JOURNAL OF
EMERGING INVESTIGATORS

Article

A novel approach for predicting Alzheimer's disease using machine learning on DNA methylation in blood

Evan C. Adami[1], Ayush Pandit[2]

[1] San Dieguito High School Academy, Encinitas, California

[2] Stanford University, Stanford, California

SUMMARY

Alzheimer's disease (AD) is the leading cause of dementia worldwide. Mild cognitive impairment (MCI) is an early stage of mental decline that can precede the development of AD. Oftentimes, the progression of MCI to AD is difficult to predict. An AD diagnosis can involve invasive brain scans and spinal fluid tests. However, one promising biomarker for the diagnosis of AD is epigenetic data, specifically the methylation level of cytosine-phosphate-guanine (CpG) sites in the DNA. AD is linked to environmental factors, with some factors causing changes in these DNA methylation levels. We hypothesized that machine learning models can use blood DNA methylation levels, sex, and age to predict between cognitive normality, MCI, and AD with at least 50% accuracy. In this paper, we generated four machine learning models and two dataset dimensionality reduction methods in order to test this hypothesis. We trained the models on data from CpG sites from whole blood. When predicting if a patient would be cognitively normal, have MCI, or have AD, we achieved an accuracy of 53.33%, which is 20% greater than random guessing. We achieved this accuracy using a gradient boosting decision trees model in combination with a logistic regression method for feature selection. While this accuracy is low, the feature selection method that we developed may be useful in future research, as it identifies CpG sites that are most correlated with AD. Since peripheral blood is easily accessible through blood draw, our model represents a practical way to assist in the diagnosis of AD.

INTRODUCTION

Alzheimer's disease (AD) affects an estimated 6.5 million people in America (1). Over 10% of those above the age of 65 have AD (1). Alzheimer's negatively affects memory, cognitive function, and sensory processing, preventing people from carrying out daily tasks (2). AD is currently incurable, but medications and therapies exist to treat the disease and increase cognitive function (2). An early diagnosis of AD can help to select these treatments more effectively (3). However, it is difficult to definitively diagnose AD in its early stages, as tests involving brain imaging, brain tissue, and spinal taps are invasive and expensive (4).

Recently, AD was associated with high DNA methylation (DNAm) levels at cytosine-phosphate-guanine (CpG) sites in the brain (5, 6). Specifically, epigenome-wide association studies have revealed hundreds of differentially methylated regions in neurons that are associated with AD (7). DNAm is an epigenetic mechanism that controls whether a gene will be expressed or not (8). Since gene expression patterns in cognitively normal (CN) brains are different from patterns in AD brains, gene expression is an important area of study for AD diagnosis (9). Machine learning (ML) is a tool that uses large datasets in order to learn and predict outcomes and can have powerful applications in healthcare and disease diagnosis (10). These applications extend to the diagnosis of AD. Using ML, one study was able to achieve up to a 92% successful classification rate when predicting if a person would have Alzheimer's based on the CpG DNAm data from their brain (11). This study highlights the diagnostic potential of epigenetic biomarkers in the brain; however, its practical implications are limited because brain tissue is difficult to collect and analyze, as mentioned previously.

Alternatively, epigenome-wide association studies of peripheral blood have identified several differentially methylated CpG sites that are correlated with AD progression and cognitive decline, especially in the HOXB6 gene (12–14). These studies demonstrate that alternative to brain tissue, blood tissue contains promising and accessible biomarkers for AD. Our research aims to expand upon these findings by utilizing ML. An epigenetic change in blood DNAm levels might be the key for an early AD diagnosis as both the disease and DNAm levels are heritable and linked to environmental factors (15). Furthermore, data on blood DNAm is becoming increasingly accessible (16). We hypothesized that ML models could predict CN, mild cognitive impairment (MCI), or AD with at least 50% accuracy using DNAm data from blood. Given the limited amount of current data, a 50% prediction accuracy represents a substantial improvement over guessing between three classes.

