Journal of
Clinical Medicine

MDPI

Article

Neural Network Aided Detection of Huntington Disease

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: Huntington Disease (HD) is a degenerative neurological disease that causes a significant impact on the quality of life of the patient and eventually death. In this paper we present an approach to create a biomarker using as an input DNA CpG methylation data to identify HD patients. DNA CpG methylation is a well-known epigenetic marker for disease state. Technological advances have made it possible to quickly analyze hundreds of thousands of CpGs. This large amount of information might introduce noise as potentially not all DNA CpG methylation levels will be related to the presence of the illness. In this paper, we were able to reduce the number of CpGs considered from hundreds of thousands to 237 using a non-linear approach. It will be shown that using only these 237 CpGs and non-linear techniques such as artificial neural networks makes it possible to accurately differentiate between control and HD patients. An underlying assumption in this paper is that there are no indications suggesting that the process is linear and therefore non-linear techniques, such as artificial neural networks, are a valid tool to analyze this complex disease. The proposed approach is able to accurately distinguish between control and HD patients using DNA CpG methylation data as an input and non-linear forecasting techniques. It should be noted that the dataset analyzed is relatively small. However, the results seem relatively consistent and the analysis can be repeated with larger data-sets as they become available.

Keywords: Huntington disease; DNA methylation; neural networks

Citation: Alfonso Perez, G.; Caballero Villarraso, J. Neural Network Aided Detection of Huntington Disease. *J. Clin. Med.* **2022**, *11*, 2110. https://doi.org/10.3390/jcm11082110

Academic Editor: Vida Abedi

Received: 2 March 2022
Accepted: 8 April 2022
Published: 10 April 2022

1. Introduction

Huntington disease (HD) is a neurological progressive disorder [1–4]. The typical onset of the illness is in mid-adult life [5–7] causing uncontrolled movements as well as declining cognitive and reasoning skills. The disease is associated with a mutation of a gene in Chromosome 4 [8–10] related to the gene encoding for the protein huntingtin [11–13]. There are also other proteins associated with the illness. Vonsattel [14] estimates that death typically occurs approximately 12 to 15 years after the onset of symptoms but some other authors have mentioned a slightly longer period, approximately 15 to 20 years [15,16].

Ross [17] identified three clinical stages of the disease: (1) early-stage, (2) middle-stage and (3) late-state. In the early-stage phase the symptoms are relatively minor with some moderate decrease in motor skills (including some involuntary movements) as well as increased irritability. In the middle-stage phase typically the symptoms are more apparent with a visible decrease in motor and cognitive skills. The late-stage is the third and final stage. In this phase the patient tends to have severe reduction in motor and cognitive skills with in many cases the patient unable to leave the bed or communicate. Regrettably, there is no cure for HD.

Currently there is genetic testing available for HD [18–20], which is typically only carried out when there is significant clinical evidence or family history suggesting the presence of HD. There are also economic costs to take into account when carrying out tests. This paper presents a complementary approach for the detection of HD using DNA methylation data [21–23]. DNA methylation data has been associated with many diseases,

particularly in illnesses such as different types of cancers. DNA methylation analysis is a relatively inexpensive and simple technique.

In simple terms, DNA CpG methylation consists of the addition of a methyl group to a cytosine-phosphate-guanine group as illustrated in Figure 1. DNA methylation is a well-known epigenetic change [24–26]. Current laboratory equipment can quickly analyze more than 450,000 CpGs per patient. It should be noted that the resulting data will consist of a percentage value ranging from 1, meaning that it is fully methylated, to 0, meaning that it is entirely unmethylated. It should also be noted that there is a new generation of equipment that can analyze in excess of 800,000 but this equipment is not yet as widely used as the 450,000 CpGs equipment.

