mathematics

MDPI

Article

Alzheimer Identification through DNA Methylation and Artificial Intelligence Techniques

Gerardo Alfonso Perez [1,*] and Javier Caballero Villarraso [1,2]

[1] Department of Biochemistry and Molecular Biology, University of Cordoba, 14071 Cordoba, Spain; bc2cavij@uco.es
[2] Biochemical Laboratory, Reina Sofia University Hospital, 14004 Cordoba, Spain
* Correspondence: ga284@cantab.net

Abstract: A nonlinear approach to identifying combinations of CpGs DNA methylation data, as biomarkers for Alzheimer (AD) disease, is presented in this paper. It will be shown that the presented algorithm can substantially reduce the amount of CpGs used while generating forecasts that are more accurate than using all the CpGs available. It is assumed that the process, in principle, can be non-linear; hence, a non-linear approach might be more appropriate. The proposed algorithm selects which CpGs to use as input data in a classification problem that tries to distinguish between patients suffering from AD and healthy control individuals. This type of classification problem is suitable for techniques, such as support vector machines. The algorithm was used both at a single dataset level, as well as using multiple datasets. Developing robust algorithms for multi-datasets is challenging, due to the impact that small differences in laboratory procedures have in the obtained data. The approach that was followed in the paper can be expanded to multiple datasets, allowing for a gradual more granular understanding of the underlying process. A 92% successful classification rate was obtained, using the proposed method, which is a higher value than the result obtained using all the CpGs available. This is likely due to the reduction in the dimensionality of the data obtained by the algorithm that, in turn, helps to reduce the risk of reaching a local minima.

Keywords: algorithm; identification; Alzheimer

Citation: Alfonso Perez, G.; Caballero Villarraso, J. Alzheimer Identification through DNA Methylation and Artificial Intelligence Techniques. *Mathematics* **2021**, *9*, 2482. https://doi.org/10.3390/math9192482

Academic Editors: Monica Bianchini and Maria Lucia Sampoli

Received: 6 September 2021
Accepted: 30 September 2021
Published: 4 October 2021

1. Introduction

Alzheimer (AD) is a relatively common neurological disorder associated with a decline in cognitive skills [1,2] and memory [3–5]. The causes of Alzheimer are not yet well understood, even as some processes of the development of amyloid plaque seems to be a major part of the disease [6]. The development of biomarkers [7] for the detection of AD is of clear importance. Over the last few decades, there has been a sharp increase in the amount of information publicly available, with researchers graciously making their data public. This, coupled with advances, such as the possibility to simultaneously estimate the methylation [8] levels of thousands of CpGs in the DNA, has created a large amount of information. CpG refers to having a guanine nucleotide after a cytosine nucleotide in a section of the DNA sequence. CpGs can be methylated, i.e., having an additional methyl group added. The level of methylation in the DNA is a frequently used marker for multiple illnesses [9–12], as well as a estimator of the biological age of the patient; hence, it has become an important biomarker [13]. The computational task is rather challenging. Current equipment can quickly analyze the level of methylation of in excess of 450,000 CpGs [14–16], with the latest generation of machines able to roughly double that amount [17]. As previously mentioned, methylation data has been linked to many diseases [18–20] and it is a logical research area for AD biomarkers. An additional challenge is that, at least in principle, there could be a highly non-linear process that is not necessarily accurately described by traditional regression analysis. The scope would then, hence, be to try to identify techniques that select a combination of the CpGs to be analyzed

and then a non-linear algorithm that is able to predict whether the patient analyzed has the disease. However, on the other hand, it would not appear reasonable to totally discard the information presented in linear analysis. In the following sections, a mixed approach is presented. It will be shown that the approach is able to generate predictions (classifications between the control and patients suffering from Alzheimer).

1.1. Forecasting and Classification Models

Prediction and/or classification tasks are frequently found in many scientific and engineering fields with a large amount of potential artificial intelligence related techniques. The specific topics covered are rather diverse, including weather forecasts [21], plane flight time deviation [22], distributed networks [23], and many others [24–26]. One frequently used set of techniques are artificial neural networks. These techniques are extensively used in many fields. There are, however, several alternatives, which have received less attention in the existing literature (for instance, k-nearest neighbors and support vector machines). It should be noted that the k-nearest neighbor technique is frequently used in data pre-processing for instance in situations, in which the dataset has some missing values and the researcher needs to estimate those (typically as a previous step before using them as an input into a more complex model).

In our case the non-linear basic classification algorithm chosen was support vector machines (SVM) [27–29]. The basic idea of SVM is dividing the data into hyperplanes [30] and trying to decrease the measures of the classification error. This is achieved by following the usual supervised learning, in which a proportion of the data are used for training the SVM, while other portion (not used during the training phase) is used for testing purposes only, in order to avoid to avoid the issue of overfitting [5,31]. This technique has been applied in the context of Alzheimer for the classification of MRI images [32,33]. Some SVM models have been proposed in the context of CpGs methylation related to AD [34].

1.2. CpG DNA Methylation

A CpG is a dinucleotide pair (composed by cytosine a phosphate and guanine), while methylation refers to the addition of a methyl group to the DNA. Methylation levels are typically expressed as a percentage with 0 indicating completely unmethylated and 1 indicating 100% methylated. CpG DNA methylation levels are frequently used as epigenetic biomarkers [35,36]. Methylation levels change as an individual ages and this has been used to build biological clocks [37]. Individuals with some illnesses such as some cancers and Alzheimer present deviations in their levels of methylations.