In this work, we constructed four ML models to classify patients as either CN, MCI, or AD. While MCI can be an indicator of other late-onset disorders besides AD, this study focuses only on its connection to the progression of AD. We trained the ML models on a large set of data that includes the DNAm levels at 410,942 unique CpG sites in whole blood from each patient in a study. CpG sites are methylated on a scale from 0 to 100 percent. When abnormal levels of methylation are present at certain CpG sites, this can indicate that a patient may have AD. We reduced the dimensionality of the data before training via a feature selection method using logistic regression. The method identifies the 410 CpG sites that are most correlated with AD.

The logistic regression feature selection method is shown to greatly reduce the noise of the data and improve accuracies. After training on the dataset, a gradient boosting decision trees (GBDT) model achieved the highest accuracy of 53.33%, and highest area under the curve (AUC) of 0.69

when predicting on a subset of the data reserved for testing. Random guessing would have only yielded a 33.33% accuracy and an AUC of 0.5. These results confirmed our hypothesis that ML can achieve up to a 50% accuracy in predicting AD using blood DNAm data alone. This suggests that a relationship does exist between blood-based epigenetics and AD, and it could lead to a practical healthcare application in diagnosing AD from blood samples using both machine learning and DNAm biomarkers.

RESULTS

For the dataset that we obtained from the Gene Expression Omnibus (17), each case includes the sex, age, disease state 1, disease state 2, and 410,942 unique measured CpG sites with their respective methylation levels. Disease state 1 is a person's measured disease state at the initial time of data collection, whether that be CN, MCI, or AD. Disease state 2 is a person's disease state one year after disease state 1 was measured. For example, an individual's disease state 1 could be MCI while their disease state 2 is AD, if they progressed to Alzheimer's within a year.

We tested the predictive capabilities of a k-nearest neighbors, logistic regression, support vector machine (SVM), and GBDT model. ML models were evaluated by their accuracy score, a confusion matrix, receiver operating characteristic (ROC) curve, and AUC. An ROC curve plots a model's true positive rate versus its false positive rate. AUC measures the area under this curve. AUC values of 0.5 represent random chance, and AUC values closer to 1 are optimal. We first investigated if ML models could use disease state 1 as a training feature to predict disease state 2. In this way, we could determine if ML would be a useful tool in predicting the progression of existing cognitive decline and AD, not just a tool for diagnosing AD. When predicting disease state 2 using the full training dataset, the ML models all performed with surprisingly high accuracies of up to 88.33%.

While these results appeared promising, we found them to be flawed upon further inspection. We discovered that models were predicting disease state 2 to be exactly what disease state 1 was every single time. For example, if an individual's first disease state was MCI, then the model would always predict that in one year the patient would still have MCI, regardless of their DNAm levels. Due to the nature of the dataset, this artificially inflated the model's accuracy. However, these results do not reveal any relationship between CpG site DNAm levels and Alzheimer's disease, as our models placed no weight on those important features when predicting disease state 2. Therefore, the results are inapplicable to our studies. Training the models with access to disease state 1 made them too dependent on a single feature due to overfitting, allowing them to ignore the DNAm data.

We removed disease state 1 from the training features, then examined ML models' ability to predict disease state 1, disregarding disease state 2. This is representative of the diagnostic capabilities of ML, as our models would now be making predictions of an individual's cognitive disease state based solely upon their sex, age, and blood DNAm levels. When predicting disease state 1, the machine learning models struggled to achieve high accuracies. The models performed worst when trained on the full training dataset, since accuracies for all models ranged from about 30% to 35%. This is not better than random guessing, which would yield about 33.33% accuracy. The models did not perform

