0.18
rs769449	45,410,002	0.58	0.58	0.67	0.3	0.41	0.58	0.19
rs9282609	45,409,113	0.49	0.5	0.49	1	0.66	0.5	0
rs1486677963	45,409,283	0.49	0.5	0.49	1	0.66	0.5	0
rs769448	45,409,579	0.49	0.5	0.49	1	0.66	0.46	0
rs61357706	45,410,273	0.49	0.5	0.49	1	0.66	0.5	0
rs74253333	45,410,444	0.49	0.5	0.49	1	0.66	0.52	0
rs115299243	45,410,548	0.49	0.5	0.49	1	0.66	0.5	0
rs769451	45,410,911	0.51	0.5	0	0	0	0.5	0
rs201672011	45,411,064	0.49	0.5	0.49	1	0.66	0.5	0
rs769452	45,411,110	0.49	0.5	0.49	1	0.66	0.5	0
rs61228756	45,411,941	0.51	0.5	0	0	0	0.63	0
rs769455	45,412,040	0.49	0.5	0.49	1	0.66	0.5	0
rs3200542	45,412,079	0.44	0.43	0.17	0.04	0.06	0.43	-0.23

AD prediction accuracy results when running SVM on ADNI SNPs from *TOMM40* gene with linear kernel.

rsID	Location in chr19	Accuracy	Balanced accuracy	Precision	Recall	F1 score	ROC AUC	MCC
rs58185379	45,406,673	0.52	0.59	0.82	0.46	0.59	0.59	0.15
rs34095326	45,395,844	0.4	0.58	0.92	0.2	0.33	0.58	0.19
rs11668327	45,398,633	0.7	0.57	0.77	0.85	0.81	0.57	0.15
rs59841965	45,406,798	0.76	0.55	0.76	1	0.86	0.55	0.27
rs71337246	45,397,512	0.56	0.52	0.75	0.61	0.67	0.52	0.03
rs2238680	45,398,264	0.56	0.52	0.75	0.61	0.67	0.52	0.03
rs386539078	45,398,716	0.56	0.52	0.75	0.61	0.67	0.52	0.03
rs140684051	45,399,456	0.74	0.52	0.74	0.98	0.85	0.52	0.09
rs183743534	45,404,866	0.75	0.52	0.75	1	0.86	0.52	0.19
rs59915866	45,397,229	0.72	0.51	0.74	0.97	0.84	0.51	0.03
rs61679753	45,400,747	0.72	0.51	0.74	0.97	0.84	0.51	0.03
rs111784051	45,402,262	0.72	0.51	0.74	0.97	0.84	0.51	0.03
rs17850098	45,394,819	0.74	0.5	0.74	1	0.85	0.5	0
rs185865	45,394,969	0.74	0.5	0.74	1	0.85	0.39	0
rs146539357	45,395,171	0.74	0.5	0.74	1	0.85	0.5	0
rs78245864	45,395,193	0.74	0.5	0.74	1	0.85	0.51	0
rs1969476125	45,395,266	0.74	0.5	0.74	1	0.85	0.6	0
rs141224510	45,395,318	0.26	0.5	0	0	0	0.5	0
rs2075649	45,395,330	0.74	0.5	0.74	1	0.85	0.46	0
rs60321974	45,395,619	0.74	0.5	0.74	1	0.85	0.6	0
rs11556507	45,395,714	0.74	0.5	0.74	1	0.85	0.39	0
rs73936968	45,395,816	0.74	0.5	0.74	1	0.85	0.52	0
rs151285748	45,395,875	0.74	0.5	0.74	1	0.85	0.5	0
rs34404554	45,395,909	0.74	0.5	0.74	1	0.85	0.6	0
rs16979513	45,396,144	0.74	0.5	0.74	1	0.85	0.6	0
rs778934950	45,396,219	0.74	0.5	0.74	1	0.85	0.61	0
rs116040278	45,396,240	0.74	0.5	0.74	1	0.85	0.5	0
rs138280231	45,396,257	0.74	0.5	0.74	1	0.85	0.5	0
rs142608136	45,396,258	0.74	0.5	0.74	1	0.85	0.5	0
rs4803768	45,396,276	0.74	0.5	0.74	1	0.85	0.55	0
rs111884388	45,396,318	0.74	0.5	0.74	1	0.85	0.5	0
rs59007384	45,396,665	0.74	0.5	0.74	1	0.85	0.6	0
rs386539077	45,396,673	0.26	0.5	0	0	0	0.49	0
rs480228	45,396,899	0.54	0.5	0.74	0.58	0.65	0.5	0
rs77301115	45,396,973	0.74	0.5	0.74	1	0.85	0.51	0
rs73936970	45,397,171	0.74	0.5	0.74	1	0.85	0.51	0
rs112849259	45,397,307	0.74	0.5	0.74	1	0.85	0.51	0
rs28480204	45,397,343	0.74	0.5	0.74	1	0.85	0.5	0
rs182472499	45,397,506	0.74	0.5	0.74	1	0.85	0.5	0
rs116881820	45,397,952	0.74	0.5	0.74	1	0.85	0.51	0
rs115676124	45,397,965	0.26	0.5	0	0	0	0.5	0
rs114083252	45,398,014	0.74	0.5	0.74	1	0.85	0.51	0
rs1313349	45,398,206	0.74	0.5	0.74	1	0.85	0.5	0

rs1313347	45,398,319	0.74	0.5	0.74	1	0.85	0.5	0
rs140021317	45,398,360	0.74	0.5	0.74	1	0.85	0.5	0
rs77726367	45,398,457	0.26	0.5	0	0	0	0.5	0
rs79398853	45,398,785	0.74	0.5	0.74	1	0.85	0.51	0
rs1014387798	45,398,817	0.74	0.5	0.74	1	0.85	0.47	0
rs145752462	45,399,001	0.74	0.5	0.74	1	0.85	0.5	0
rs118111371	45,399,023	0.74	0.5	0.74	1	0.85	0.5	0
rs147707133	45,399,165	0.74	0.5	0.74	1	0.85	0.5	0
rs117100783	45,399,186	0.74	0.5	0.74	1	0.85	0.5	0
rs75687619	45,399,344	0.74	0.5	0.74	1	0.85	0.51	0
rs139399286	45,399,356	0.74	0.5	0.74	1	0.85	0.5	0
rs76366838	45,399,896	0.74	0.5	0.74	1	0.85	0.51	0
rs180854461	45,399,922	0.74	0.5	0.74	1	0.85	0.5	0
rs188605845	45,400,113	0.74	0.5	0.74	1	0.85	0.5	0
rs191946858	45,400,486	0.26	0.5	0	0	0	0.5	0
rs114536010	45,400,725	0.74	0.5	0.74	1	0.85	0.51	0
rs283817	45,400,775	0.74	0.5	0.74	1	0.85	0.5	0
rs116874600	45,400,871	0.74	0.5	0.74	1	0.85	0.51	0
rs113886004	45,401,159	0.74	0.5	0.74	1	0.85	0.5	0
rs73052317	45,401,211	0.26	0.5	0	0	0	0.5	0
rs143500700	45,401,270	0.74	0.5	0.74	1	0.85	0.52	0
rs191880358	45,401,392	0.74	0.5	0.74	1	0.85	0.5	0
rs137983845	45,401,579	0.74	0.5	0.74	1	0.85	0.5	0
rs57826936	45,401,666	0.74	0.5	0.74	1	0.85	0.47	0
rs118170342	45,401,868	0.74	0.5	0.74	1	0.85	0.51	0
rs139988932	45,401,918	0.74	0.5	0.74	1	0.85	0.5	0
rs814575	45,402,368	0.26	0.5	0	0	0	0.5	0
rs814574	45,402,470	0.74	0.5	0.74	1	0.85	0.5	0
rs34878901	45,402,477	0.74	0.5	0.74	1	0.85	0.51	0
rs113112231	45,402,516	0.74	0.5	0.74	1	0.85	0.5	0
rs140918487	45,402,546	0.74	0.5	0.74	1	0.85	0.5	0
rs76841546	45,402,589	0.74	0.5	0.74	1	0.85	0.51	0
rs35568738	45,402,718	0.74	0.5	0.74	1	0.85	0.51	0
rs573199	45,403,119	0.74	0.5	0.74	1	0.85	0.5	0
rs116860749	45,403,216	0.74	0.5	0.74	1	0.85	0.51	0
rs1160985	45,403,412	0.74	0.5	0.74	1	0.85	0.5	0
rs77100236	45,403,458	0.74	0.5	0.74	1	0.85	0.51	0
rs56951511	45,403,858	0.74	0.5	0.74	1	0.85	0.5	0
rs1160984	45,403,924	0.74	0.5	0.74	1	0.85	0.51	0
rs34459630	45,404,000	0.74	0.5	0.74	1	0.85	0.5	0
rs59019406	45,404,431	0.74	0.5	0.74	1	0.85	0.5	0
rs117264457	45,404,432	0.26	0.5	0	0	0	0.5	0
rs543763	45,404,579	0.74	0.5	0.74	1	0.85	0.5	0
rs74253332	45,404,691	0.74	0.5	0.74	1	0.85	0.58	0
rs116977783	45,404,721	0.74	0.5	0.74	1	0.85	0.51	0
rs950159943	45,404,857	0.74	0.5	0.74	1	0.85	0.51	0
rs181585594	45,404,883	0.74	0.5	0.74	1	0.85	0.5	0

rs144738835	45,404,926	0.74	0.5	0.74	1	0.85	0.5	0
rs57188354	45,404,972	0.74	0.5	0.74	1	0.85	0.5	0
rs61583573	45,405,062	0.74	0.5	0.74	1	0.85	0.5	0
rs149311267	45,405,113	0.74	0.5	0.74	1	0.85	0.51	0
rs4803769	45,405,521	0.74	0.5	0.74	1	0.85	0.51	0
rs187250392	45,405,552	0.26	0.5	0	0	0	0.5	0
rs117843462	45,405,634	0.74	0.5	0.74	1	0.85	0.51	0
rs191178282	45,405,683	0.26	0.5	0	0	0	0.5	0
rs140853179	45,405,778	0.74	0.5	0.74	1	0.85	0.51	0
rs144737872	45,405,818	0.74	0.5	0.74	1	0.85	0.5	0
rs139361502	45,405,929	0.74	0.5	0.74	1	0.85	0.5	0
rs200337138	45,406,450	0.74	0.5	0.74	1	0.85	0.5	0
rs112328660	45,401,952	0.72	0.49	0.73	0.98	0.84	0.49	-0.07
rs141864196	45,405,499	0.72	0.49	0.73	0.98	0.84	0.49	-0.07
rs56892245	45,405,931	0.72	0.49	0.73	0.98	0.84	0.49	-0.07
rs73052321	45,404,121	0.71	0.48	0.73	0.97	0.83	0.48	-0.1