1.3. Paper Structure

In the next section a related literature review is carried out given an overview of articles in prediction and classification. The literature review is followed by the materials and methods section, in which the main algorithm is explained. In this section, there is also a subsection describing the analyzed data. In Section 4 the results are presented. This section is divided into two subsection the first one describing the results for a single dataset and the second subsection describing the results when a multi dataset approach is followed. The last two sections are the discussion and the conclusions.

2. Literature Review

As previously mentioned, the CpG DNA methylation data were used in a variety of biomedical applications, such as the creation of biological clocks. For instance, Horvath [38] created an accurate CpG DNA methylation clock. Horvath managed to reduce the dimensionality of the data from hundred of thousands of CpGs analyzed per patient to a few hundred. This biological clock is able to predict the age of patients (in years) with rather high accuracy using as inputs the methylation data of a few hundred CpGs. A related article is [39], in which the authors used neural networks to predict the forensic

age of individuals. The authors showed how using machine learning techniques could improve the accuracy of the age forecast, compared to traditional (linear) models.

Park et al. [40] is an interesting article focusing on DNA methylation and AD. The authors of this article found a link between DNA methylation and AD but similar to Horvath paper did not use machine learning techniques. Machine learning techniques have been applied with some success. For instance, [41] used neural networks to analyze the relationship between gene-promoters methylation and biomarkers (one carbon metabolism in patients). Another interesting model was created by [42]. In this model the authors use a combination of DNA methylation and gene expression data to predict AD. The approached followed by the authors in this paper is different from the one that we pursued as they increased the amount of input data (including gene expression), while we focus on trying to reduce the dimensionality of the existing data i.e., select CpGs.

While most of the existing literature focuses on neural networks, there are also some interesting applications of other techniques such as for instance support vector machines (SVM). For instance, [43] used SVM for the classification of histones. SVM have also been used for classification purposes in some illnesses such as colorectal cancer [44]. Even if SVM appears to be a natural choice for classification problems there seems to be less existing literature applying it to DNA methylation data in the context of AD identification.

3. Materials And Methods

One of the main objectives of this paper is to be able to accurately generate classification forecasts differentiating between individuals with Alzheimer's disease (AD) and control cases.The algorithm was built with the intention to be easily expandable from one to multiple data sets. A categorical variable y_i was created to classify individuals.

$$y_j = \begin{cases} 0 \ if \ \ Control \\ 1 \ if \ \ AD \end{cases} \tag{1}$$

In this way, a vector $Y = \{Y_1, Y_2, \ldots, Y_{nc}\}$ can be constructed classifying all the existing cases according to the disease estate (control or AD). In this notation nc denotes the total number, including both control and AD, of cases considered. Every case analyzed (j) has an associated vector X^j containing all the methylation levels of each CpG.

$$X^j = \begin{Bmatrix} X^1 \\ X^2 \\ . \\ . \\ . \\ X^{mn} \end{Bmatrix} \tag{2}$$

This notation is used in order to clearly differentiate between the vector (X_j) containing all the methylation data for a single individual (all CpGs) from the vector (X_i) containing all the cases for a given CpG.

$$X_i = \{X_1, X_2, ..., X_{nc}\} \tag{3}$$

In a matrix notation the complete methylation data can be expressed as follows

$$X = \begin{pmatrix} X_1^1 & X_2^1 & ... & X_{nc}^1 \\ X_1^2 & X_2^2 & ... & X_{nc}^2 \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ X_1^{mn} & X_2^{mn} & ... & X_{nc}^{mn} \end{pmatrix} \tag{4}$$

For clarity purposes it is perhaps convenient shoving a hypothetical (oversimplified) example, in which 4 patients $(nc = 4)$ are analyzed (2 control and 2 AD) and that only 5

CpGs were included per patient ($mn = 5$). In this hypothetical example:

$$Y = \{0, 0, 1, 1\} \tag{5}$$

As an example, the methylation data for patient 1 could be:

$$X^1 = \left\{ \begin{array}{c} 0.9832 \\ 0.6145 \\ 0.1254 \\ 0.7845 \\ 0.6548 \end{array} \right\} \tag{6}$$

Similarly, the methylation data for a single CpG for all patients can be expressed as:

$$X_i = \{0.9832, 0.3215, 0.6574, 0.6584\} \tag{7}$$

And the methylation data for all patients (matrix form) would be as follows:

$$X = \begin{pmatrix} 0.9832 & 0.3215 & 0.6574 & 0.6584 \\ 0.6145 & 0.6548 & 0.8475 & 0.7487 \\ 0.1254 & 0.6587 & 0.3254 & 0.6514 \\ 0.7845 & 0.3514 & 0.6254 & 0.6584 \\ 0.6548 & 0.6547 & 0.6587 & 0.6555 \end{pmatrix} \tag{8}$$

The proposed algorithm has two distinct steps. In the first step an initial filtering is carried out. This step reduced the dimensionality of the problem. The second step is the main algorithm. Both steps are described in the following subsections.

Initial filtering

1. $\forall X_i$ estimate a linear regression with Y as the dependent variable. Save the p-value for each X_i.
2. Filter off the X_i with (p-value) < 0.005.

$$\{X_1, X_2, \ldots, X_{mn}\} \rightarrow \{X_1, X_2, \ldots, X_m\} \tag{9}$$

with $m < mn$.