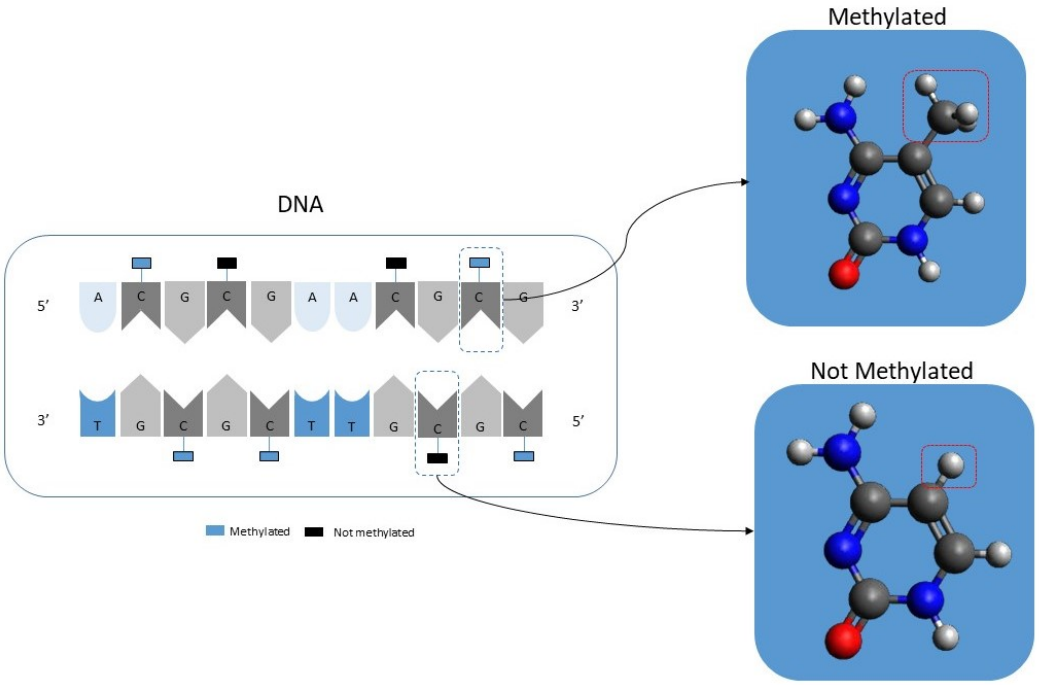

Figure 1. Illustration showing the concept of DNA methylation.

There is a significant amount of literature using DNA CpG methylation data in fields such as aging [27–29], cancer [30–32], Alzheimer [33–35] and Multiple Sclerosis [36]. A common approach in the existing literature is trying to identify relevant CpGs using linear methods. However, in principle there is no indication that the underlying DNA methylation process of aging or of any these illnesses needs to follow a linear behaviour. There are some papers using non-linear methods. For instance, Vidaki [37] analyzed DNA methylation data using neural networks for forensic age purposes. Marchevsky [38] used a similar approach but in this case applied to the classification of different types of lung cancers. In fact, one of the most frequent applications is in the classification of different types of cancers or in differentiating between control and cancer patients [39–44]. This approach has also been applied in the context of some neurological illnesses, such as Alzheimer [45,46].

Huntington disease has attracted less interest in the existing literature than other neurological diseases such as Alzheimer. However, there are some interesting articles exploring the disease in the context of DNA methylation [47–49]. To the best of the knowledge of the authors of this article, the existing literature covering Huntington in the context of DNA methylation follows a linear approach.

2. Aims

One of the main aims of this article is to provide alternative approaches to detect Huntington Disease using available and relatively straightforward techniques based on DNA methylation. Currently there are no treatments for HD but we are relatively optimistic

that eventually there will be some treatment break through. It is acknowledged that there remains significant technical hurdles but when treatments are developed it would be useful to have techniques for screening.