AD prediction accuracy results when running SVM on ADNI SNPs from *TOMM40* gene with linear kernel, with random downsampling.

rsID	Location in chr19	Accuracy	Balanced accuracy	Precision	Recall	F1 score	ROC AUC	MCC
rs16979513	45,396,144	0.56	0.56	0.64	0.26	0.37	0.56	0.15
rs58185379	45,406,673	0.56	0.56	0.58	0.41	0.48	0.56	0.13
rs74253332	45,404,691	0.55	0.55	0.53	0.74	0.62	0.55	0.11
rs76841546	45,402,589	0.53	0.52	1	0.04	0.07	0.52	0.14
rs183743534	45,404,866	0.51	0.52	0.5	1	0.67	0.52	0.13
rs117843462	45,405,634	0.53	0.52	1	0.04	0.07	0.52	0.14
rs59841965	45,406,798	0.51	0.52	0.5	1	0.67	0.52	0.13
rs1160985	45,403,412	0.51	0.51	0.5	0.67	0.57	0.51	0.03
rs56951511	45,403,858	0.51	0.51	0.5	0.67	0.57	0.51	0.03
rs59019406	45,404,431	0.51	0.51	0.5	0.67	0.57	0.51	0.03
rs57188354	45,404,972	0.51	0.51	0.5	0.67	0.57	0.51	0.03
rs17850098	45,394,819	0.49	0.5	0.49	1	0.66	0.5	0
rs185865	45,394,969	0.49	0.5	0.49	1	0.66	0.46	0
rs146539357	45,395,171	0.49	0.5	0.49	1	0.66	0.5	0
rs78245864	45,395,193	0.51	0.5	0	0	0	0.5	0
rs1969476125	45,395,266	0.49	0.5	0.49	1	0.66	0.43	0
rs141224510	45,395,318	0.49	0.5	0.49	1	0.66	0.5	0
rs2075649	45,395,330	0.49	0.5	0.49	1	0.66	0.46	0
rs60321974	45,395,619	0.49	0.5	0.49	1	0.66	0.44	0
rs11556507	45,395,714	0.49	0.5	0.49	1	0.66	0.46	0
rs73936968	45,395,816	0.49	0.5	0.49	1	0.66	0.52	0
rs34095326	45,395,844	0.49	0.5	0.49	1	0.66	0.56	0
rs151285748	45,395,875	0.49	0.5	0.49	1	0.66	0.5	0
rs34404554	45,395,909	0.49	0.5	0.49	1	0.66	0.44	0
rs778934950	45,396,219	0.49	0.5	0.49	1	0.66	0.46	0
rs116040278	45,396,240	0.49	0.5	0.49	1	0.66	0.5	0
rs138280231	45,396,257	0.49	0.5	0.49	1	0.66	0.5	0
rs142608136	45,396,258	0.49	0.5	0.49	1	0.66	0.5	0
rs4803768	45,396,276	0.49	0.5	0.49	1	0.66	0.5	0
rs111884388	45,396,318	0.49	0.5	0.49	1	0.66	0.5	0
rs59007384	45,396,665	0.49	0.5	0.49	1	0.66	0.44	0
rs386539077	45,396,673	0.49	0.5	0.49	1	0.66	0.52	0
rs480228	45,396,899	0.49	0.5	0.49	1	0.66	0.49	0
rs73936970	45,397,171	0.49	0.5	0.49	1	0.66	0.5	0
rs28480204	45,397,343	0.49	0.5	0.49	1	0.66	0.5	0
rs182472499	45,397,506	0.49	0.5	0.49	1	0.66	0.5	0
rs71337246	45,397,512	0.49	0.5	0.49	1	0.66	0.47	0
rs115676124	45,397,965	0.49	0.5	0.49	1	0.66	0.5	0
rs114083252	45,398,014	0.49	0.5	0.49	1	0.66	0.5	0
rs1313349	45,398,206	0.49	0.5	0.49	1	0.66	0.5	0
rs2238680	45,398,264	0.49	0.5	0.49	1	0.66	0.47	0
rs1313347	45,398,319	0.49	0.5	0.49	1	0.66	0.5	0
rs140021317	45,398,360	0.49	0.5	0.49	1	0.66	0.5	0

rs77726367	45,398,457	0.49	0.5	0.49	1	0.66	0.5	0
rs386539078	45,398,716	0.49	0.5	0.49	1	0.66	0.46	0
rs1014387798	45,398,817	0.49	0.5	0.49	1	0.66	0.56	0
rs145752462	45,399,001	0.49	0.5	0.49	1	0.66	0.5	0
rs118111371	45,399,023	0.49	0.5	0.49	1	0.66	0.5	0
rs147707133	45,399,165	0.49	0.5	0.49	1	0.66	0.5	0
rs117100783	45,399,186	0.49	0.5	0.49	1	0.66	0.5	0
rs139399286	45,399,356	0.49	0.5	0.49	1	0.66	0.5	0
rs140684051	45,399,456	0.51	0.5	0	0	0	0.52	0
rs180854461	45,399,922	0.49	0.5	0.49	1	0.66	0.5	0
rs188605845	45,400,113	0.49	0.5	0.49	1	0.66	0.5	0
rs191946858	45,400,486	0.49	0.5	0.49	1	0.66	0.5	0
rs61679753	45,400,747	0.49	0.5	0.49	1	0.66	0.45	0
rs283817	45,400,775	0.49	0.5	0.49	1	0.66	0.5	0
rs116874600	45,400,871	0.51	0.5	0	0	0	0.5	0
rs113886004	45,401,159	0.49	0.5	0.49	1	0.66	0.5	0
rs73052317	45,401,211	0.51	0.5	0	0	0	0.5	0
rs143500700	45,401,270	0.49	0.5	0.49	1	0.66	0.5	0
rs191880358	45,401,392	0.49	0.5	0.49	1	0.66	0.52	0
rs137983845	45,401,579	0.49	0.5	0.49	1	0.66	0.5	0
rs57826936	45,401,666	0.49	0.5	0.49	1	0.66	0.54	0
rs118170342	45,401,868	0.51	0.5	0	0	0	0.44	0
rs139988932	45,401,918	0.49	0.5	0.49	1	0.66	0.5	0
rs112328660	45,401,952	0.49	0.5	0.49	1	0.66	0.5	0
rs111784051	45,402,262	0.49	0.5	0.49	1	0.66	0.45	0
rs814575	45,402,368	0.49	0.5	0.49	1	0.66	0.5	0
rs814574	45,402,470	0.49	0.5	0.49	1	0.66	0.5	0
rs113112231	45,402,516	0.49	0.5	0.49	1	0.66	0.5	0
rs140918487	45,402,546	0.49	0.5	0.49	1	0.66	0.5	0
rs35568738	45,402,718	0.49	0.5	0.49	0.96	0.65	0.5	0
rs573199	45,403,119	0.49	0.5	0.49	1	0.66	0.5	0
rs116860749	45,403,216	0.49	0.5	0.49	1	0.66	0.46	0
rs77100236	45,403,458	0.49	0.5	0.49	1	0.66	0.5	0
rs1160984	45,403,924	0.49	0.5	0.49	0.96	0.65	0.5	0
rs34459630	45,404,000	0.49	0.5	0.49	1	0.66	0.5	0
rs73052321	45,404,121	0.51	0.5	0	0	0	0.48	0
rs117264457	45,404,432	0.49	0.5	0.49	1	0.66	0.5	0
rs543763	45,404,579	0.49	0.5	0.49	1	0.66	0.5	0
rs116977783	45,404,721	0.51	0.5	0	0	0	0.5	0
rs950159943	45,404,857	0.49	0.5	0.49	1	0.66	0.46	0
rs181585594	45,404,883	0.49	0.5	0.49	1	0.66	0.5	0
rs144738835	45,404,926	0.49	0.5	0.49	1	0.66	0.5	0
rs61583573	45,405,062	0.49	0.5	0.49	1	0.66	0.51	0
rs149311267	45,405,113	0.51	0.5	0	0	0	0.5	0
rs141864196	45,405,499	0.51	0.5	0	0	0	0.52	0
rs4803769	45,405,521	0.49	0.5	0.49	1	0.66	0.46	0
rs187250392	45,405,552	0.49	0.5	0.49	1	0.66	0.5	0

rs191178282	45,405,683	0.49	0.5	0.49	1	0.66	0.5	0
rs140853179	45,405,778	0.49	0.5	0.49	1	0.66	0.5	0
rs144737872	45,405,818	0.49	0.5	0.49	1	0.66	0.5	0
rs139361502	45,405,929	0.51	0.5	0	0	0	0.5	0
rs56892245	45,405,931	0.49	0.5	0.49	1	0.66	0.5	0
rs200337138	45,406,450	0.49	0.5	0.49	1	0.66	0.5	0
rs77301115	45,396,973	0.45	0.46	0.47	0.93	0.62	0.46	-0.2
rs112849259	45,397,307	0.45	0.46	0.47	0.93	0.62	0.46	-0.2
rs116881820	45,397,952	0.45	0.46	0.47	0.93	0.62	0.46	-0.2
rs79398853	45,398,785	0.45	0.46	0.47	0.93	0.62	0.46	-0.2
rs75687619	45,399,344	0.45	0.46	0.47	0.93	0.62	0.46	-0.2
rs76366838	45,399,896	0.45	0.46	0.47	0.93	0.62	0.46	-0.2
rs114536010	45,400,725	0.45	0.46	0.47	0.93	0.62	0.46	-0.2
rs34878901	45,402,477	0.45	0.46	0.46	0.67	0.55	0.46	-0.09
rs59915866	45,397,229	0.45	0.45	0	0	0	0.45	-0.24
rs11668327	45,398,633	0.45	0.45	0.33	0.11	0.17	0.45	-0.14