Main algorithm

1. Create a vector grid (D) with the each component representing the dimension (group of X_i) includes in the simulation. Two grids are included, a fine grid with relative small differences in the values of the elements (representing the dimensions that the researcher considers more likely) and a broad grid with large differences in values.

$$Fine\ grid = \{n_1, n_1 + \Delta n_s, n_1 + 2\Delta n_s, \ldots, n_1 + l\Delta n_s\} \tag{10}$$

$$Broad\ grid = \{(n_1 + l\Delta n_s) + \Delta n_l, (n_1 + l\Delta n_s) + 2\Delta n_l, \ldots (n_1 + l\Delta n_s) + p\Delta n_l\}. \tag{11}$$

The values inside the above grids represent the X_i selected. As an example, n_1 represents X_1. Δn_l and Δn_s are the constant step increases in the fine and broad grids, respectively. For instance, $n_1 + \Delta n_l$ and $n_1 + 2\Delta n_l$ are the second and third elements in the fine grid. The actual X_i elements related to this second and third values depend on the actual value of Δn_l. If $\Delta n_l = 1$ then the second and third elements related to X_2 and X_3, respectively, while if $\Delta n_l = 2$, then they relate to X_3 and X_5, respectively. Where $\Delta n_l > \Delta n_s$, each of these values, i.e., $n_1 + \Delta n_s$ is the number of x_i chosen. $l \in \mathbb{Z}^+$ is a constant that specifies (together with n_l) the total size of the fine grid, while $p \in \mathbb{Z}^+$

is the analogous term for the broad grid. For simplicity purposes the case of a fine grid, starting a X_1, followed by a broad grid has been shown but this is not a required constraint. The intent is giving discretion to the researcher to apply the fine grid to the area that is considered more important. This is an attempt to bring the expertise of the researcher into the algorithm. In equation 12 it can be seen the combination of these two grids (D).

$$\begin{aligned} D = \{n_1, n_1 + \Delta n_s, n_1 + 2\Delta n_s, \ldots, n_1 + l\Delta n_s, (n_1 + l\Delta n_s) + \Delta n_l, \\ (n_1 + l\Delta n_s) + 2\Delta n_l, \ldots, (n_1 + l\Delta n_s) + p\Delta n_l\}. \end{aligned} \tag{12}$$

For clarity purposes, let simplify the notation:

$$D = \{S_j\} = \{S_1, S_2, \ldots, S_m\} \tag{13}$$

where Equations (12) and (13) are identical. "S" is a more compact notation with for instance S_1 and S_2 representing n_1 and $n_1 + \Delta n_s$, respectively.

2. Create a mapping between each $x_i = \{X_1, \ldots, X_m\} = \{X_i\}$, where each X_i is a vector, and 10 decile regions. The group of X_i with the highest 10% of the p-value are included in the first decile and assigned a probability of 100%. The group of X_i with the second highest 10% of the p-value are included in the second decile and assigned a probability of 90%. This process is repeated for all deciles creating a mapping.

$$\{X_1, \ldots, X_m\} \rightarrow B\{1.0, 0.9, 0.8, \ldots, 0.1\} \tag{14}$$

Where B is a vector of probabilities. In this way, the X_i with the largest p-values are more likely to be included.

3. For each S_j generate $\forall X_i$, i=1,...,m, a random number R_i with $(0 \leq R_i \leq 1)$. If $R_i > B\{X_i\}$ then X_i is not included in the preliminary S_j group of X_is. Otherwise it is included. In this way a filtering is carried out.

$$\{X_1, \ldots, X_m\} \rightarrow \{X_1, \ldots, X_{m*}\} \; \forall S_j \tag{15}$$

4. Randomly S_j elements of m^* are chosen.
5. Estimate the Hit Ratio (HR)

$$HR = \frac{CE}{TE} \tag{16}$$

where TE is the total number of classification estimations and CE is the number of correct classification estimates.

6. Repeat steps (3) to (6) k times for each S_j. In this way there is a mapping:

$$\{S_1, \ldots, S_m\} \rightarrow \{HR(S_1), \ldots, HR(S_m)\} \tag{17}$$

Remark 1. *An alternative approach would be choosing the starting distribution S_j as the one after which the mean value of the HR does not statistically increase at a 5% confidence level.*

7. Define new search interval between the two highest success rates:

$$max\{HR(S_1), \ldots, HR(_m)\} \rightarrow S^1_{max} \tag{18}$$

$$max\{HR(S_1), \ldots, HR(_m)\} < S^1_{max} \rightarrow \; S^1_{max-1} \tag{19}$$

Iteration 1 (Iter=1) ends, identifying interval:

$$\{S^1_{max}, S^1_{max-1}\} \tag{20}$$

Remark 2. *It is assumed, for simplicity, without loss of generality that* $S^1_{max} < S^1_{max-1}$. *If that it is not the case then the interval needs to be switched* ($\{S^1_{max-1}, S^1_{max}\}$).

8. Divide the interval identified in the previous step into $k-1$ steps.

$$\{S_1, \ldots, S_k\} \tag{21}$$

where $S_1 = S^1_{max}$ and $S_k = S^1_{max-1}$

9. Create a new mapping estimating the new hit rates (following the same approach as in previous steps)

$$\{S_1, \ldots, S_k\} = \{HR(S_1), \ldots, HR(S_k)\} \tag{22}$$

10. Repeat $Iter_t$ times until the maximum number of iterations ($Iter_{max}$) is reached.

$$Iter_t \geq Iter_{max} \tag{23}$$

or until the desire hit rate ($HR_{desired}$) is reached

$$HR(S) \leq HR_{desired} \tag{24}$$

or until no further HR improvement is achieved. Select S^t_{max}.