3. Materials and Methods

A classification variable Y_i was defined for each case as follows.

$$y_i = \begin{cases} 0 \ if \ \ Control \\ 1 \ if \ \ Huntington \end{cases} \tag{1}$$

Therefore for m cases analyzed there is a vector Y

$$Y = \{y_1, \ldots, y_m\} \tag{2}$$

There is also an associated vector for each variable y_i containing the methylation levels for n CpGs.

$$X_p = \begin{Bmatrix} X_p^1 \\ X_p^2 \\ . \\ . \\ . \\ X_p^n \end{Bmatrix} \tag{3}$$

Hence, the dataset can be visualized as follows:

$$\begin{pmatrix} y_1 & y_2 & \ldots & y_m \\ X_1^1 & X_2^1 & \ldots & X_m^1 \\ X_1^2 & X_2^2 & \ldots & X_m^2 \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ X_1^n & X_2^n & \ldots & X_m^n \end{pmatrix} \tag{4}$$

The dimensionality of the problem can be defined as n.

3.1. Algorithm

First, the dimensionality of the dataset is reduced. Each CpG is used (individually) as an input for a classification algorithm. The steps are as follows:

1. Select a classification algorithm φ using each CpG (individually) as an input and the classification variable as output $\varphi(X^i, y)$. In this notation X^i refers to the vector containing the methylation data for all the cases analyzed for a single CpG.

$$x^i = \{X_1^i, X_2^i, \ldots, X_m^i\} \tag{5}$$

2. Separate the data into a training and a testing dataset. For clarity purposes the training and testing datasets are labeled A and B respectively.

$$A = \begin{pmatrix} y_1 & y_2 & \cdots & y_k \\ X_1^1 & X_2^1 & \cdots & X_k^1 \\ X_1^2 & X_2^2 & \cdots & X_k^2 \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ X_1^n & X_2^n & \cdots & X_k^n \end{pmatrix} \quad (6)$$

$$B = \begin{pmatrix} y_{k+1} & y_{k+2} & \cdots & y_m \\ X_{k+1}^1 & X_{k+2}^1 & \cdots & X_m^1 \\ X_{k+1}^2 & X_{k+2}^2 & \cdots & X_m^2 \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ X_{k+1}^n & X_{k+2}^n & \cdots & X_m^n \end{pmatrix} \quad (7)$$

3. Train the non-linear algorithm with the training dataset ($\varphi(A)$).
4. Estimate classification forecasts

$$YP = \{YP_{k+1}, \ldots, YP_m\} \quad (8)$$

using the testing dataset and the trained algorithm ($\varphi(B)$)

5. Estimate the accuracy of the forecast (YP) comparing it with the actual values $\{y_{k+1}, y_{k+2}, \ldots, y_m\}$
   (a) For $l = k + 1$ to m

$$if \begin{cases} YP_l = Y_l \; then \; a_l = 1 \\ else \; a_l = 0 \end{cases} \quad (9)$$

   (b) Estimate the accuracy

$$F^i = \left\{ \sum_{l=k+1}^{m} a_l \right\} \frac{1}{(m-k)} \quad (10)$$

6. Repeat steps 2 to 5, k times.
7. Estimate the average of the accuracy $\{F_1^i, \ldots, F_k^i\}$.

$$MF^i = \frac{1}{k} \sum F^i \quad (11)$$

8. Repeat steps 1 to 8 (estimating forecasting accuracy individually for each CpG).

$$MF = \{MF^1, \ldots, MF^n\} \quad (12)$$

9. Define a cut off level (MF_c).
10. Exclude from the analysis all $MF^i < MF_c$.
11. Create a new list of CpGs according to the condition shown in the previous step.

$$MF^{new} = \{MF_*^1, \ldots, MF_*^{nn}\} \; with \; nn \leq n. \quad (13)$$

Note: the dimensionality has been reduced from n to nn.