AD prediction accuracy results when running SVM on ADNI SNPs from *SIGLEC11* gene with linear kernel.

rsID	Location in chr19	Accuracy	Balanced accuracy	Precision	Recall	F1 score	ROC AUC	MCC
rs61467868	50,456,770	0.49	0.53	0.76	0.44	0.56	0.53	0.05
rs57860877	50,457,346	0.49	0.53	0.76	0.44	0.56	0.53	0.05
rs73932071	50,459,806	0.28	0.51	1	0.02	0.03	0.51	0.07
rs4802641	50,452,341	0.74	0.5	0.74	1	0.85	0.5	0
rs114819375	50,452,606	0.74	0.5	0.74	1	0.85	0.5	0
rs143688215	50,452,641	0.74	0.5	0.74	1	0.85	0.5	0
rs138574928	50,452,960	0.74	0.5	0.74	1	0.85	0.5	0
rs200448773	50,453,203	0.74	0.5	0.74	1	0.85	0.5	0
rs56579996	50,453,317	0.74	0.5	0.74	1	0.85	0.43	0
rs201942673	50,453,351	0.74	0.5	0.74	1	0.85	0.5	0
rs73932016	50,453,889	0.74	0.5	0.74	1	0.85	0.51	0
rs61141600	50,453,957	0.74	0.5	0.74	1	0.85	0.51	0
rs62126307	50,454,086	0.74	0.5	0.74	1	0.85	0.46	0
rs57663431	50,454,348	0.74	0.5	0.74	1	0.85	0.5	0
rs117180821	50,454,375	0.26	0.5	0	0	0	0.52	0
rs2076155786	50,454,383	0.26	0.5	0	0	0	0.52	0
rs111516788	50,454,489	0.26	0.5	0	0	0	0.5	0
rs182406893	50,454,529	0.74	0.5	0.74	1	0.85	0.5	0
rs73932017	50,454,723	0.26	0.5	0	0	0	0.51	0
rs73576644	50,454,801	0.74	0.5	0.74	1	0.85	0.5	0
rs7247753	50,454,973	0.74	0.5	0.74	1	0.85	0.5	0
rs45438992	50,455,180	0.74	0.5	0.74	1	0.85	0.5	0
rs112796514	50,455,202	0.26	0.5	0	0	0	0.5	0
rs10405621	50,455,351	0.74	0.5	0.74	1	0.85	0.46	0
rs188736707	50,455,676	0.26	0.5	0	0	0	0.5	0
rs140029606	50,455,837	0.74	0.5	0.74	1	0.85	0.5	0
rs1354287441	50,455,963	0.74	0.5	0.74	1	0.85	0.5	0
rs151333014	50,456,083	0.74	0.5	0.74	1	0.85	0.5	0
rs115420067	50,456,219	0.26	0.5	0	0	0	0.5	0
rs77312057	50,456,299	0.74	0.5	0.74	1	0.85	0.5	0
rs57719743	50,456,518	0.74	0.5	0.74	1	0.85	0.53	0
rs60765786	50,456,602	0.74	0.5	0.74	1	0.85	0.5	0
rs192228506	50,456,605	0.74	0.5	0.74	1	0.85	0.5	0
rs150046182	50,456,731	0.74	0.5	0.74	1	0.85	0.5	0
rs56666020	50,456,821	0.74	0.5	0.74	1	0.85	0.53	0
rs192269876	50,456,931	0.74	0.5	0.74	1	0.85	0.5	0
rs56139473	50,457,187	0.74	0.5	0.74	1	0.85	0.51	0
rs74605432	50,457,249	0.74	0.5	0.74	1	0.85	0.49	0
rs190335185	50,457,331	0.74	0.5	0.74	1	0.85	0.5	0
rs58768527	50,457,394	0.74	0.5	0.74	1	0.85	0.51	0
rs77695494	50,457,489	0.74	0.5	0.74	1	0.85	0.5	0
rs58603144	50,457,551	0.74	0.5	0.74	1	0.85	0.5	0
rs147552723	50,457,768	0.74	0.5	0.74	1	0.85	0.5	0

rs117126572	50,457,876	0.26	0.5	0	0	0	0.52	0
rs76691680	50,457,915	0.26	0.5	0	0	0	0.52	0
rs117971487	50,457,927	0.26	0.5	0	0	0	0.52	0
rs73576651	50,458,360	0.74	0.5	0.74	1	0.85	0.5	0
rs117428283	50,458,411	0.26	0.5	0	0	0	0.52	0
rs181578199	50,458,542	0.74	0.5	0.74	1	0.85	0.48	0
rs186318895	50,458,546	0.74	0.5	0.74	1	0.85	0.5	0
rs150362483	50,458,743	0.74	0.5	0.74	1	0.85	0.5	0
rs77695568	50,458,850	0.26	0.5	0	0	0	0.52	0
rs8103552	50,458,931	0.74	0.5	0.74	1	0.85	0.46	0
rs112178012	50,458,932	0.74	0.5	0.74	1	0.85	0.5	0
rs149359449	50,459,048	0.74	0.5	0.74	1	0.85	0.5	0
rs283514	50,459,231	0.26	0.5	0	0	0	0.52	0
rs186219825	50,459,745	0.74	0.5	0.74	1	0.85	0.5	0
rs77620599	50,460,002	0.74	0.5	0.74	1	0.85	0.5	0
rs57566803	50,460,914	0.74	0.5	0.74	1	0.85	0.5	0
rs113440966	50,460,959	0.74	0.5	0.74	1	0.85	0.5	0
rs148890842	50,460,985	0.74	0.5	0.74	1	0.85	0.5	0
rs184846975	50,461,021	0.74	0.5	0.74	1	0.85	0.51	0
rs143608588	50,461,117	0.74	0.5	0.74	1	0.85	0.5	0
rs10423405	50,461,207	0.74	0.5	0.74	1	0.85	0.5	0
rs184556939	50,461,224	0.74	0.5	0.74	1	0.85	0.5	0
rs61652084	50,461,447	0.74	0.5	0.74	1	0.85	0.5	0
rs7259039	50,461,732	0.74	0.5	0.74	1	0.85	0.5	0
rs57682421	50,461,735	0.74	0.5	0.74	1	0.85	0.5	0
rs148610893	50,461,799	0.74	0.5	0.74	1	0.85	0.5	0
rs1865042	50,461,893	0.74	0.5	0.74	1	0.85	0.51	0
rs1811375	50,462,244	0.74	0.5	0.74	1	0.85	0.54	0
rs62113133	50,462,298	0.74	0.5	0.74	1	0.85	0.46	0
rs200466360	50,462,889	0.74	0.5	0.74	1	0.85	0.48	0
rs202060100	50,463,026	0.74	0.5	0.74	1	0.85	0.52	0
rs201740168	50,463,030	0.74	0.5	0.74	1	0.85	0.48	0
rs200380048	50,463,034	0.74	0.5	0.74	1	0.85	0.52	0
rs144427989	50,463,661	0.74	0.5	0.74	1	0.85	0.5	0
rs79085301	50,463,670	0.74	0.5	0.74	1	0.85	0.5	0
rs80184294	50,463,727	0.74	0.5	0.74	1	0.85	0.5	0
rs149855251	50,463,937	0.74	0.5	0.74	1	0.85	0.5	0
rs78673790	50,463,982	0.26	0.5	0	0	0	0.57	0
rs80021244	50,464,023	0.74	0.5	0.74	1	0.85	0.5	0
rs148651187	50,464,040	0.74	0.5	0.74	1	0.85	0.51	0
rs13343845	50,459,226	0.29	0.49	0.67	0.07	0.12	0.49	-0.05
rs79972908	50,458,488	0.28	0.48	0.6	0.05	0.09	0.48	-0.08
rs199756638	50,463,044	0.26	0.47	0.5	0.03	0.06	0.47	-0.12

AD prediction accuracy results when running SVM on ADNI SNPs from *SIGLEC11* gene with linear kernel, with random downsampling.