A few points need to be highlighted. It is important to reduce the number of combinations to a manageable size. For instance, assuming that there are "m" X_i (after the initial filtering of p-Values) there would be $\binom{m}{r}$ combinations of size r. The well known equation (25) can be used.

$$\sum_{r=0}^{m} \binom{m}{r} = 2^m \; \forall m \in \mathbf{N}^+ \tag{25}$$

Assuming that at least one of the X_i is selected:

$$\sum_{r=0}^{m} \binom{m}{r} = \sum_{r=1}^{m} \binom{m}{r} + \binom{m}{0} = 2^m \tag{26}$$

$$\sum_{r=1}^{m} \binom{m}{r} = 2^m - 1 \tag{27}$$

For large m values the -1 term is negligible.
In the initial step the problem of having to calculate the estimations for 2^m combinations is simplified into calculating a $q2^q$ combinations with $q < m$. If for example, $q = m/10$, then the problem is reduced form 2^{10q} to 102^q combinations. It can be proven that:

$$2^{10q} > 10 \cdot 2^q \; \forall q \geq 2 \tag{28}$$

Proof. Using induction. Base case (q=2). $2^{10(k)} = 2^{20} = 1,048,576; 10 \cdot 2^q = 10 \cdot 2^2 = 40$. $1,048,576 > 40$. Therefore, the base case is confirmed. Assume:

$$2^{10k} > 10 \cdot 2^k \; for \; some \; k \geq 2 \tag{29}$$

induction hypothesis

$$2^{10(k+1)} > 10 \cdot 2^{k+1} \tag{30}$$

$$2^{10(k+1)} = 2^{10k}2^{10} > 10 \cdot 2^k 2^{10} = 10 \cdot 2^k 22^9 = 10 \cdot 2^{k+1}2^9 > 10 \cdot 2^{k+1} \tag{31}$$

which completes the proof by induction. □

Data

The methylation data set (Table 1) were obtained from the GEO database and the corresponding accession codes are shown in the table. The methylation data in these two experiments was obtained following similar approaches and both experiments used an Illumina machine. The raw data were structured in a matrix form. For clarity purposes a sample for an specific individual is shown in Table 2. In this table it can be seen the methylation level for all 481,868 CpGs analyzed for a single patient. In the second column it can be seen the identification number for each specific CpG, while in the third column the level of methylation for each specific CpG is shown. Please notice that this is a percentage value ranging from 0 (no methylation) to 1 (fully methylated). Additionally, each patient in the database will be classified according to a binary variable showing if the patient has Alzheimer of if he/she is a healthy control individual. The binary classification variable can be seen in the last row of the table (it is either a 0 or a 1).

Table 1. Methylation data sets included in the analysis.

GEO Code	Cases	Tissue	Illness
GSE66351	190	Glian and neuron	AD and control
GSE80970	286	Pre-frontal cortex and gyrus	AD and control

Table 2. Single patient methylation data.

Number	CpG (Indetifier)	Methylation Level
1	cg13869341	0.89345
2	cg14008030	0.71088
...	...	...
481,868	cg05999368	0.51372
AD/Control		0

Hence, the problem becomes a classification problem, in which the algorithm has to identify how many and which CpGs to use in order to appropriately classify the individuals in the two categories (AD and healthy). A oversimplified sample (not accurate for classification purposes but rather clear for explanation purposes) is shown in Table 3. In this (unrealistic) case only two CpGs were selected for each patient.

Table 3. Single patient methylation data.

Number	CpG (Indetifier)	Methylation Level
2	cg14008030	0.71088
481,868	cg05999368	0.51372
AD/Control		0

It is perhaps easier to conceptualize if the number and the CpG identifier are omitted and several patients are shown (Table 4). This table shows the results (for illustration purposes only) of an unrealistic case, in which the algorithm selects only two CpGs for each patient. Three patient in total are shown, two are control patients and one has AD. This clearly illustrates the objective of the algorithm, which is Selectric the CpGs (rows in this notation) to classify each patient (columns in this notation) according to a binary variable (last row in this notation).

Table 4. Multiple patient methylation data.

Patient 1	Patient 2	Patient 3
0.71088	0.63174	0.72582
0.51372	0.62145	0.43212
0	1	0

In this notation, the Table 4 is the solution generated by the algorithm when presented with the original data of the form shown in Table 5. Table 5 shows all the potential input variables X_i^j (to be selected) where, as previously mentioned, "i" identifies all the potential CpGs per patient and the index "j" identifies the patient. The variable Y_i is the binary variable associated with each patient differentiating between healthy an AD individuals. When expressed in this notation, it is easy to see that the problem boils down to a classification problem, suitable for techniques such as support vector machines.

Table 5. Multiple patient methylation data (general data structure).

Patient 1	Patient 2	Patient 3
X_1^1	X_1^2	X_1^3
X_2^1	X_2^2	X_2^3
...	...	...
$X_{481,868}^1$	$X_{481,868}^2$	$X_{481,868}^3$
Y_1	Y_2	Y_3

4. Results

4.1. Single Data Set

Initially a first estimation using all the available CpGs and a support vector machine classifier was used. The age of the patient (Table 6) was one of the main factors affecting the accuracy of the patient classification using the data set GSE 66351. Controlling for age allowed for better HR rates. Controlling for other variables, such as gender, cell type, or brain region did not appear to improve the classification accuracy . Three different kernels were used (linear, Gaussian, and polynomial), with the best results obtained when using the linear kernel.

Table 6. Hit Rate (HR) of SVM with 3 different kernels for Alzheimer classification (versus control patients), using all the CpGs available (481,778) and controlling for different factors, such as age, gender, cell type, or brain region (GSE 66351 test data).

Controls	HR (Linear)	HR (Gaussian)	HR (Polynomial)	CpGs
None	0.8211	0.7921	0.8167	All
Age	0.8947	0.8142	0.8391	All
Gender	0.8211	0.7921	0.8167	All
Cell type	0.8211	0.7921	0.8167	All
Brain Region	0.8211	0.7921	0.8167	All

In the initial filtering stage the linear regression between each CpGs (X_i) and the vector classification (identifying patients suffering from Alzheimer and control patients was carried out and the p-values stored. CpGs with p-values higher than 0.05 were excluded. The remaining 41,784 CpGs were included in the analysis. It can be seen in Table 7 that as in the previous case controlling for age did improve the HR.The linear kernel was used.