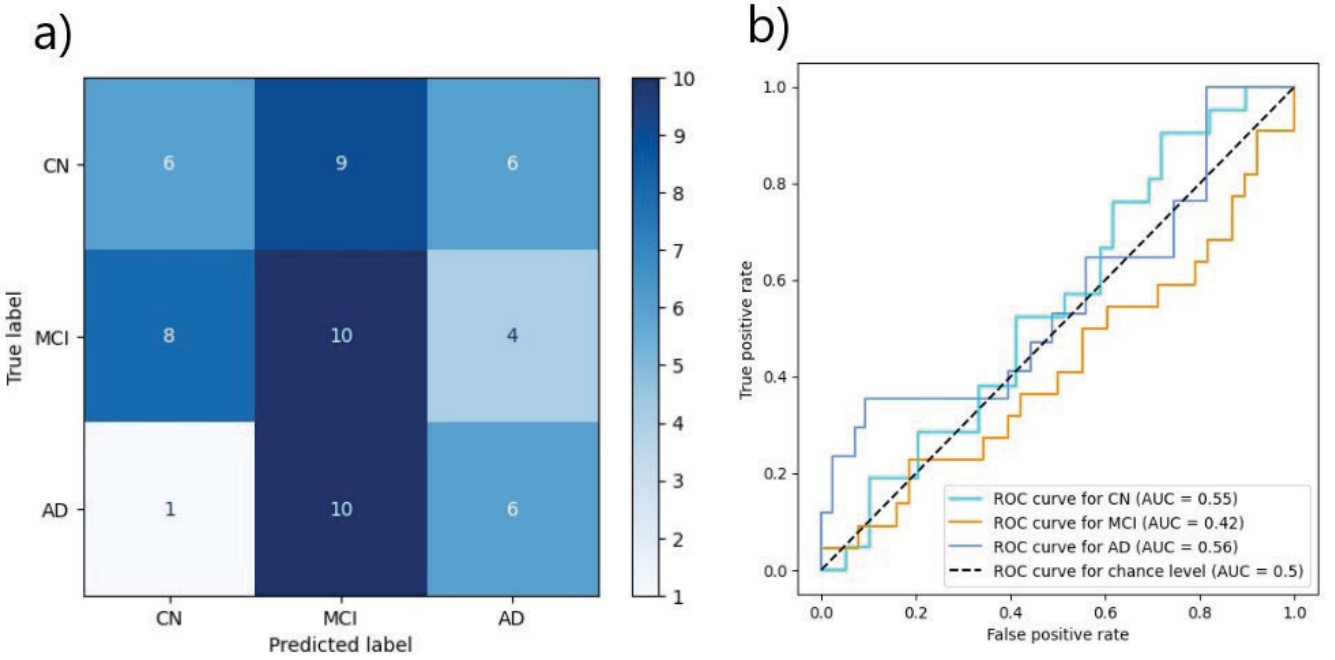

Figure 1: Metrics for predicting disease state 1 after PCA. These metrics are from the GBDT model that was trained on data reduced using PCA. a) Confusion matrix showing the true label of disease state versus the label of disease state predicted by the GBDT model. For example, the box in the center represents the number of times that the model correctly predicted that a sample would have MCI when they truly had MCI (10 times). The box in the center of the bottom row represents the number of times that the model incorrectly predicted that a sample would have MCI when they truly had AD (10 times). The confusion matrix of a perfect predictive model would only have values in the top left, center, and bottom right boxes. b) ROC curves for CN, MCI, and AD plotted against the ROC curve for chance level.

better when trained on the dataset that was reduced using principal component analysis (PCA). PCA consolidates the data into principal components, a minimum number of features that can be used to explain a certain amount of the data's variance. We found that using PCA to explain 80% of variance yielded the highest accuracies. However, model accuracies did not exceed 40% when trained on PCA-reduced data. We plotted a confusion matrix and ROC AUC graph for an optimized gradient boosting decision trees (GBDT) model that achieved 36.66% accuracy **(Figure 1)**.