In the second part of the algorithm a combinatorial approach was followed. The starting point of this second part is the already filtered CpG list with the previously mentioned dimensionality reduction from n to nn. The steps of the second part are as follows:

1. Starting with the reduced list of CpGs. As an example, patient p will now have associated the following CpGs.

$$\begin{Bmatrix} X_p^{*1} \\ X_p^{*2} \\ . \\ . \\ . \\ X_p^{*nn} \end{Bmatrix} \tag{14}$$

Notice again the reduction in the dimensionality from n to nn ($nn < n$).

2. The data, as in the first part of the algorithm, was divided into a training and a testing datasets denoted this time as A^* and B^*.

$$A^* = \begin{pmatrix} y_1 & y_2 & \cdots & y_k \\ X_1^{*1} & X_2^{*1} & \ldots & X_k^{*1} \\ X_1^{*2} & X_2^{*2} & \ldots & X_k^{*2} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ X_1^{*nn} & X_2^{*nn} & \ldots & X_k^{*nn} \end{pmatrix} \tag{15}$$

$$B^* = \begin{pmatrix} y_{k+1} & y_{k+2} & \cdots & y_m \\ X_{k+1}^{*1} & X_{k+2}^{*1} & \ldots & X_m^{*1} \\ X_{k+1}^{*2} & X_{k+2}^{*2} & \ldots & X_m^{*2} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ X_{k+1}^{*nn} & X_{k+2}^{*nn} & \ldots & X_m^{*nn} \end{pmatrix} \tag{16}$$

3. Train the non-linear algorithm with the training dataset ($\varphi(A^*)$).
4. Estimate classification forecasts

$$YP^* = \{YP_{k+1}^*, \ldots, YP_m^*\} \tag{17}$$

using the reduced testing dataset and the trained algorithm ($\varphi(B^*)$).

5. Estimate the accuracy of the forecast (YP^*) comparing it with the actual values $\{Y_{k+1}, Y_{k+2}, \ldots, Y_m\}$

   (a) For $l = k + 1$ to m

$$if \begin{cases} YP_l^* = Y_l \; then \; a_l = 1 \\ else \; a_l = 0 \end{cases} \tag{18}$$

   (b) Estimate the accuracy

$$F^* = \left\{ \sum_{l=k+1}^{m} a_l \right\} \frac{1}{(m-k)} \tag{19}$$

6. Repeat steps 2 to 5, k times.
7. Estimate the average of the accuracy $\{F_1^*, \ldots, F_k^*\}$.

$$MF^* = \frac{1}{k}\sum F^* \tag{20}$$

8. Reduce the number of of CpGs considered by one (randomly selected). Hence, the dimensionality is reduced from nn to nn-1. As an example, the initial reduced CpG list for patient p was:

$$\begin{Bmatrix} X_p^{*1} \\ X_p^{*2} \\ . \\ . \\ . \\ X_p^{*nn} \end{Bmatrix} \tag{21}$$

After this step the new CpG list is:

$$\begin{Bmatrix} X_p^{**1} \\ X_p^{**2} \\ . \\ . \\ . \\ X_p^{**(nn-1)} \end{Bmatrix} \tag{22}$$

9. Repeat steps 2 to 5 with the new CpG list (of dimensionality $nn-1$).
10. Estimate the average (MF^{**}) of the accuracy $\{F_1^*, \ldots, F_k^*\}$.
11. Choose between the previous and the current configuration
    (a) If $MF^{**} > MF^*$, then accept the CpG list used to obtain MF^{**} as the current best list. $MF^{Current} = MF^{**}$.
    (b) If $MF^{**} \leq MF^*$, then reject the CpG list used to obtain MF^{**} and continue using the previous list. $MF^{Current} = MF^*$.
12. Repeat steps 8 to 11 until:
    (a) The number of iterations reaches a predetermined level ($iter_{max}$) or
    (b) $MF^{current} \leq MF_p$, where MF_p is a predetermined acceptable value for the accuracy level.