Table 7. HR of SVM for Alzheimer classification (versus control patients), using all CpGs with p-values < 0.05 (41,784) and controlling for different factors, such as age, gender, cell type, or brain region (GSE 66351 test data).

Controls	Hit Rate	CpGs
None	0.7263	41,784
Age	0.8424	41,784
Gender	0.7263	41,784
Cell type	0.7263	41,784
Brain Region	0.7263	41,784

In Figure 1 it is shown that it is possible to achieve high HR using a subset of the CpGs. This HR is higher than the one obtained using all CpGs. As in all the previous cases, the HR rate showed is the out-of-sample HR, i.e., the HR obtained using the testing data that were not used during the training phase. The SVM was trained with approximately 50% of the data contained in the GSE 66351 data set. The testing and training datasets were divided in a manner that roughly maintained the same proportion of control and AD individuals in both datasets. 10-fold cross validation was carried out to try to ensure model robustness. The SVM used linear kernel. The analysis in this figure was carried out controlling for age, gender, cell type and brain region. As in previous cases, the only factor that appears to have an impact on the calculation, besides the level of methylation of the CpGs, was the age. In total, 190 cases of this database was used for either training or testing purposes. The maximum HR obtained was 0.9684, obtained while using 1000 CpGs.

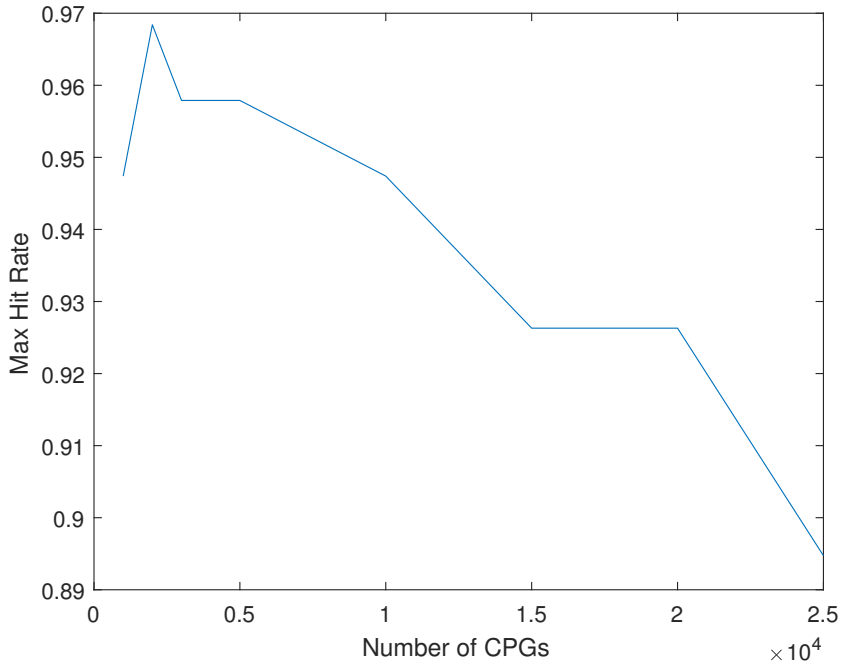

Figure 1. Max Hit Rate (HR) versus number of CpGs included in the analysis

Figure 2 shows the alternative approach mentioned in the methodology, rather than the maximum HR rate obtained the figure shows the average HR obtained at each level(number of CpGS) and its related confidence interval (5%). It is clear from both Figures 1 and 2 that regardless of the approach followed it appears that after a certain amount of CpGs adding additional CpGs to the analysis does not further increased the HR.

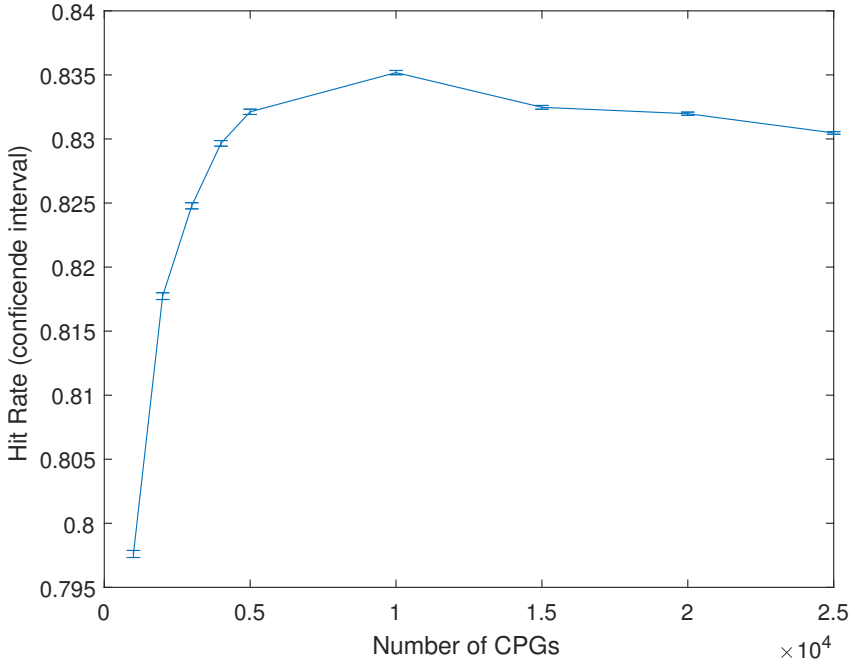

Figure 2. Average Hit Rate (HR) and confidence interval (5%) versus number of CpGs included in the analysis

4.2. Multiple Data Sets

One of the practical issues when carrying out this type of analysis is the lack of consistency between databases, even when there are following similar empirical approaches. As an example, in the case of the GSE66351 dataset a total of 41,784 CpGs were found to be statistically significant (after data pre-processing). Of these 41,784 CpGs only 18.98% (7929) were found to be statistically significant (same *p*-value) in the GSE80970 dataset. This is likely due to subtle different in experimental procedures. In order to overcome this issue only the 7929 CpGs statistically significant CpGs were used when analyzing these two combined datasets. Besides this different pre-filtering step the rest of the algorithm used was as described in the previous section. Both data sets were combined and divided into a training and a test data set.