rsID	Location in chr19	Accuracy	Balanced accuracy	Precision	Recall	F1 score	ROC AUC	MCC
rs117180821	50,454,375	0.62	0.61	1	0.22	0.36	0.61	0.36
rs2076155786	50,454,383	0.62	0.61	1	0.22	0.36	0.61	0.36
rs117126572	50,457,876	0.62	0.61	1	0.22	0.36	0.61	0.36
rs76691680	50,457,915	0.62	0.61	1	0.22	0.36	0.61	0.36
rs117971487	50,457,927	0.62	0.61	1	0.22	0.36	0.61	0.36
rs117428283	50,458,411	0.62	0.61	1	0.22	0.36	0.61	0.36
rs79972908	50,458,488	0.62	0.61	1	0.22	0.36	0.61	0.36
rs56579996	50,453,317	0.56	0.57	0.53	0.89	0.67	0.57	0.18
rs62126307	50,454,086	0.56	0.57	0.53	0.89	0.67	0.57	0.18
rs10405621	50,455,351	0.56	0.57	0.53	0.89	0.67	0.57	0.18
rs8103552	50,458,931	0.56	0.57	0.53	0.89	0.67	0.57	0.18
rs79085301	50,463,670	0.53	0.53	0.51	0.85	0.64	0.53	0.09
rs80184294	50,463,727	0.53	0.53	0.51	0.85	0.64	0.53	0.09
rs4802641	50,452,341	0.49	0.5	0.49	1	0.66	0.5	0
rs114819375	50,452,606	0.49	0.5	0.49	1	0.66	0.5	0
rs143688215	50,452,641	0.49	0.5	0.49	1	0.66	0.5	0
rs138574928	50,452,960	0.49	0.5	0.49	1	0.66	0.5	0
rs200448773	50,453,203	0.49	0.5	0.49	1	0.66	0.5	0
rs201942673	50,453,351	0.49	0.5	0.49	1	0.66	0.5	0
rs73932016	50,453,889	0.49	0.5	0.49	1	0.66	0.5	0
rs61141600	50,453,957	0.49	0.5	0.49	1	0.66	0.5	0
rs57663431	50,454,348	0.49	0.5	0.49	1	0.66	0.5	0
rs111516788	50,454,489	0.49	0.5	0.49	1	0.66	0.5	0
rs182406893	50,454,529	0.49	0.5	0.49	1	0.66	0.52	0
rs73932017	50,454,723	0.51	0.5	0	0	0	0.5	0
rs73576644	50,454,801	0.49	0.5	0.49	1	0.66	0.52	0
rs7247753	50,454,973	0.49	0.5	0.49	1	0.66	0.5	0
rs45438992	50,455,180	0.49	0.5	0.49	1	0.66	0.5	0
rs188736707	50,455,676	0.49	0.5	0.49	1	0.66	0.48	0
rs140029606	50,455,837	0.49	0.5	0.49	1	0.66	0.5	0
rs1354287441	50,455,963	0.49	0.5	0.49	1	0.66	0.5	0
rs151333014	50,456,083	0.49	0.5	0.49	1	0.66	0.5	0
rs115420067	50,456,219	0.49	0.5	0.49	1	0.66	0.5	0
rs77312057	50,456,299	0.49	0.5	0.49	1	0.66	0.5	0
rs57719743	50,456,518	0.51	0.5	0	0	0	0.49	0
rs60765786	50,456,602	0.49	0.5	0.49	1	0.66	0.52	0
rs192228506	50,456,605	0.49	0.5	0.49	1	0.66	0.5	0
rs150046182	50,456,731	0.51	0.5	0	0	0	0.5	0
rs61467868	50,456,770	0.51	0.5	0	0	0	0.49	0
rs56666020	50,456,821	0.51	0.5	0	0	0	0.49	0
rs192269876	50,456,931	0.49	0.5	0.49	1	0.66	0.5	0
rs56139473	50,457,187	0.49	0.5	0.49	1	0.66	0.5	0

rs190335185	50,457,331	0.49	0.5	0.49	1	0.66	0.5	0
rs57860877	50,457,346	0.51	0.5	0	0	0	0.49	0
rs58768527	50,457,394	0.49	0.5	0.49	1	0.66	0.5	0
rs77695494	50,457,489	0.49	0.5	0.49	1	0.66	0.5	0
rs58603144	50,457,551	0.49	0.5	0.49	1	0.66	0.46	0
rs147552723	50,457,768	0.49	0.5	0.49	1	0.66	0.5	0
rs73576651	50,458,360	0.49	0.5	0.49	1	0.66	0.52	0
rs181578199	50,458,542	0.51	0.5	0	0	0	0.5	0
rs186318895	50,458,546	0.49	0.5	0.49	1	0.66	0.5	0
rs150362483	50,458,743	0.49	0.5	0.49	1	0.66	0.5	0
rs77695568	50,458,850	0.49	0.5	0.49	1	0.66	0.39	0
rs112178012	50,458,932	0.49	0.5	0.49	1	0.66	0.5	0
rs149359449	50,459,048	0.49	0.5	0.49	1	0.66	0.5	0
rs13343845	50,459,226	0.49	0.5	0.49	1	0.66	0.59	0
rs283514	50,459,231	0.49	0.5	0.49	1	0.66	0.5	0
rs186219825	50,459,745	0.49	0.5	0.49	1	0.66	0.5	0
rs73932071	50,459,806	0.51	0.5	0	0	0	0.5	0
rs77620599	50,460,002	0.49	0.5	0.49	1	0.66	0.52	0
rs57566803	50,460,914	0.49	0.5	0.49	1	0.66	0.48	0
rs113440966	50,460,959	0.49	0.5	0.49	1	0.66	0.5	0
rs148890842	50,460,985	0.49	0.5	0.49	1	0.66	0.5	0
rs184846975	50,461,021	0.51	0.5	0	0	0	0.5	0
rs143608588	50,461,117	0.51	0.5	0	0	0	0.5	0
rs10423405	50,461,207	0.49	0.5	0.49	1	0.66	0.5	0
rs61652084	50,461,447	0.49	0.5	0.49	1	0.66	0.5	0
rs7259039	50,461,732	0.49	0.5	0.49	1	0.66	0.5	0
rs57682421	50,461,735	0.49	0.5	0.49	1	0.66	0.5	0
rs148610893	50,461,799	0.49	0.5	0.49	1	0.66	0.52	0
rs1865042	50,461,893	0.51	0.5	0	0	0	0.5	0
rs1811375	50,462,244	0.49	0.5	0.49	1	0.66	0.49	0
rs62113133	50,462,298	0.51	0.5	0	0	0	0.57	0
rs200466360	50,462,889	0.51	0.5	0	0	0	0.52	0
rs201740168	50,463,030	0.49	0.5	0.49	1	0.66	0.48	0
rs200380048	50,463,034	0.49	0.5	0.49	1	0.66	0.54	0
rs199756638	50,463,044	0.51	0.5	0	0	0	0.46	0
rs144427989	50,463,661	0.49	0.5	0.49	1	0.66	0.5	0
rs149855251	50,463,937	0.51	0.5	0	0	0	0.48	0
rs78673790	50,463,982	0.51	0.5	0	0	0	0.57	0
rs80021244	50,464,023	0.49	0.5	0.49	1	0.66	0.5	0
rs148651187	50,464,040	0.51	0.5	0	0	0	0.5	0
rs112796514	50,455,202	0.47	0.48	0.48	0.96	0.64	0.48	-0.14
rs74605432	50,457,249	0.47	0.48	0.48	0.96	0.64	0.48	-0.14
rs184556939	50,461,224	0.47	0.48	0.48	0.96	0.64	0.48	-0.14
rs202060100	50,463,026	0.42	0.43	0.45	0.85	0.59	0.43	-0.29

AD prediction accuracy results when running SVM on ADNI SNPs from *EXOC3L2* gene with linear kernel.