The ROC curves for all three classes loosely aligned with the ROC curve that represents random guessing, reaffirming that the accuracies achieved by the models in this stage were largely due to chance. The greatest AUC value was 0.56, only slightly above chance level at 0.5. The confusion matrix shows that the GBDT model frequently predicted MCI, which is most likely because MCI cases made up the largest portion of the training data. The confusion matrices for the other three models we tested—k-nearest neighbors, SVM, and logistic regression—all looked like that of the GBDT model, with MCI being the most predicted class. From these results, we saw that PCA was ineffective in reducing the dataset to a reasonable size that models could use to be trained. There was too much noise in the large number of features, so the important features could not be extracted.

The logistic regression method that we devised for reducing dimensionality proved to be substantially more effective. When the models were trained on the dataset that was feature-selected using our logistic regression method, all accuracies increased. The GBDT model consistently had the best performance of all four models, achieving an accuracy of 53.33% when optimized. We plotted a confusion matrix and ROC AUC graph for this GBDT model **(Figure 2)**.

Based on the ROC AUC graph, the model's accuracy was not due to chance; this contrasts to the ROC AUC graph when the model was trained on the full dataset or PCA-reduced data, which was due to chance. The higher scores for AUC, reaching 0.68 when predicting MCI, indicated that the model was considerably better at distinguishing between classes. From the confusion matrix, we can see that the model no longer predicted MCI the most. Instead, there was a better spread of predictions, with 53.33% of the predictions being correct.

DISCUSSION

We suspect that the ML models' low accuracy when predicting on the full dataset or PCA-reduced data was because of overfitting, meaning that the models would only be able to predict accurately on the exact set of data that was used for training. The improved accuracy when predicting on only a select amount of CpG sites, rather than the whole dataset or on principal components, suggests that there is a relationship between blood DNAm levels of specific genes and cognitive decline. In fact, the results of our blood-based epigenetic study extend the results of a similar study on brain epigenetics. In that study, the researchers found that feature-selecting CpG sites in the brain to train ML models was the most effective form of dimensionality reduction, and they were able to greatly increase the accuracy of their classifier in predicting if a patient had AD (11). Furthermore, it is important