One of the main differences in the results, besides the actual HR, is that including the age of the patient in the algorithm (using these reduced starting CpG pools) did not appear to substantially increase the forecasting accuracy of the model. The best results when using this approach were obtained when using 4300 CpGs with a combined HR (out of sample) of 0.9202 (Table 8). The list of the 4300 CpGs can be found in the supplementary material.

Table 8. HR of SVM for AD vs. control patients using 4300 CpGs.

Controls	Hit Rate	CpGs
GSE66351	0.8710	4300
GSE80970	0.9517	4300
All	0.9202	4300

Following the standard practice [45] the sensitivity, specificity, positive predictive value (PPV) and negative predictive ratio (NPV) were calculated for all the testing data combined as well as for the testing data in the GSE66351 and GSE80970 separately, Table 9, using the obtained model (4300 CpGs) All the cases included in the analysis are out-of-

sample cases, i.e., not previously used during the training of the support vector machine. It is important to obtain models that are able to generalize well across different data sets.

Table 9. Classification ratios (out-of-sample), including positive predictive value (PPV) and negative predictive ratio (NPV).

Ratio	All	GSE66351	GSE80970
Sensitivity	0.9007	0.8333	0.9506
Specificity	0.9485	0.9394	0.9531
PPV	0.9621	0.9615	0.9625
NPV	0.8679	0.7561	0.9385

5. Discussion

In this paper, an algorithm for the selection of DNA methylation CpG data is presented. A substantial reduction on the number of CpGs analyzed is achieved, while the classification precision is higher than when using all CpGs available. The algorithm is designed to be scalable. In this way, as more data set of Alzheimer DNA methylation become available, the analysis can be gradually expanding. There appear to be substantial differences in the data contained in the data sets analyzed. This is likely due to relatively small experimental procedures. There results obtained (two data sets) are reasonably precise with a sensitivity of 0.9007 and a specificity of 0.9485, while the PPV and the NPV were 0.9621 and 0.8679, respectively. It was also appreciated that when using large amounts of CpGs controlling for age was a crucial steps. However, as the number of CpGs selected by the algorithm decreased, the importance of controlling for age also decreased. Given the large amount of possible combinations of CpGs it is of clear importance to develop algorithm for their selection. As an example, it is clearly not feasible to calculate all the possible combinations of a data set composed by 450,000 CpGs.

The results highlight the necessity to reduce the dimensionality of the data. This is not only in order to facilitate the computations but from a purely statistical point of view, as well. Ideally the number of factors considered should be of the same order of magnitude than the number of samples. In this situation there is a large amount of factors (+450,000) per individual but a relatively small number of individuals. Besides some very specific trails, such as the ongoing SARS-CoV-2 (COVID-19) trials of some vaccines, it is very unlikely to have a cohort of patients and control individuals approaching 450,000. The accuracy of the forecasts increases when the dimensionality of the data are reduced. This is likely due to a reduction of the risk of the algorithm reaching a local minima.

Several methodological decisions were made in order to try to improve the generalization power of the model, i.e., the ability to generate accurate forecast when faced with new data. One of this decisions was to have a large (50%) testing dataset and to have a process that can accommodate for multiple datasets as they become available.

6. Conclusions

Having techniques that can determine if an individual has Alzheimer disease is likely going to become increasingly important. This area of research has, arguably, not received enough attention in the past. This is probably due to the fact that there was no treatment available. This has recently changed, with the FDA approving [46–49] the first drug for the treatment of Alzheimer disease (there were drugs before targeting some of the effects of the illness but not the actual illness itself).

The results, for instance, in Table 9, suggest that the approached followed can generate an accurate forecast (out-of-sample), when using a multi dataset approach, which is a significant development, with, for instance, the sensitivity and the specificity reaching, respectively, 0.9007 and 0.9485 values, when using 4300 CpGs. The obtained positive predictive value (PPV) and the negative predictive value (NPV) were also relatively high, coming in at 0.9621 and 0.8679, respectively. The results also indicate (Figures 1 and 2) that

increasing the number of CpGs does not improve the forecast. This is very likely related to the issue of local minima.

It is also important to remark that, as more data becomes available, the algorithm could be used to classify between healthy and AD patients following a less invasive approach. Most of the currently available methylation data are related to brain tissue that requires an invasive procedure to be obtained. However, methylation datasets in numerous other illnesses already exist, using blood. As blood-based datasets become available, the algorithm presented in this paper can be easily applied to those, potentially becoming an additional practical tool for diagnosis of the illness. There are also several interesting lines of future work. For instance, the addition of new datasets as they become gradually available.