rsID	Location in chr19	Accuracy	Balanced accuracy	Precision	Recall	F1 score	ROC AUC	MCC
rs59647713	45,736,003	0.62	0.55	0.76	0.71	0.74	0.55	0.09
rs112759099	45,726,845	0.32	0.54	1	0.08	0.16	0.54	0.15
rs28645301	45,724,692	0.31	0.53	1	0.07	0.13	0.53	0.14
rs28564302	45,724,868	0.31	0.53	1	0.07	0.13	0.53	0.14
rs386809738	45,724,961	0.31	0.53	1	0.07	0.13	0.53	0.14
rs60269219	45,724,963	0.31	0.53	1	0.07	0.13	0.53	0.14
rs59356929	45,725,127	0.31	0.53	1	0.07	0.13	0.53	0.14
rs58213824	45,725,185	0.31	0.53	1	0.07	0.13	0.53	0.14
rs57294488	45,725,481	0.31	0.53	1	0.07	0.13	0.53	0.14
rs73568222	45,725,975	0.31	0.53	1	0.07	0.13	0.53	0.14
rs57399322	45,726,106	0.31	0.53	1	0.07	0.13	0.53	0.14
rs10423031	45,726,224	0.31	0.53	1	0.07	0.13	0.53	0.14
rs58715307	45,726,458	0.31	0.53	1	0.07	0.13	0.53	0.14
rs10423753	45,726,563	0.31	0.53	1	0.07	0.13	0.53	0.14
rs113728460	45,726,654	0.31	0.53	1	0.07	0.13	0.53	0.14
rs142415915	45,726,745	0.31	0.53	1	0.07	0.13	0.53	0.14
rs146871722	45,726,758	0.31	0.53	1	0.07	0.13	0.53	0.14
rs113045530	45,726,821	0.31	0.53	1	0.07	0.13	0.53	0.14
rs143154520	45,726,869	0.31	0.53	1	0.07	0.13	0.53	0.14
rs111691933	45,726,964	0.31	0.53	1	0.07	0.13	0.53	0.14
rs112909419	45,726,968	0.31	0.53	1	0.07	0.13	0.53	0.14
rs112668741	45,726,976	0.31	0.53	1	0.07	0.13	0.53	0.14
rs111462669	45,727,167	0.31	0.53	1	0.07	0.13	0.53	0.14
rs1969927370	45,727,275	0.31	0.53	1	0.07	0.13	0.53	0.14
rs117316672	45,727,276	0.31	0.53	1	0.07	0.13	0.53	0.14
rs60048477	45,727,362	0.31	0.53	1	0.07	0.13	0.53	0.14
rs10405194	45,727,622	0.31	0.53	1	0.07	0.13	0.53	0.14
rs1387808030	45,727,671	0.31	0.53	1	0.07	0.13	0.53	0.14
rs60406788	45,727,930	0.31	0.53	1	0.07	0.13	0.53	0.14
rs59172754	45,728,059	0.31	0.53	1	0.07	0.13	0.53	0.14
rs57767166	45,728,123	0.31	0.53	1	0.07	0.13	0.53	0.14
rs58647388	45,728,231	0.31	0.53	1	0.07	0.13	0.53	0.14
rs60081440	45,728,238	0.31	0.53	1	0.07	0.13	0.53	0.14
rs57787576	45,728,406	0.31	0.53	1	0.07	0.13	0.53	0.14
rs58846289	45,728,576	0.31	0.53	1	0.07	0.13	0.53	0.14
rs61625909	45,728,595	0.31	0.53	1	0.07	0.13	0.53	0.14
rs57002525	45,728,695	0.31	0.53	1	0.07	0.13	0.53	0.14
rs59741163	45,728,806	0.31	0.53	1	0.07	0.13	0.53	0.14
rs57403313	45,728,942	0.31	0.53	1	0.07	0.13	0.53	0.14
rs56675703	45,729,123	0.31	0.53	1	0.07	0.13	0.53	0.14
rs61552519	45,729,200	0.31	0.53	1	0.07	0.13	0.53	0.14
rs57035271	45,729,275	0.31	0.53	1	0.07	0.13	0.53	0.14
rs56879892	45,729,587	0.31	0.53	1	0.07	0.13	0.53	0.14

rs111269631	45,729,813	0.31	0.53	1	0.07	0.13	0.53	0.14
rs10424245	45,729,948	0.31	0.53	1	0.07	0.13	0.53	0.14
rs112405270	45,730,238	0.31	0.53	1	0.07	0.13	0.53	0.14
rs61463967	45,730,470	0.31	0.53	1	0.07	0.13	0.53	0.14
rs188601304	45,722,334	0.75	0.52	0.75	1	0.86	0.52	0.19
rs186531134	45,726,149	0.75	0.52	0.75	1	0.86	0.52	0.19
rs145952987	45,729,924	0.75	0.52	0.75	1	0.86	0.52	0.19
rs62118504	45,734,751	0.59	0.52	0.75	0.66	0.7	0.52	0.04
rs59259486	45,719,790	0.7	0.51	0.74	0.92	0.82	0.51	0.02
rs57045381	45,731,564	0.3	0.51	0.8	0.07	0.12	0.51	0.04
rs140013593	45,732,661	0.3	0.51	0.8	0.07	0.12	0.51	0.04
rs58258155	45,732,725	0.3	0.51	0.8	0.07	0.12	0.51	0.04
rs181222539	45,732,960	0.28	0.51	1	0.02	0.03	0.51	0.07
rs1970013094	45,733,214	0.3	0.51	0.8	0.07	0.12	0.51	0.04
rs189049349	45,733,309	0.28	0.51	1	0.02	0.03	0.51	0.07
rs10402739	45,733,897	0.3	0.51	0.8	0.07	0.12	0.51	0.04
rs1970025005	45,734,195	0.3	0.51	0.8	0.07	0.12	0.51	0.04
rs77003151	45,715,976	0.74	0.5	0.74	1	0.85	0.58	0
rs143393432	45,715,996	0.74	0.5	0.74	1	0.85	0.51	0
rs11667509	45,716,192	0.74	0.5	0.74	1	0.85	0.59	0
rs11667430	45,716,197	0.74	0.5	0.74	1	0.85	0.58	0
rs185860487	45,716,283	0.74	0.5	0.74	1	0.85	0.5	0
rs200836902	45,716,357	0.74	0.5	0.74	1	0.85	0.5	0
rs189063316	45,716,364	0.74	0.5	0.74	1	0.85	0.52	0
rs57437338	45,716,678	0.74	0.5	0.74	1	0.85	0.45	0
rs73034885	45,716,899	0.26	0.5	0	0	0	0.51	0
rs1969806189	45,717,169	0.74	0.5	0.74	1	0.85	0.5	0
rs151046811	45,717,248	0.74	0.5	0.74	1	0.85	0.52	0
rs142428371	45,717,296	0.74	0.5	0.74	1	0.85	0.52	0
rs1287006835	45,717,427	0.74	0.5	0.74	1	0.85	0.5	0
rs57354345	45,717,615	0.74	0.5	0.74	1	0.85	0.52	0
rs187175824	45,717,625	0.74	0.5	0.74	1	0.85	0.5	0
rs182773180	45,717,719	0.74	0.5	0.74	1	0.85	0.5	0
rs144670664	45,717,785	0.74	0.5	0.74	1	0.85	0.52	0
rs187458474	45,717,943	0.74	0.5	0.74	1	0.85	0.52	0
rs189424543	45,718,433	0.74	0.5	0.74	1	0.85	0.5	0
rs57112300	45,718,624	0.29	0.5	0.75	0.05	0.1	0.5	0.01
rs1305380718	45,718,720	0.74	0.5	0.74	1	0.85	0.5	0
rs188515109	45,718,852	0.74	0.5	0.74	1	0.85	0.5	0
rs59566001	45,719,065	0.26	0.5	0	0	0	0.54	0
rs73568206	45,719,106	0.26	0.5	0	0	0	0.5	0
rs115906914	45,719,138	0.74	0.5	0.74	1	0.85	0.5	0
rs141724506	45,719,193	0.74	0.5	0.74	1	0.85	0.5	0
rs56916276	45,719,426	0.74	0.5	0.74	1	0.85	0.51	0
rs73568208	45,719,463	0.26	0.5	0	0	0	0.5	0
rs76738189	45,719,493	0.74	0.5	0.74	1	0.85	0.52	0
rs149229776	45,720,307	0.74	0.5	0.74	1	0.85	0.5	0

rs1160873375	45,720,764	0.74	0.5	0.74	1	0.85	0.51	0
rs58935155	45,721,596	0.74	0.5	0.74	1	0.85	0.5	0
rs79531457	45,722,020	0.74	0.5	0.74	1	0.85	0.5	0
rs73568210	45,722,077	0.26	0.5	0	0	0	0.5	0
rs114505591	45,722,186	0.74	0.5	0.74	1	0.85	0.5	0
rs117755077	45,722,230	0.29	0.5	0.75	0.05	0.1	0.5	0.01
rs111406553	45,722,265	0.26	0.5	0	0	0	0.5	0
rs60528995	45,722,517	0.74	0.5	0.74	1	0.85	0.53	0
rs74459616	45,722,529	0.74	0.5	0.74	1	0.85	0.51	0
rs144640143	45,722,616	0.74	0.5	0.74	1	0.85	0.52	0
rs59333887	45,722,743	0.74	0.5	0.74	1	0.85	0.53	0
rs73568215	45,722,773	0.26	0.5	0	0	0	0.5	0
rs192547359	45,723,142	0.74	0.5	0.74	1	0.85	0.5	0
rs183665617	45,723,228	0.74	0.5	0.74	1	0.85	0.5	0
rs111664793	45,723,233	0.26	0.5	0	0	0	0.5	0
rs1969881112	45,723,376	0.74	0.5	0.74	1	0.85	0.5	0
rs149548393	45,723,380	0.74	0.5	0.74	1	0.85	0.51	0
rs79890446	45,723,446	0.74	0.5	0.74	1	0.85	0.49	0
rs140668794	45,723,450	0.74	0.5	0.74	1	0.85	0.5	0
rs147898480	45,723,518	0.74	0.5	0.74	1	0.85	0.5	0
rs12461144	45,723,706	0.74	0.5	0.74	1	0.85	0.45	0
rs112102023	45,723,714	0.26	0.5	0	0	0	0.48	0
rs78222968	45,723,832	0.74	0.5	0.74	1	0.85	0.5	0
rs10406604	45,723,986	0.26	0.5	0	0	0	0.53	0
rs73034893	45,724,044	0.74	0.5	0.74	1	0.85	0.51	0
rs141046425	45,724,110	0.74	0.5	0.74	1	0.85	0.5	0
rs1194606521	45,724,296	0.26	0.5	0	0	0	0.53	0
rs1555756029	45,724,297	0.26	0.5	0	0	0	0.53	0
rs1969894901	45,724,561	0.26	0.5	0	0	0	0.54	0
rs1426173634	45,724,633	0.26	0.5	0	0	0	0.54	0
rs12978617	45,724,658	0.26	0.5	0	0	0	0.54	0
rs184287728	45,724,732	0.74	0.5	0.74	1	0.85	0.5	0
rs28607628	45,724,743	0.74	0.5	0.74	1	0.85	0.51	0
rs144619413	45,725,081	0.74	0.5	0.74	1	0.85	0.52	0
rs181890181	45,725,109	0.74	0.5	0.74	1	0.85	0.49	0
rs141681064	45,725,149	0.74	0.5	0.74	1	0.85	0.52	0
rs185288032	45,725,199	0.26	0.5	0	0	0	0.5	0
rs80074203	45,725,247	0.74	0.5	0.74	1	0.85	0.5	0
rs148021310	45,725,250	0.74	0.5	0.74	1	0.85	0.5	0
rs386422402	45,725,448	0.74	0.5	0.74	1	0.85	0.5	0
rs181959846	45,725,614	0.74	0.5	0.74	1	0.85	0.5	0
rs56715955	45,725,739	0.74	0.5	0.74	1	0.85	0.51	0
rs185002523	45,725,878	0.74	0.5	0.74	1	0.85	0.5	0
rs142986624	45,725,945	0.74	0.5	0.74	1	0.85	0.5	0
rs181170401	45,726,006	0.74	0.5	0.74	1	0.85	0.5	0
rs147468361	45,726,047	0.74	0.5	0.74	1	0.85	0.5	0
rs191434584	45,726,190	0.74	0.5	0.74	1	0.85	0.51	0