3.2. Data

DNA methylation data was obtained from the GEO database with the accession code GSE 147004 [49]. The dataset contains DNA methylation data for 76 samples, including 24 control (healthy), 19 HD pre-manifest and 33 HD manifest. The manifest and the pre-manifest sets were grouped together. The dataset contains 485,512 CpG DNA methylation data per patient. The samples were obtained from blood (buffy coat). Age and body fat index data are also available. As previously mentioned the methylation data is expressed as a percentage value (from 0 to 1) with a value of 1 suggesting full methylation. Healthy (control) cases were assigned the categorical variable 0 while HD patients were assigned the categorical variable 1. For clarity purposes some potential values for "A" are shown below.

$$A = \begin{pmatrix} 0 & 0 & \dots & 1 \\ 0.651094 & 0.650451 & \dots & 0.634303 \\ 0.960434 & 0.954877 & \dots & 0.957124 \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ 0.077337 & 0.063247 & \dots & 0.090948 \end{pmatrix} \quad (23)$$

where the values in the first row identify healthy cases (with a "0" categorical value) and HD patients (with a "1" categorical value). All the other rows represent the methylation level of different CpGs expressed as a percentage value. For instance, the second row is associated with one CpG (cg00000029), the third row with a different one (cg00001108) and so on. Some DNA methylation values from an illustrative patient can be seen below.

$$\begin{pmatrix} cg00000029 & 0.651094 \\ cg00000108 & 0.960434 \\ cg00000109 & 0.899284 \\ . & . \\ . & . \\ . & . \end{pmatrix} \quad (24)$$

3.3. Artificial Neural Networks

The classification technique used was an artificial neural network. Neural networks are a flexible approach that have been used successfully in multiple disciplines, including illness identification using DNA methylation data. One of the advantages is that neural networks do not require previous knowledge of the process to be model.It should be noted that the algorithm was constructed in a generic way to allow for the use of other classification techniques. An artificial neural network (ANN) is a well-known technique, inspired by the human brain. The basic component of an ANN is an artificial neuron which in basic terms is a mathematical function translating some input signal into an output signal. The artificial neuron has a related weight associated with it. This weight is a value that it is calibrated during a training phase. There are many training algorithms. The objective of these training algorithms is to minimize the classification error when comparing the actual output value with the output generated by the neural network. Artificial neurons are typically arranged in layers. One critical factor when deciding the architecture of the neural network is to decide the number of layers. In this paper we tested several ANN configurations with the number of layers ranging from 1 to 10. There is no clear definition of the concept of deep learning but it is typically assumed that a neural network with several layers can be considered deep learning. The analysis was carried out, using the standard approach, dividing the dataset in a training dataset and a testing dataset. The training data set contained approximately two thirds of the cases (66.6%) and the testing data set one third (33.3%). Unless otherwise stated the forecasting accuracy refers to than in the testing dataset. Each hidden layer contained 100 sigmoid neurons and the maximum number of iterations was 1000. The analysis was also repeated using only the pre-manifest and control cases (excluding the manifest cases). In this second approach the number of cases is lower. In order to focus on out-of-sample precision the training and the testing data set were divided into two data sets of roughly equal dimensions.

3.4. Similarities and Differences with Previously Published Research

Although they differ quite a bit from our field of application, some authors have also carried out a methodological approach similar to that of our study, having used computer-assisted diagnostic strategies for the detection of neurodegenerative diseases. For this purpose, they have used, for example, the pooled analysis of information from clinical information, such as Lones et al. who designed an algorithm based on the collection of