Author Contributions: Conceptualization, G.A.P.; methodology, G.A.P. and J.C.V.; software, G.A.P.; validation, G.A.P. and J.C.V.; formal analysis, G.A.P. and J.C.V.; investigation, G.A.P. and J.C.V.; resources, G.A.P. and J.C.V.; data curation, G.A.P. and J.C.V.; writing—original draft preparation, G.A.P.; writing—review and editing, G.A.P. and J.C.V.; visualization, G.A.P. and J.C.V.; supervision, G.A.P. and J.C.V.; project administration, G.A.P. and J.C.V.; funding acquisition, G.A.P. and J.C.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All the data used in this paper is publicly available at the GEO Database (https://www.ncbi.nlm.nih.gov/geo/, accessed on July 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Olivari, B.S.; Baumgart, M.; Taylor, C.A.; McGuire, L.C. Population measures of subjective cognitive decline: A means of advancing public health policy to address cognitive health. *Alzheimer's Dement. Transl. Res. Clin. Interv.* **2021**, *7*, e12142.
2. Donohue, M.C.; Sperling, R.A.; Salmon, D.P.; Rentz, D.M.; Raman, R.; Thomas, R.G.; Weiner, M.; Aisen, P.S. The preclinical Alzheimer cognitive composite: Measuring amyloid-related decline. *JAMA Neurol.* **2014**, *71*, 961–970.
3. Morris, R.G.; Kopelman, M.D. The memory deficits in Alzheimer-type dementia: A review. *Q. J. Exp. Psychol.* **1986**, *38*, 575–602.
4. Greene, J.D.; Hodges, J.R.; Baddeley, A.D. Autobiographical memory and executive function in early dementia of Alzheimer type. *Neuropsychologia* **1995**, *33*, 1647–1670.
5. Sahakian, B.J.; Morris, R.G.; Evenden, J.L.; Heald, A.; Levy, R.; Philpot, M.; Robbins, T.W. A comparative study of visuospatial memory and learning in Alzheimer-type dementia and Parkinson's disease. *Brain* **1988**, *111*, 695–718.
6. Serrano-Pozo, A.; Frosch, M.P.; Masliah, E.; Hyman, B.T. Neuropathological alterations in Alzheimer disease. *Cold Spring Harb. Perspect. Med.* **2011**, *1*, a006189.
7. Blennow, K.; Hampel, H.; Weiner, M.; Zetterberg, H. Cerebrospinal fluid and plasma biomarkers in Alzheimer disease. *Nat. Rev. Neurol.* **2010**, *6*, 131–144.
8. Hsieh, C.L. Dependence of transcriptional repression on CpG methylation density. *Mol. Cell. Biol.* **1994**, *14*, 5487–5494.
9. Cooper, D.N.; Krawczak, M. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Hum. Genet.* **1989**, *83*, 181–188.
10. Vertino, P.M.; Yen, R.; Gao, J.; Baylin, S.B. De novo methylation of CpG island sequences in human fibroblasts overexpressing DNA (cytosine-5-)-methyltransferase. *Mol. Cell. Biol.* **1996**, *16*, 4555–4565.
11. Gudjonsson, J.E.; Krueger, G. A role for epigenetics in psoriasis: Methylated cytosine–guanine sites differentiate lesional from nonlesional skin and from normal skin. *J. Investig. Dermatol.* **2012**, *132*, 506–508.
12. Cornélie, S.; Wiel, E.; Lund, N.; Lebuffe, G.; Vendeville, C.; Riveau, G.; Vallet, B.; Ban, E. Cytosine-phosphate-guanine (CpG) motifs are sensitizing agents for lipopolysaccharide in toxic shock model. *Intensive Care Med.* **2002**, *28*, 1340–1347.
13. Mikeska, T.; Craig, J.M. DNA methylation biomarkers: Cancer and beyond. *Genes* **2014**, *5*, 821–864.
14. Pidsley, R.; Wong, C.C.; Volta, M.; Lunnon, K.; Mill, J.; Schalkwyk, L.C. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genom.* **2013**, *14*, 1–10.
15. Marabita, F.; Almgren, M.; Lindholm, M.E.; Ruhrmann, S.; Fagerström-Billai, F.; Jagodic, M.; Sundberg, C.J.; Ekström, T.J.; Teschendorff, A.E.; Tegnér, J.; et al. An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. *Epigenetics* **2013**, *8*, 333–346.
16. Kuan, P.F.; Wang, S.; Zhou, X.; Chu, H. A statistical framework for Illumina DNA methylation arrays. *Bioinformatics* **2010**, *26*, 2849–2855.
17. You, L.; Han, Q.; Zhu, L.; Zhu, Y.; Bao, C.; Yang, C.; Lei, W.; Qian, W. Decitabine-mediated epigenetic reprograming enhances anti-leukemia efficacy of CD123-targeted chimeric antigen receptor T-cells. *Front. Immunol.* **2020**, *11*, 1787.