rs77783265	45,726,549	0.74	0.5	0.74	1	0.85	0.5	0
rs148761251	45,726,701	0.74	0.5	0.74	1	0.85	0.5	0
rs192926463	45,727,043	0.74	0.5	0.74	1	0.85	0.5	0
rs150312307	45,727,190	0.74	0.5	0.74	1	0.85	0.5	0
rs185070442	45,727,443	0.74	0.5	0.74	1	0.85	0.52	0
rs12975661	45,727,496	0.74	0.5	0.74	1	0.85	0.5	0
rs143150894	45,727,571	0.74	0.5	0.74	1	0.85	0.5	0
rs1484323	45,727,902	0.74	0.5	0.74	1	0.85	0.46	0
rs57259996	45,728,546	0.74	0.5	0.74	1	0.85	0.47	0
rs188147127	45,728,651	0.74	0.5	0.74	1	0.85	0.5	0
rs180682471	45,728,671	0.74	0.5	0.74	1	0.85	0.5	0
rs1491279199	45,728,880	0.74	0.5	0.74	1	0.85	0.52	0
rs74253349	45,729,533	0.74	0.5	0.74	1	0.85	0.51	0
rs148814104	45,729,565	0.74	0.5	0.74	1	0.85	0.5	0
rs183330131	45,730,389	0.74	0.5	0.74	1	0.85	0.5	0
rs191036928	45,730,525	0.74	0.5	0.74	1	0.85	0.5	0
rs144251579	45,730,542	0.74	0.5	0.74	1	0.85	0.5	0
rs182375409	45,730,594	0.74	0.5	0.74	1	0.85	0.5	0
rs141492594	45,730,795	0.74	0.5	0.74	1	0.85	0.52	0
rs183418915	45,731,093	0.74	0.5	0.74	1	0.85	0.5	0
rs73939819	45,731,302	0.74	0.5	0.74	1	0.85	0.5	0
rs772723687	45,731,339	0.26	0.5	0	0	0	0.48	0
rs115648030	45,731,348	0.26	0.5	0	0	0	0.48	0
rs75426681	45,731,515	0.26	0.5	0	0	0	0.48	0
rs430319	45,731,762	0.74	0.5	0.74	1	0.85	0.5	0
rs346769	45,731,858	0.74	0.5	0.74	1	0.85	0.5	0
rs184344638	45,732,125	0.74	0.5	0.74	1	0.85	0.5	0
rs60537807	45,732,839	0.26	0.5	0	0	0	0.48	0
rs117473794	45,732,931	0.26	0.5	0	0	0	0.51	0
rs148613044	45,732,972	0.29	0.5	0.75	0.05	0.1	0.5	0.01
rs60507663	45,733,201	0.26	0.5	0	0	0	0.48	0
rs10402508	45,733,782	0.57	0.5	0.74	0.66	0.7	0.5	-0.01
rs186122312	45,733,794	0.74	0.5	0.74	1	0.85	0.5	0
rs190887667	45,733,925	0.74	0.5	0.74	1	0.85	0.5	0
rs116033882	45,734,152	0.26	0.5	0	0	0	0.48	0
rs184169354	45,734,194	0.74	0.5	0.74	1	0.85	0.5	0
rs150191999	45,734,397	0.74	0.5	0.74	1	0.85	0.47	0
rs182987942	45,734,409	0.74	0.5	0.74	1	0.85	0.5	0
rs644177	45,734,433	0.26	0.5	0	0	0	0.46	0
rs182768545	45,734,611	0.74	0.5	0.74	1	0.85	0.5	0
rs115530236	45,734,660	0.74	0.5	0.74	1	0.85	0.51	0
rs187447862	45,734,862	0.74	0.5	0.74	1	0.85	0.5	0
rs79744739	45,735,252	0.74	0.5	0.74	1	0.85	0.5	0
rs138339429	45,735,303	0.74	0.5	0.74	1	0.85	0.47	0
rs193056445	45,735,377	0.74	0.5	0.74	1	0.85	0.5	0
rs73570362	45,735,649	0.74	0.5	0.74	1	0.85	0.5	0
rs141551411	45,735,772	0.74	0.5	0.74	1	0.85	0.5	0

rs187564987	45,735,794	0.74	0.5	0.74	1	0.85	0.5	0
rs150907907	45,735,900	0.74	0.5	0.74	1	0.85	0.5	0
rs112225752	45,736,058	0.74	0.5	0.74	1	0.85	0.5	0
rs144226760	45,736,212	0.74	0.5	0.74	1	0.85	0.51	0
rs181322752	45,736,469	0.74	0.5	0.74	1	0.85	0.5	0
rs147792159	45,736,659	0.74	0.5	0.74	1	0.85	0.5	0
rs75727214	45,737,149	0.74	0.5	0.74	1	0.85	0.52	0
rs60598859	45,737,218	0.74	0.5	0.74	1	0.85	0.55	0
rs182074053	45,737,388	0.74	0.5	0.74	1	0.85	0.5	0
rs182639370	45,722,328	0.72	0.49	0.73	0.98	0.84	0.49	-0.07
rs59242878	45,736,189	0.72	0.49	0.73	0.98	0.84	0.49	-0.07
rs74444899	45,724,060	0.71	0.48	0.73	0.97	0.83	0.48	-0.1
rs386809736	45,723,379	0.45	0.47	0.71	0.42	0.53	0.47	-0.05
rs202180860	45,723,570	0.64	0.46	0.72	0.83	0.77	0.46	-0.09
rs573244969	45,733,117	0.28	0.46	0.57	0.07	0.12	0.46	-0.12

AD prediction accuracy results when running SVM on ADNI SNPs from *EXOC3L2* gene with linear kernel, with random downsampling.

rsID	Location in chr19	Accuracy	Balanced accuracy	Precision	Recall	F1 score	ROC AUC	MCC
rs60528995	45,722,517	0.56	0.56	1	0.11	0.2	0.56	0.24
rs57354345	45,717,615	0.51	0.52	0.5	0.85	0.63	0.52	0.04
rs1969894901	45,724,561	0.51	0.52	0.5	0.89	0.64	0.52	0.05
rs1426173634	45,724,633	0.51	0.52	0.5	0.89	0.64	0.52	0.05
rs12978617	45,724,658	0.51	0.52	0.5	0.89	0.64	0.52	0.05
rs73568222	45,725,975	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs57399322	45,726,106	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs58715307	45,726,458	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs10423753	45,726,563	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs113728460	45,726,654	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs142415915	45,726,745	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs146871722	45,726,758	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs113045530	45,726,821	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs143154520	45,726,869	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs111691933	45,726,964	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs112909419	45,726,968	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs112668741	45,726,976	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs111462669	45,727,167	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs1969927370	45,727,275	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs117316672	45,727,276	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs60048477	45,727,362	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs1387808030	45,727,671	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs1484323	45,727,902	0.51	0.52	0.5	1	0.67	0.52	0.13
rs60406788	45,727,930	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs59172754	45,728,059	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs57767166	45,728,123	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs58647388	45,728,231	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs60081440	45,728,238	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs57787576	45,728,406	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs58846289	45,728,576	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs61625909	45,728,595	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs57002525	45,728,695	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs59741163	45,728,806	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs57403313	45,728,942	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs61552519	45,729,200	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs57035271	45,729,275	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs56879892	45,729,587	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs111269631	45,729,813	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs10424245	45,729,948	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs61463967	45,730,470	0.53	0.52	0.6	0.11	0.19	0.52	0.07
rs182768545	45,734,611	0.51	0.52	0.5	1	0.67	0.52	0.13
rs79744739	45,735,252	0.53	0.52	1	0.04	0.07	0.52	0.14
rs138339429	45,735,303	0.53	0.52	1	0.04	0.07	0.52	0.14