information related to movement disorders in patients with Parkinson's disease (PD). To this end, they performed continuous monitoring of dyskinesia in six patients with PD using a device that comprised a tri-axial accelerometer and tri-axial gryoscope [50]. Other authors have also carried out machine-learning approaches based on diagnostic imaging information, such as Elahifasaee et al., who designed an algorithm for the classification of diagnostic images compatible with Alzheimer's disease (AD). To do this, they used a feature decomposition and kernel discriminant analysis (KDA) applying it to information from MR brain images from 830 subjects comprising 198 AD patients [51]. Other more recent studies have also carried out a methodological approach more similar to ours, having used strategies based on artificial intelligence for the detection of neurodegenerative diseases, although also based on clinical or neuroimaging information [52,53]. However, very few investigations use this methodology for the design of diagnostic algorithms based on information from molecular studies. Bahado-Singh et al. devised a predictive model for the diagnosis of cerebral palsy using information about DNA epigenetic profiles. These authors are the first to mention the concept of deep learning that we have discussed previously [54]. Something more similar to our research would be the work published a few months ago by Sh et al. because like us, these authors use information from the GEO database. Using a machine-learning model, they have identified the role of natural killer T cells (NKT) and granulocyte macrophage progenitor (GMP) in the aetiology of AD. To do so, they relied on information from mRNA data from blood from 711 subjects, including the control group (238 patients), mild cognitive impairment (189 patients), and AD (284 patients) [55]. Nevertheless, there are no studies with these methodological approaches that are based on epigenetic information and are focused on Huntington's disease (HD), so the present study would be a first in this regard. In accordance with what we have commented on previously, there are studies that use artificial intelligence formulas as a diagnostic resource, but they are based on information from neuroimaging tests [56,57]. Perhaps the closest thing are studies based of genomic information. Lovrecic et al devised a diagnostic algorithm based on the expression of 12 candidate genes [58]. A decade later, the same research group used machine learning techniques to study these genes and discovered that two of them (ARFGEF2 and GOLGA8G) were significantly up-regulated [59]. All the same, as we initially stated, the use of artificial intelligence strategies based on epigenetic information for the diagnosis of HD was an unprecedented topic until nowadays.

4. Results

The results for the first part of the algorithm can be seen in Figure 2. The most accurate classifications were obtained when using a four layers ANN. Further increases in the number of layers did not appear to increase the accuracy of the forecasts. It can be seen that the initial increase in the number of layers did improve the accuracy but after reaching four layers the process seems to have reached a plateau. It should be noted that the computational time required to carry out this analysis was rather substantial. For instance, it required 3.45 days to obtain the results for a one-layer ANN architecture and 10.38 days for a 10-layer architecture. The training process, as shown in Table 1, required significant time. However after training, the application to data from a new patient requires negligible time (a few seconds). The scaled conjugate gradient training algorithm generated better forecasts than other training algorithms such as one-step secant backpropagation or resilient backpropagation. All the calculations were done with an Intel(R) Core(TM) i5-4590 3.3 GHz computer. There are some options to reduce the computational type. For instance, the algorithm was designed in order to make it easily parallelizable, particularly the dimensionality reduction part. This algorithm can be distributed in several computers in a cluster with each computer analyzing a different group of CpGs.

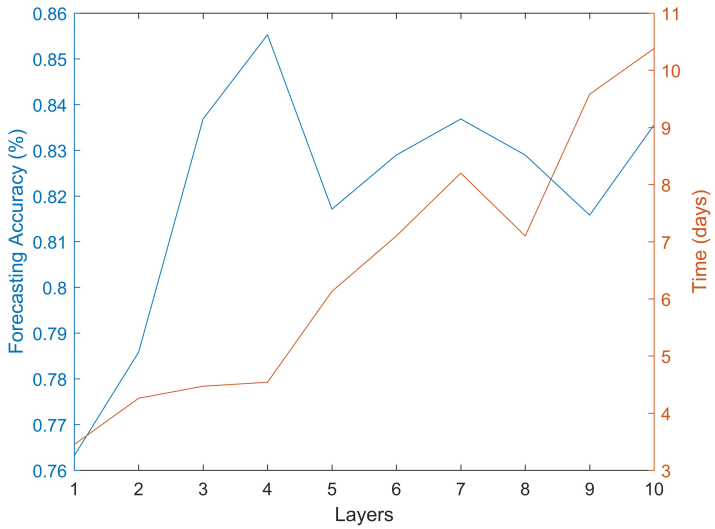

Figure 2. Forecasting accuracy and required computational time using different ANN architectures.