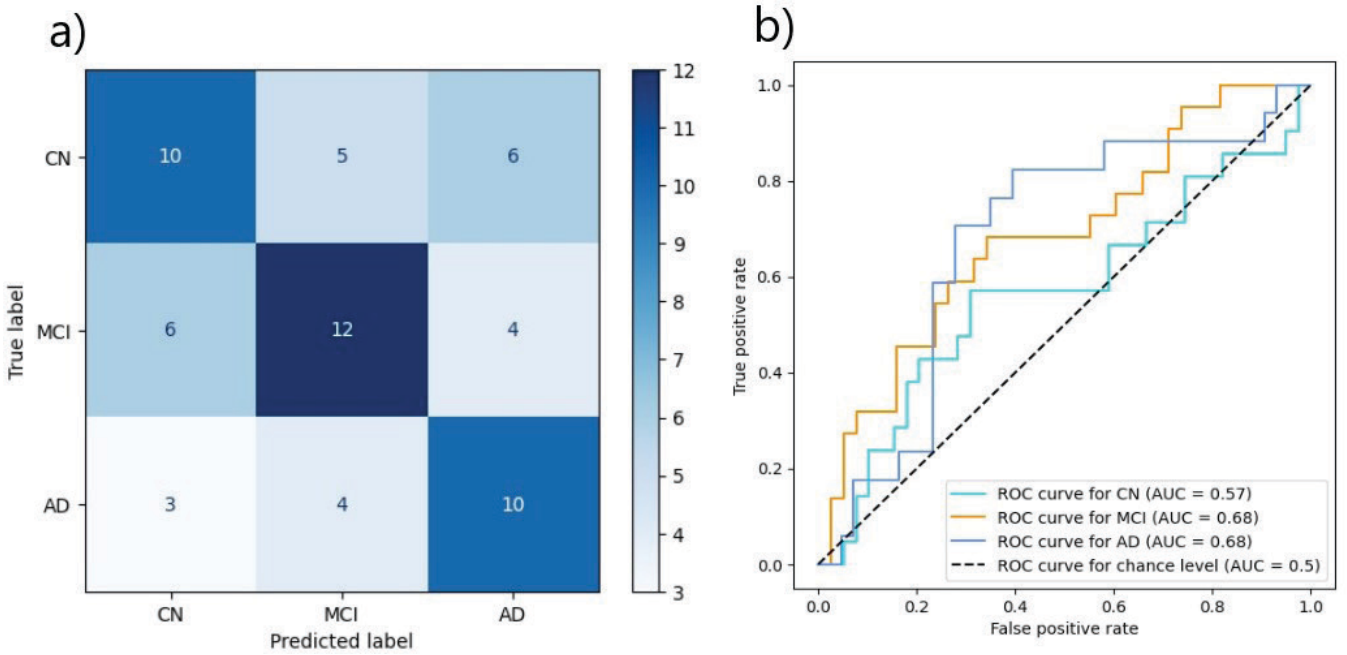

Figure 2: Metrics for predicting disease state 1 after feature selection. These metrics are from the GBDT model that was trained on data reduced using the logistic regression method for feature selection. a) Confusion matrix showing the true label of disease state versus the label of disease state predicted by the GBDT model. For example, the box in the center represents the number of times that the model correctly predicted that a sample would have MCI when they truly had MCI (12 times). The box in the center of the bottom row represents the number of times that the model incorrectly predicted that a sample would have MCI when they truly had AD (4 times). The confusion matrix of a perfect predictive model would only have values in the top left, center, and bottom right boxes. b) ROC curves for CN, MCI, and AD plotted against the ROC curve for chance level.

to note that many of the CpG sites that were feature selected by our logistic regression method were also differentially methylated CpG sites in the *HOXB6* gene that have been previously associated with AD (14). While ML may not yet be capable of diagnosing AD based on peripheral blood DNAm data due to its accuracy of only 53.33%, ML does have a place in identifying AD correlated CpG sites. Knowing what genes are affected by AD is the first step in making early AD predictions through noninvasive blood draws and analysis. We have demonstrated that ML, especially a GBDT model, shows promise in making these predictions. Our models were most accurate in predicting if a patient would have MCI. Identifying this state of cognitive decline at an early point in time is crucial in administering valuable therapies to patients (3). Our ML method may find a use here. Furthermore, our logistic regression method for feature selection and dimensionality reduction is effective in selecting which CpG sites are important in AD diagnosis. By selecting features in chunks of 1,000, the process is designed to be scalable as more blood DNAm data on Alzheimer's becomes available. Our results highlight the necessity of feature selection in order to create a reasonable amount of data for ML models to train on, with limited noise and overfitting.

The largest limitation that we faced was the availability of data on blood epigenetics and AD. The dataset that we trained on contained data for only 300 samples, which is not enough to apply our findings to the general population. With access to more data, our ML models might have achieved greater predictive accuracy, although we cannot be certain of this. A second limitation of our work is that we did not control for other conditions that could affect one's blood DNAm levels. For example, heart disease has been shown in the past to be correlated with blood DNAm levels at CpG sites (18). If some of those who had AD in our dataset also had heart disease, then our models may have been incorrectly associating features with AD that were more correlated with heart disease instead. This is another reason that more extensive data is needed for this area of research, so that confounding variables can be better controlled. A third limitation that we faced was the complexity of our models. It would be important in the future to test more complex ML models, such as neural networks, to determine if they can predict AD with a higher accuracy. However, it is likely that a neural network would overfit due to the high ratio of feature to sample data. Finally, it is important to note that MCI does not always lead to AD. One may remain in a stable state of MCI and never progress to a worse cognitive state. Since our data was minimal, our ML method was not effective in predicting if one would go through this progression, as evidenced when we tried to predict disease state 2. This connection between MCI and AD would be important to explore using ML in future research.