rs77003151	45,715,976	0.51	0.5	0	0	0	0.54	0
rs143393432	45,715,996	0.49	0.5	0.49	1	0.66	0.5	0
rs11667509	45,716,192	0.49	0.5	0.49	1	0.66	0.48	0
rs11667430	45,716,197	0.51	0.5	0	0	0	0.54	0
rs185860487	45,716,283	0.49	0.5	0.49	1	0.66	0.5	0
rs200836902	45,716,357	0.49	0.5	0.49	1	0.66	0.5	0
rs189063316	45,716,364	0.49	0.5	0.49	1	0.66	0.48	0
rs57437338	45,716,678	0.49	0.5	0.49	1	0.66	0.53	0
rs73034885	45,716,899	0.51	0.5	0	0	0	0.48	0
rs1969806189	45,717,169	0.49	0.5	0.49	1	0.66	0.5	0
rs151046811	45,717,248	0.49	0.5	0.49	1	0.66	0.48	0
rs142428371	45,717,296	0.49	0.5	0.49	1	0.66	0.48	0
rs1287006835	45,717,427	0.51	0.5	0	0	0	0.5	0
rs187175824	45,717,625	0.49	0.5	0.49	1	0.66	0.5	0
rs182773180	45,717,719	0.51	0.5	0	0	0	0.5	0
rs144670664	45,717,785	0.49	0.5	0.49	1	0.66	0.48	0
rs187458474	45,717,943	0.49	0.5	0.49	1	0.66	0.48	0
rs189424543	45,718,433	0.51	0.5	0	0	0	0.5	0
rs57112300	45,718,624	0.49	0.5	0.49	1	0.66	0.52	0
rs1305380718	45,718,720	0.49	0.5	0.49	1	0.66	0.5	0
rs188515109	45,718,852	0.49	0.5	0.49	1	0.66	0.5	0
rs59566001	45,719,065	0.49	0.5	0.49	1	0.66	0.52	0
rs73568206	45,719,106	0.49	0.5	0.49	1	0.66	0.5	0
rs115906914	45,719,138	0.49	0.5	0.49	1	0.66	0.5	0
rs141724506	45,719,193	0.49	0.5	0.49	1	0.66	0.5	0
rs56916276	45,719,426	0.49	0.5	0.49	1	0.66	0.5	0
rs73568208	45,719,463	0.49	0.5	0.49	1	0.66	0.5	0
rs76738189	45,719,493	0.51	0.5	0	0	0	0.52	0
rs59259486	45,719,790	0.51	0.5	0	0	0	0.5	0
rs149229776	45,720,307	0.51	0.5	0	0	0	0.5	0
rs1160873375	45,720,764	0.49	0.5	0.49	1	0.66	0.5	0
rs58935155	45,721,596	0.51	0.5	0	0	0	0.48	0
rs79531457	45,722,020	0.51	0.5	0	0	0	0.5	0
rs73568210	45,722,077	0.49	0.5	0.49	1	0.66	0.5	0
rs114505591	45,722,186	0.49	0.5	0.49	1	0.66	0.5	0
rs117755077	45,722,230	0.51	0.5	0	0	0	0.5	0
rs111406553	45,722,265	0.49	0.5	0.49	1	0.66	0.5	0
rs182639370	45,722,328	0.49	0.5	0.49	1	0.66	0.5	0
rs188601304	45,722,334	0.49	0.5	0.49	1	0.66	0.5	0
rs74459616	45,722,529	0.49	0.5	0.49	1	0.66	0.5	0
rs144640143	45,722,616	0.49	0.5	0.49	1	0.66	0.48	0
rs59333887	45,722,743	0.49	0.5	0.49	1	0.66	0.52	0
rs73568215	45,722,773	0.49	0.5	0.49	1	0.66	0.5	0
rs192547359	45,723,142	0.49	0.5	0.49	1	0.66	0.5	0
rs183665617	45,723,228	0.49	0.5	0.49	1	0.66	0.5	0
rs111664793	45,723,233	0.49	0.5	0.49	1	0.66	0.5	0
rs1969881112	45,723,376	0.49	0.5	0.49	1	0.66	0.5	0

rs149548393	45,723,380	0.49	0.5	0.49	1	0.66	0.5	0
rs79890446	45,723,446	0.49	0.5	0.49	1	0.66	0.54	0
rs140668794	45,723,450	0.51	0.5	0	0	0	0.5	0
rs147898480	45,723,518	0.49	0.5	0.49	1	0.66	0.5	0
rs202180860	45,723,570	0.49	0.5	0.49	1	0.66	0.53	0
rs12461144	45,723,706	0.49	0.5	0.49	1	0.66	0.48	0
rs112102023	45,723,714	0.49	0.5	0.49	1	0.66	0.5	0
rs78222968	45,723,832	0.49	0.5	0.49	1	0.66	0.5	0
rs10406604	45,723,986	0.49	0.5	0.49	1	0.66	0.45	0
rs73034893	45,724,044	0.51	0.5	0	0	0	0.48	0
rs74444899	45,724,060	0.49	0.5	0.49	1	0.66	0.48	0
rs141046425	45,724,110	0.49	0.5	0.49	1	0.66	0.5	0
rs1194606521	45,724,296	0.49	0.5	0.49	1	0.66	0.53	0
rs1555756029	45,724,297	0.51	0.5	0	0	0	0.53	0
rs28645301	45,724,692	0.49	0.5	0.49	1	0.66	0.48	0
rs184287728	45,724,732	0.51	0.5	0	0	0	0.5	0
rs28607628	45,724,743	0.49	0.5	0.49	1	0.66	0.46	0
rs28564302	45,724,868	0.49	0.5	0.49	1	0.66	0.48	0
rs386809738	45,724,961	0.49	0.5	0.49	1	0.66	0.48	0
rs60269219	45,724,963	0.49	0.5	0.49	1	0.66	0.48	0
rs144619413	45,725,081	0.51	0.5	0	0	0	0.48	0
rs181890181	45,725,109	0.51	0.5	0	0	0	0.5	0
rs59356929	45,725,127	0.49	0.5	0.49	1	0.66	0.48	0
rs141681064	45,725,149	0.51	0.5	0	0	0	0.48	0
rs58213824	45,725,185	0.49	0.5	0.49	1	0.66	0.48	0
rs185288032	45,725,199	0.49	0.5	0.49	1	0.66	0.52	0
rs80074203	45,725,247	0.49	0.5	0.49	1	0.66	0.5	0
rs148021310	45,725,250	0.51	0.5	0	0	0	0.5	0
rs57294488	45,725,481	0.49	0.5	0.49	1	0.66	0.48	0
rs181959846	45,725,614	0.49	0.5	0.49	1	0.66	0.5	0
rs56715955	45,725,739	0.49	0.5	0.49	1	0.66	0.44	0
rs185002523	45,725,878	0.49	0.5	0.49	1	0.66	0.5	0
rs142986624	45,725,945	0.51	0.5	0	0	0	0.5	0
rs181170401	45,726,006	0.49	0.5	0.49	1	0.66	0.5	0
rs147468361	45,726,047	0.51	0.5	0.5	0.04	0.07	0.5	0
rs186531134	45,726,149	0.49	0.5	0.49	1	0.66	0.5	0
rs191434584	45,726,190	0.49	0.5	0.49	1	0.66	0.5	0
rs10423031	45,726,224	0.49	0.5	0.49	1	0.66	0.48	0
rs77783265	45,726,549	0.51	0.5	0	0	0	0.5	0
rs148761251	45,726,701	0.49	0.5	0.49	1	0.66	0.5	0
rs112759099	45,726,845	0.51	0.5	0	0	0	0.52	0
rs192926463	45,727,043	0.49	0.5	0.49	1	0.66	0.5	0
rs150312307	45,727,190	0.49	0.5	0.49	1	0.66	0.5	0
rs185070442	45,727,443	0.49	0.5	0.49	1	0.66	0.48	0
rs12975661	45,727,496	0.49	0.5	0.49	1	0.66	0.5	0
rs143150894	45,727,571	0.51	0.5	0	0	0	0.5	0
rs10405194	45,727,622	0.49	0.5	0.49	1	0.66	0.48	0

rs57259996	45,728,546	0.49	0.5	0.49	0.93	0.64	0.5	-0.01
rs188147127	45,728,651	0.51	0.5	0	0	0	0.5	0
rs180682471	45,728,671	0.49	0.5	0.49	1	0.66	0.5	0
rs1491279199	45,728,880	0.51	0.5	0	0	0	0.5	0
rs56675703	45,729,123	0.49	0.5	0.49	1	0.66	0.48	0
rs74253349	45,729,533	0.49	0.5	0.49	1	0.66	0.44	0
rs148814104	45,729,565	0.49	0.5	0.49	1	0.66	0.5	0
rs145952987	45,729,924	0.49	0.5	0.49	1	0.66	0.5	0
rs112405270	45,730,238	0.49	0.5	0.49	1	0.66	0.48	0
rs183330131	45,730,389	0.49	0.5	0.49	1	0.66	0.5	0
rs191036928	45,730,525	0.51	0.5	0	0	0	0.5	0
rs144251579	45,730,542	0.49	0.5	0.49	1	0.66	0.5	0
rs182375409	45,730,594	0.49	0.5	0.49	1	0.66	0.5	0
rs141492594	45,730,795	0.51	0.5	0	0	0	0.48	0
rs183418915	45,731,093	0.49	0.5	0.49	1	0.66	0.5	0
rs73939819	45,731,302	0.49	0.5	0.49	1	0.66	0.5	0
rs772723687	45,731,339	0.49	0.5	0.49	1	0.66	0.48	0
rs115648030	45,731,348	0.49	0.5	0.49	1	0.66	0.48	0
rs75426681	45,731,515	0.49	0.5	0.49	1	0.66	0.48	0
rs57045381	45,731,564	0.51	0.5	0	0	0	0.48	0
rs430319	45,731,762	0.49	0.5	0.49	1	0.66	0.5	0
rs346769	45,731,858	0.49	0.5	0.49	1	0.66	0.5	0
rs184344638	45,732,125	0.49	0.5	0.49	0.96	0.65	0.5	0
rs140013593	45,732,661	0.51	0.5	0	0	0	0.48	0
rs58258155	45,732,725	0.51	0.5	0	0	0	0.48	0
rs60537807	45,732,839	0.49	0.5	0.49	1	0.66	0.44	0
rs117473794	45,732,931	0.51	0.5	0	0	0	0.48	0
rs181222539	45,732,960	0.49	0.5	0.49	0.96	0.65	0.5	0
rs148613044	45,732,972	0.51	0.5	0	0	0	0.44	0
rs573244969	45,733,117	0.51	0.5	0	0	0	0.57	0
rs60507663	45,733,201	0.49	0.5	0.49	1	0.66	0.44	0
rs1970013094	45,733,214	0.51	0.5	0	0	0	0.52	0
rs189049349	45,733,309	0.49	0.5	0.49	1	0.66	0.46	0
rs10402508	45,733,782	0.49	0.5	0.49	1	0.66	0.53	0
rs186122312	45,733,794	0.49	0.5	0.49	1	0.66	0.5	0
rs10402739	45,733,897	0.51	0.5	0	0	0	0.48	0
rs190887667	45,733,925	0.49	0.5	0.49	1	0.66	0.5	0
rs116033882	45,734,152	0.49	0.5	0.49	1	0.66	0.48	0
rs184169354	45,734,194	0.49	0.5	0.49	1	0.66	0.5	0
rs1970025005	45,734,195	0.49	0.5	0.49	1	0.66	0.48	0
rs150191999	45,734,397	0.51	0.5	0	0	0	0.5	0
rs182987942	45,734,409	0.49	0.5	0.49	1	0.66	0.5	0
rs644177	45,734,433	0.49	0.5	0.49	1	0.66	0.46	0
rs115530236	45,734,660	0.49	0.5	0.49	1	0.66	0.5	0
rs62118504	45,734,751	0.51	0.5	0	0	0	0.52	0
rs187447862	45,734,862	0.49	0.5	0.49	1	0.66	0.5	0
rs193056445	45,735,377	0.49	0.5	0.49	1	0.66	0.5	0