The second part of the algorithm further increased the accuracy of the forecasts. The best results are, similar to the previous case, obtained when using an artificial neural network with four layers (Table 1). Deeper artificial neural networks, such as the one using ten hidden layers, did not improve the results obtained using four layers. The sensitivity and specificity (Table 2) were 0.95 and 0.80 respectively with a final list of 237 CpGs, representing a very substantial reduction from the initial 485,512 available CpGs. The complete 237 CpG list can be found in the supplementary material. Controlling for age and body mass index did not impact the classifications obtained.

Table 1. Forecasting precision obtained with the different neural network configurations (after the second part of the algorithm). The second column shows the results using control, pre-manifest and manifest cases while the third column includes only control and pre-manifest cases. The fourth column shows the computational time required for training the neural network.

N. Layers	Max Precision (Control & Manifest & Pre-Manifest)	Max Precision (Control & Pre-Manifest)	Training Time (Days)
1	0.80	0.76	3.45
2	0.84	0.81	3.78
3	0.88	0.86	4.12
4	0.92	0.81	4.61
5	0.88	0.76	5.82
6	0.88	0.71	6.17
7	0.84	0.71	7.56
8	0.80	0.67	8.43
9	0.80	0.67	9.62
10	0.84	0.62	10.38

Table 2. Forecasting accuracy results.The second column shows the results using control, pre-manifest and manifest cases while the third column includes only control and pre-manifest cases.

Field	Control & Manifest & Pre-Manifest (%)	Control & Pre-Manifest (%)
Correct classification	0.92	0.86
Sensitivity	0.95	0.88
Specificity	0.80	0.80

5. Discussion

Huntington disease is a degenerative illness currently without a cure. However, it is an area of very active research and it is possible that in the future there will be some treatments. Currently there are some specific genetic tests that can identify the illness however they are typically only prescribed when there are clear indications of the illness such as clinical evidences or family history. When treatments become available it is likely that early detection becomes crucially important. In this regard it would be interesting to be able to detect the illness in general blood tests as early as possible. Blood DNA methylation data can be obtained through an inexpensive a relatively quick test that can be carried out and used to test for indications of multiple different illness, such as cancer, and it is likely that in the future this type of test will become more widespread. Using the same basic blood DNA methylation data when testing for other illnesses it may be possible to test for indications of HD as well.

Increasing our understanding of the DNA methylation dynamics in the context of Huntington, such as for instance identifying relevant CpGs as well as improving our search algorithms, can encourage other researchers to obtain more DNA methylation data which in turn can be used to develop more accurate models, in this way creating a positive feedback loop. This is particularly important because while there is a significant existing body of research covering the topic there is much less research than in other degenerative neurological diseases, such as Alzheimer.

From a computational point of view the results show that increasing the complexity of the models beyond a certain point did not translate into an increase in the forecasting accuracy. The best results were obtained using four layers. It is however possible that, using larger datasets, the complexity of the models i.e., the number of layers, might need to be further increased but there is clearly an upper limit. There is also a clear trade-off between the complexity of the model and the required computational time, with some of the models tested requiring in excess of ten days of computing power. Controlling for age and body mass composition did not appear to change the forecasts. However, this might be due to a relatively small data set.

The case of pre-manifest cases was also analyzed independently. It was shown that the accuracy of the classification was relatively high when using only pre-manifest and control cases (excluding HD manifest cases). It should be noted that the accuracy when using this approach (pre-manifest and control only) was high, but lower than that obtained using all cases (control, pre-manifest and manifest), which might be due to a relatively small sample size.