As more treatments become available for AD, reliable and early diagnosis of AD will become increasingly valuable for cognitively impaired patients. Our accuracy of 53.33% when predicting if a patient would have MCI or AD should only improve as more data on this field becomes available and as ML techniques progress. The results also indicate a relationship between differential CpG site methylation and AD, which will be useful in further research and additionally in identifying hereditary or environmental risk factors for AD. The feature selection process proposed in this paper using logistic regression could be replicated and applied to new datasets as well. Correlations between blood-based epigenetics and other diseases besides AD could also be explored using similar ML methods. ML holds much potential for the future of disease diagnosis.

MATERIALS AND METHODS

Dataset

The dataset used in our work is downloaded from the Gene Expression Omnibus (17) under the accession ID: GSE144858. The data was collected in an epigenome-wide association study of DNAm patterns in whole blood, from patients with AD and MCI. Blood samples were analyzed using the Illumina Infinium Human Methylation 450K BeadChip array to find the DNAm levels. The dataset includes a total of 300 sample cases **(Table 1)**. Each case includes a unique accession ID, initial disease state, a second disease state measured one year after the initial disease state was obtained, sex, and age. There are 120 males and 180 females in the dataset. Furthermore, each case has 410,942 measured CpG sites and their respective methylation levels as a decimal value ranging from 0 (no methylation) to 1 (100% methylation). The methylation levels were measured at the time that disease state 1 was recorded. Each CpG site is given a unique Illumina ID corresponding to its location in the genome. Using the pandas Python library, we parsed and organized the data into a data frame: each row represents a unique case, marked by an accession ID, and each column represents a unique feature **(Table 2)**. We converted non-numerical features into numbers. Male is represented by a 0 and female is represented by a 1. CN is represented by a 0, MCI is represented by a 1, and AD is represented by a 2. The final data frame is 300 rows × 410,947 columns. We used an 80/20 train/test split on the preprocessed data using the scikit-learn Python library in order to prepare for model training. This means that 80% of the data would be used for training the ML models, and 20% of the data would be used for testing the models.

Dimensionality Reduction

As the dataset contained many features (410,942 measured CpG sites per sample), it was likely that our ML models would overfit on the training data. In order to prevent this, we used a method involving logistic regression to select features and reduce the dimensionality of our data. We trained a logistic regression model on 1,000 features of the dataset at a time

Disease classification	disease state 1	disease state 2
CN	96	96
MCI	111	69
AD	93	135
Total	300	300

Table 1: Number of cases for disease states 1 and 2. The table shows the number of cases for each disease classification during two different times of measurement, in order to track disease progression. The disease classifications are cognitively normal (CN), mild cognitive impairment (MCI), or Alzheimer's disease (AD). Disease state 1 is the initial measured disease classification, and disease state 2 is the measured disease classification one year after disease state 1 was measured.

	accession	sex	age	disease state 1	disease state 2	cg00000029	cg00000108	cg00000109	cg00000165	…
0	GSM4299202	1	76	1	1	0.610080	0.892013	0.679462	0.380888	…
1	GSM4299203	0	86	2	2	0.461774	0.874571	0.743489	0.315226	…
2	GSM4299204	1	80	2	2	0.532801	0.864435	0.750426	0.408930	…
3	GSM4299205	0	81	0	0	0.525638	0.888415	0.766076	0.363258	…
4	GSM4299206	1	67	1	1	0.565315	0.876623	0.719490	0.355817	…

Table 2: First five rows of preprocessed data frame. This table shows the first five rows of the preprocessed DNAm data obtained from the Gene Expression Omnibus (17). Each accession ID represents a unique individual in the dataset. For sex, 0 = male and 1 = female. For the disease states, 0 = CN, 1 = MCI, and 2 = AD. The data frame contains 410,942 measured CpG sites and their respective methylation levels for each individual. The CpG sites are identified by their Illumina IDs, and their methylation levels are represented as a decimal value ranging from 0 (no methylation) to 1 (100% methylation). The first four CpG sites are shown in the table.