18. Rhee, I.; Jair, K.W.; Yen, R.W.C.; Lengauer, C.; Herman, J.G.; Kinzler, K.W.; Vogelstein, B.; Baylin, S.B.; Schuebel, K.E. CpG methylation is maintained in human cancer cells lacking DNMT1. *Nature* **2000**, *404*, 1003–1007.
19. Feng, W.; Shen, L.; Wen, S.; Rosen, D.G.; Jelinek, J.; Hu, X.; Huan, S.; Huang, M.; Liu, J.; Sahin, A.A.; et al. Correlation between CpG methylation profiles and hormone receptor status in breast cancers. *Breast Cancer Res.* **2007**, *9*, 1–13.
20. Lin, R.K.; Hsu, H.S.; Chang, J.W.; Chen, C.Y.; Chen, J.T.; Wang, Y.C. Alteration of DNA methyltransferases contributes to 5 CpG methylation and poor prognosis in lung cancer. *Lung Cancer* **2007**, *55*, 205–213.
21. Haupt, S.E.; Cowie, J.; Linden, S.; McCandless, T.; Kosovic, B.; Alessandrini, S. Machine learning for applied weather prediction. In Proceedings of the 2018 IEEE 14th International Conference on e-Science (e-Science), Amsterdam, The Netherlands, 29 October–1 November 2018; pp. 276–277.
22. Stefanovič, P.; Štrimaitis, R.; Kurasova, O. Prediction of flight time deviation for lithuanian airports using supervised machine learning model. *Comput. Intell. Neurosci., 2020. https://doi.org/10.1155/2020/8878681.*
23. Rafiee, P.; Mirjalily, G. Distributed Network Coding-Aware Routing Protocol Incorporating Fuzzy-Logic-Based Forwarders in Wireless Ad hoc Networks. *J. Netw. Syst. Manag.* **2020**, *28*, 1279–1315.
24. Roshani, M.; Phan, G.; Roshani, G.H.; Hanus, R.; Nazemi, B.; Corniani, E.; Nazemi, E. Combination of X-ray tube and GMDH neural network as a nondestructive and potential technique for measuring characteristics of gas-oil–water three phase flows. *Measurement* **2021**, *168*, 108427.
25. Pourbemany, J.; Essa, A.; Zhu, Y. Real Time Video based Heart and Respiration Rate Monitoring. *arXiv* **2021**, arXiv:2106.02669.
26. Alfonso, G.; Carnerero, A.D.; Ramirez, D.R.; Alamo, T. Stock forecasting using local data. *IEEE Access* **2020**, *9*, 9334–9344.
27. Joachims, T. *SVM-Light: Support Vector Machine*, version 6.02; University of Dortmund: Dortmund, Germany, 1999.
28. Meyer, D.; Leisch, F.; Hornik, K. The support vector machine under test. *Neurocomputing* **2003**, *55*, 169–186.
29. Wang, L. *Support Vector Machines: Theory and Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2005; Volume 177.
30. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567.
31. Li, X.; Wang, L.; Sung, E. A study of AdaBoost with SVM based weak learners. In Proceedings of the 2005 IEEE International Joint Conference on Neural Networks, Montreal, QC, Canada, 31 July–4 August 2005; Volume 1, pp. 196–201.
32. Magnin, B.; Mesrob, L.; Kinkingnéhun, S.; Pélégrini-Issac, M.; Colliot, O.; Sarazin, M.; Dubois, B.; Lehéricy, S.; Benali, H. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. *Neuroradiology* **2009**, *51*, 73–83.
33. Wang, S.; Lu, S.; Dong, Z.; Yang, J.; Yang, M.; Zhang, Y. Dual-tree complex wavelet transform and twin support vector machine for pathological brain detection. *Appl. Sci.* **2016**, *6*, 169.
34. Fetahu, I.S.; Ma, D.; Rabidou, K.; Argueta, C.; Smith, M.; Liu, H.; Wu, F.; Shi, Y.G. Epigenetic signatures of methylated DNA cytosine in Alzheimer's disease. *Sci. Adv.* **2019**, *5*, eaaw2880.
35. Tost, J. DNA methylation: An introduction to the biology and the disease-associated changes of a promising biomarker. *Mol. Biotechnol.* **2010**, *44*, 71–81.
36. Rauch, T.A.; Wang, Z.; Wu, X.; Kernstine, K.H.; Riggs, A.D.; Pfeifer, G.P. DNA methylation biomarkers for lung cancer. *Tumor Biol.* **2012**, *33*, 287–296.
37. Horvath, S.; Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **2018**, *19*, 371–384.
38. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **2013**, *14*, 1–20.
39. Vidaki, A.; Ballard, D.; Aliferi, A.; Miller, T.H.; Barron, L.P.; Court, D.S. DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. *Forensic Sci. Int. Genet.* **2017**, *28*, 225–236.
40. Mastroeni, D.; Grover, A.; Delvaux, E.; Whiteside, C.; Coleman, P.D.; Rogers, J. Epigenetic changes in Alzheimer's disease: Decrements in DNA methylation. *Neurobiol. Aging* **2010**, *31*, 2025–2037.
41. Grossi, E.; Stoccoro, A.; Tannorella, P.; Migliore, L.; Coppedè, F. Artificial neural networks link one-carbon metabolism to gene-promoter methylation in Alzheimer's disease. *J. Alzheimer's Dis.* **2016**, *53*, 1517–1522.
42. Park, C.; Ha, J.; Park, S. Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset. *Expert Syst. Appl.* **2020**, *140*, 112873.
43. Bhasin, M.; Reinherz, E.L.; Reche, P.A. Recognition and classification of histones using support vector machine. *J. Comput. Biol.* **2006**, *13*, 102–112.
44. Zhao, D.; Liu, H.; Zheng, Y.; He, Y.; Lu, D.; Lyu, C. A reliable method for colorectal cancer prediction based on feature selection and support vector machine. *Med. Biol. Eng. Comput.* **2019**, *57*, 901–912.
45. Lalkhen, A.G.; McCluskey, A. Clinical tests: Sensitivity and specificity. *Contin. Educ. Anaesth. Crit. Care Pain* **2008**, *8*, 221–223.
46. Tanzi, R.E. FDA Approval of Aduhelm Paves a New Path for Alzheimer's Disease. *ACS Chem. Neurosci.* **2021**, *12*, 2714–2715.
47. Karlawish, J.; Grill, J.D. The approval of Aduhelm risks eroding public trust in Alzheimer research and the FDA. *Nat. Rev. Neurol.* **2021**, *17*, 523–524.
48. Ayton, S. Brain volume loss due to donanemab. *Eur. J. Neurol.* **2021**, *28*, e67–e68.
49. Vellas, B.J. The Geriatrician, the Primary Care Physician, Aducanumap and the FDA Decision: From Frustration to New Hope. *J. Nutr. Health Aging* **2021**, *25*, 821–823.