rs73570362	45,735,649	0.49	0.5	0.49	1	0.66	0.5	0
rs141551411	45,735,772	0.49	0.5	0.49	1	0.66	0.5	0
rs187564987	45,735,794	0.51	0.5	0	0	0	0.5	0
rs150907907	45,735,900	0.49	0.5	0.49	1	0.66	0.48	0
rs59647713	45,736,003	0.49	0.5	0.49	1	0.66	0.48	0
rs112225752	45,736,058	0.49	0.5	0.49	1	0.66	0.5	0
rs144226760	45,736,212	0.49	0.5	0.49	1	0.66	0.5	0
rs181322752	45,736,469	0.49	0.5	0.49	1	0.66	0.5	0
rs147792159	45,736,659	0.51	0.5	0	0	0	0.5	0
rs75727214	45,737,149	0.49	0.5	0.49	1	0.66	0.46	0
rs60598859	45,737,218	0.49	0.5	0.49	1	0.66	0.49	0
rs182074053	45,737,388	0.49	0.5	0.49	1	0.66	0.5	0
rs59242878	45,736,189	0.49	0.48	0	0	0	0.48	-0.13
rs386422402	45,725,448	0.44	0.44	0.46	0.81	0.59	0.44	-0.17
rs386809736	45,723,379	0.42	0.42	0.4	0.37	0.38	0.42	-0.17

Results of running DNABERT transformer model on each window of *APOE* gene from ADNI dataset, using random downsampling for dataset balance.

Window start pos.	Window end pos.	Train accuracy	Train F1 score	Train AUC	Test accuracy	Test F1 score	Test AUC
45,409,038	45,409,138	0.501	0.334	0.482	0.491	0.329	0.482
45,409,088	45,409,188	0.499	0.333	0.388	0.509	0.337	0.388
45,409,138	45,409,238	0.499	0.333	0.321	0.509	0.337	0.321
45,409,188	45,409,288	0.499	0.333	0.519	0.509	0.337	0.519
45,409,238	45,409,338	0.501	0.334	0.482	0.491	0.329	0.482
45,409,288	45,409,388	0.501	0.334	0.500	0.491	0.329	0.500
45,409,338	45,409,438	0.499	0.333	0.482	0.509	0.337	0.482
45,409,388	45,409,488	0.501	0.334	0.481	0.491	0.329	0.481
45,409,438	45,409,538	0.499	0.333	0.500	0.509	0.337	0.500
45,409,488	45,409,588	0.499	0.333	0.518	0.509	0.337	0.518
45,409,538	45,409,638	0.501	0.334	0.500	0.491	0.329	0.500
45,409,588	45,409,688	0.499	0.333	0.482	0.509	0.337	0.482
45,409,638	45,409,738	0.501	0.334	0.519	0.491	0.329	0.519
45,409,688	45,409,788	0.499	0.333	0.518	0.509	0.337	0.518
45,409,738	45,409,838	0.501	0.334	0.482	0.491	0.329	0.482
45,409,788	45,409,888	0.501	0.334	0.518	0.491	0.329	0.518
45,409,838	45,409,938	0.501	0.334	0.481	0.491	0.329	0.481
45,409,888	45,409,988	0.501	0.334	0.500	0.491	0.329	0.500
45,409,938	45,410,038	0.499	0.333	0.631	0.509	0.337	0.631
45,409,988	45,410,088	0.501	0.334	0.520	0.491	0.329	0.520
45,410,038	45,410,138	0.501	0.334	0.518	0.491	0.329	0.518
45,410,088	45,410,188	0.501	0.334	0.518	0.491	0.329	0.518
45,410,138	45,410,238	0.499	0.333	0.481	0.509	0.337	0.481
45,410,188	45,410,288	0.499	0.333	0.519	0.509	0.337	0.519
45,410,238	45,410,338	0.499	0.333	0.482	0.509	0.337	0.482
45,410,288	45,410,388	0.501	0.334	0.481	0.491	0.329	0.481
45,410,338	45,410,438	0.499	0.333	0.500	0.509	0.337	0.500
45,410,388	45,410,488	0.501	0.334	0.439	0.491	0.329	0.439
45,410,438	45,410,538	0.501	0.334	0.380	0.491	0.329	0.380
45,410,488	45,410,588	0.499	0.333	0.500	0.509	0.337	0.500
45,410,538	45,410,638	0.499	0.333	0.519	0.509	0.337	0.519
45,410,588	45,410,688	0.499	0.333	0.518	0.509	0.337	0.518
45,410,638	45,410,738	0.501	0.334	0.519	0.491	0.329	0.519
45,410,688	45,410,788	0.501	0.334	0.518	0.491	0.329	0.518
45,410,738	45,410,838	0.501	0.334	0.519	0.491	0.329	0.519
45,410,788	45,410,888	0.501	0.334	0.482	0.491	0.329	0.482
45,410,838	45,410,938	0.495	0.338	0.499	0.473	0.321	0.499
45,410,888	45,410,988	0.499	0.333	0.500	0.509	0.337	0.500
45,410,938	45,411,038	0.501	0.334	0.519	0.491	0.329	0.519
45,410,988	45,411,088	0.501	0.334	0.518	0.491	0.329	0.518
45,411,038	45,411,138	0.499	0.333	0.481	0.509	0.337	0.481
45,411,088	45,411,188	0.501	0.334	0.481	0.491	0.329	0.481
45,411,138	45,411,238	0.501	0.334	0.500	0.491	0.329	0.500

45,411,188	45,411,288	0.501	0.334	0.481	0.491	0.329	0.481
45,411,238	45,411,338	0.501	0.334	0.500	0.491	0.329	0.500
45,411,288	45,411,388	0.501	0.334	0.519	0.491	0.329	0.519
45,411,338	45,411,438	0.499	0.333	0.481	0.509	0.337	0.481
45,411,388	45,411,488	0.501	0.334	0.481	0.491	0.329	0.481
45,411,438	45,411,538	0.501	0.334	0.481	0.491	0.329	0.481
45,411,488	45,411,588	0.499	0.333	0.481	0.509	0.337	0.481
45,411,538	45,411,638	0.499	0.333	0.482	0.509	0.337	0.482
45,411,588	45,411,688	0.501	0.334	0.500	0.491	0.329	0.500
45,411,638	45,411,738	0.499	0.333	0.518	0.509	0.337	0.518
45,411,688	45,411,788	0.501	0.334	0.482	0.491	0.329	0.482
45,411,738	45,411,838	0.499	0.333	0.518	0.509	0.337	0.518
45,411,788	45,411,888	0.499	0.333	0.500	0.509	0.337	0.500
45,411,838	45,411,938	0.499	0.333	0.519	0.509	0.337	0.519
45,411,888	45,411,988	0.499	0.333	0.593	0.509	0.337	0.593
45,411,938	45,412,038	0.499	0.333	0.458	0.509	0.337	0.458
45,411,988	45,412,088	0.501	0.334	0.536	0.491	0.329	0.536
45,412,038	45,412,138	0.499	0.333	0.537	0.509	0.337	0.537
45,412,088	45,412,188	0.499	0.333	0.518	0.509	0.337	0.518
45,412,138	45,412,238	0.499	0.333	0.519	0.509	0.337	0.519
45,412,188	45,412,288	0.499	0.333	0.519	0.509	0.337	0.519
45,412,238	45,412,338	0.501	0.334	0.519	0.491	0.329	0.519
45,412,288	45,412,388	0.499	0.333	0.500	0.509	0.337	0.500
45,412,338	45,412,438	0.501	0.334	0.519	0.491	0.329	0.519
45,412,388	45,412,488	0.501	0.334	0.500	0.491	0.329	0.500
45,412,438	45,412,538	0.499	0.333	0.518	0.509	0.337	0.518
45,412,488	45,412,588	0.499	0.333	0.518	0.509	0.337	0.518
45,412,538	45,412,638	0.499	0.333	0.518	0.509	0.337	0.518
45,412,588	45,412,650	0.499	0.333	0.481	0.509	0.337	0.481