in order for the method to be scalable. From those 1,000 features, the model identified which feature had the largest coefficient; that is, which feature had the most influence in correctly classifying the disease state of a given sample. By the end of this process, the dataset reduced from 410,942 features to only the 410 most important features (from each chunk of 1,000), alongside sex and age. We also tried PCA and compared its results to the logistic regression method. PCA is a dimensionality reduction technique in which features are consolidated into principal components that can explain a majority of the variance in a dataset. We implemented PCA using scikit-learn and explained 80% of the variance.

Machine Learning Models

As the task is a classification problem, we started by setting up four classification models of varying complexity using scikit-learn. The goal was to test each model and see which one would have the best performance for the problem. The four models that we used were: 1) k-nearest neighbors, which converts samples into data points and classifies them based on the classes of their nearest surrounding points; 2) logistic regression, which uses weights/coefficients of features to predict a classification; 3) SVM, which maps data that is not linearly separable to a higher dimension so that it can be separated by a hyperplane; and 4) GBDT, which builds simple decision trees upon each other until classification error is minimized in a series of "boosts". We used a grid search to optimize the hyperparameters of each model. The optimized GBDT model had 200 estimators, a learning rate of 0.1, and a maximum depth of 3. We trained the models on three different training sets: the full training dataset, the training dataset reduced using PCA, and the training dataset reduced using the logistic regression method. This way, we could see what form of dimensionality reduction was most effective, if any at all. The models were then evaluated by classifying samples in the testing dataset as either 0 (CN), 1 (MCI), or 2 (AD) based off the sample's features/DNAm levels. We examined if the models could predict both disease state 2 with access to disease state 1 as a training feature, and disease state 1 alone. This would allow us to see if our models could be used not only for AD diagnosis, but also disease progression analysis.

Evaluation Metrics

Each model was assessed using its accuracy score, confusion matrix, ROC curve, and AUC. Accuracy score measures the percentage of samples that the model correctly classifies. A confusion matrix plots the true classes of samples against what class the model predicted for samples. An ROC curve plots a model's true positive rate versus its false positive rate at different thresholds. AUC measures the area under this curve. An AUC closer to 1 is better, as this means that the model only predicts true positives and never predicts false positives. An AUC of 0.5 represents a model that classifies by random chance.

ACKNOWLEDGMENTS

We are grateful to the Gene Expression Omnibus for providing the data used in this research. Additionally, a special thanks to Inspirit AI for their mentorship and resources throughout this project.

Received: March 16, 2023
Accepted: June 23, 2023
Published: September 20, 2023

REFERENCES

1. "Alzheimer's Disease Facts and Figures." *Alzheimer's Association,* 2023, www.alz.org/alzheimers-dementia/facts-figures#:~:text=More%20than%206%20million%20Americans%20of%20all%20ages%20have%20Alzheimer's,older%20(10.7%25)%20has%20Alzheimer's. Accessed 15 March 2023.
2. "Alzheimer's Disease Fact Sheet." *National Institute on Aging,* 8 Jul. 2021, www.nia.nih.gov/health/alzheimers-disease-fact-sheet. Accessed 15 March 2023.
3. "The Importance of Early Diagnosis in Alzheimer's Disease and Mild Cognitive Impairment." *Roche Diagnostics,* diagnostics.roche.com/global/en/article-listing/early-diagnosis-of-alzheimer-disease-and-mild-cognitive-impairme.html#:~:text=Confirming%20a%20diagnosis%20of%20mild,decline%20in%20their%20cognitive%20functions. Accessed 15 March 2023.
4. "Alzheimer's Disease." *Mayo Clinic,* Mayo Foundation for Medical Education and Research, 2 Feb. 2023, www.mayoclinic.org/diseases-conditions/alzheimers-disease/diagnosis-treatment/drc-20350453. Accessed 15 March

2023.
5. Smith, Rebecca G., *et al.* "A Meta-Analysis of Epigenome-Wide Association Studies in Alzheimer's Disease Highlights Novel Differentially Methylated Loci across Cortex." *Nature Communications,* vol. 12, no. 1, 2021, doi:https://doi.org/10.1038/s41467-021-23243-4.
6. Fetahu, Irfete S., *et al.* "Epigenetic Signatures of Methylated DNA Cytosine in Alzheimer's Disease." *Science Advances,* vol. 5, no. 8, 2019, https://doi.org/10.1038/s41467-021-23243-4.
7. Huang, Yanting, *et al.* "A Machine Learning Approach to Brain Epigenetic Analysis Reveals Kinases Associated with Alzheimer's Disease." *Nature Communications,* vol. 12, no. 1, 2021, https://doi.org/10.1038/s41467-021-24710-8.
8. Moore, Lisa D, *et al.* "DNA Methylation and Its Basic Function." *Neuropsychopharmacology,* vol. 38, no. 1, 2012, pp. 23–38, https://doi.org/10.1038/npp.2012.112.
9. "Gene Expression Signatures of Alzheimer's Disease." National Institute on Aging, U.S. Department of Health and Human Services, 28 May 2019, www.nia.nih.gov/news/gene-expression-signatures-alzheimers-disease. Accessed 31 July 2023.
10. Ahsan, Md Manjurul, *et al.* "Machine-Learning-Based Disease Diagnosis: A Comprehensive Review." *Healthcare,* vol. 10, no. 3, 2022, p. 541., https://doi.org/10.3390/healthcare10030541.
11. Alfonso Perez, Gerardo, and Javier Caballero Villarraso. "Alzheimer Identification through DNA Methylation and Artificial Intelligence Techniques." *Mathematics,* vol. 9, no. 19, 2021, https://doi.org/10.3390/math9192482.
12. Chouliaras, Leonidas, *et al.* "Peripheral DNA Methylation, Cognitive Decline and Brain Aging: Pilot Findings from the Whitehall II Imaging Study." *Epigenomics,* vol. 10, no. 5, 2018, pp. 585–595, https://doi.org/10.2217/epi-2017-0132.
13. Li, Qingqin S., *et al.* "Association of Peripheral Blood DNA Methylation Level with Alzheimer's Disease Progression." *Clinical Epigenetics,* vol. 13, no. 1, 15 Oct. 2021, https://doi.org/10.1186/s13148-021-01179-2.
14. Roubroeks, Janou A.Y., *et al.* "An Epigenome-Wide Association Study of Alzheimer's Disease Blood Highlights Robust DNA Hypermethylation in the HOXB6 Gene." *Neurobiology of Aging,* vol. 95, Nov. 2020, pp. 26–45, https://doi.org/10.1016/j.neurobiolaging.2020.06.023.
15. Robertson, Keith D. "DNA Methylation and Human Disease." *Nature Reviews Genetics,* vol. 6, no. 8, 1 Aug. 2005, pp. 597–610, https://doi.org/10.1038/nrg1655.
16. Houseman, E. Andres, *et al.* "DNA Methylation in Whole Blood: Uses and Challenges." *Current Environmental Health Reports,* vol. 2, no. 2, 2015, pp. 145–154, https://doi.org/10.1007/s40572-015-0050-3.
17. "Gene Expression Omnibus." National Center for Biotechnology Information, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/geo/. Accessed 15 March 2023.
18. Agha, Golareh, *et al.* "Blood Leukocyte DNA Methylation Predicts Risk of Future Myocardial Infarction and Coronary Heart Disease." *Circulation,* vol. 140, no. 8, 19 Aug. 2019, pp. 645–657, https://doi.org/10.1161/CIRCULATIONAHA.118.039357.