6. Future Research and Limitations

As a line of future research it will be interesting to have access to large data sets that will likely help further improving the accuracy of the model. The relatively small size of the data pool is one of the limitations of this paper. It would be interesting to have reasonably large sets of data at different stages of the illness (not only pre-manifest and manifest) in order to identify the progressions. This systematic, machine-learning driven approach, may prove to be important when comparing different types of potential future medications and their impact on the progression of the illness with quantifiable changes in the level of DNA methylation.

It might be possible to carry out the same type of analysis using some non-invasive biomarkers such as saliva or urine, rather than blood. This will have certain advantages with less discomfort for patients and easier collection. So far we have not found data linking DNA methylation in saliva or urine to HD but it is possible that it can be successfully used to determine the presence of the illness. Based on the experience with other illnesses it is likely that there is a different DNA methylation pattern. This would be another interesting line of future research.

The presented approach to identify relevant combinations of CpGs can be used for other diseases, as long as there is existing DNA methylation data. Similarly, the algorithm

was designed to allow for other training techniques besides artificial neural networks. This is potentially an interesting area of future research.

Another very interesting area of future research is longitudinal analysis. Analyzing DNA methylation changes as the illness progress could be used to quantitatively map the progression of the illness. Another important application of longitudinal analysis, after the above mentioned mapping is created, is as a quantitative measure of the impact of potential treatments in the progression of the illness. This is a very promising field of research but unfortunately there is currently not enough data available to be carried out and would ideally require the monitor of patients over extended periods of time. Longitudinal analysis could potentially greatly help enhancing the knowledge of the progression of the illness. Artificial intelligence techniques, such as neural networks, could be a very interesting tool for analyzing this type of complex and data driven analysis.

7. Conclusions

Huntington disease is a devastating illness. There are several research groups working on potential treatments for this illness but as of now there is no cure. We are cautiously confident that eventually there will be a treatment. As previously mentioned, we do not suggest carry out mass screening at the moment, but when a treatment is developed it will likely be important to have ways to detect the illness, particularly when using general test in patients that might be asymptomatic. It is likely that when such treatment arises early detection will be important. In this scenario, of a treatment available, such a tool could be used as pre-screening with the healthcare professional taking care of the patient to decide if it is appropriate to refer the patient to a specialist or to carry out further testing such as DNA sequencing. In this scenario extreme care should be taken when communicating with the patient, explaining clearly that the test has a degree of uncertainty and that the diagnosis is not yet confirmed. This is, once more, in the context of a potential treatment developed for the illness. The objective is to try to detect the illness as soon as possible (to increase the chances of a successful treatment) while at the same time minimizing the potential physiological impact on the patient.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jcm11082110/s1, Table S1: CpGs.

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.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Access to the data and code is facilitated through a Github repository. https://github.com/Redbluelabel/HD.git. Last accessed on 1 March 2022.

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

References

1. Caron, N.S.; Wright, G.; Hayden, M.R. Huntington Disease. Gene Reviews. 2020. Volume 57. Available online: https://europepmc.org/article/NBK/nbk1305 (accessed on 1 March 2022).
2. Frank, S. Treatment of Huntington's disease. *Neurotherapeutics* **2014**, *11*, 153–160. [CrossRef] [PubMed]
3. Sanberg, P.R.; Coyle, J.T. Scientific approaches to Huntington's disease. *CRC Crit. Rev. Clin. Neurobiol.* **1984**, *1*, 1–44. [PubMed]
4. Sturrock, A.; Leavitt, B. The clinical and genetic features of Huntington disease. *J. Geriatr. Psychiatry Neurol.* **2010**, *1*, 243–259. [CrossRef] [PubMed]

5. Bates, G.P.; Dorsey, R.; Gusella, J.F.; Hayden, M.R.; Kay, C.; Leavitt, B. R.; Nance, M.; Ross, C.A.; Scahill, R.I.; Wetzel, R. Huntington disease. *Nat. Rev. Dis. Prim.* **2015**, *1*, 1–21. [CrossRef] [PubMed]
6. Siemers, E. Huntington disease. *Arch. Neurol.* **2001**, *58*, 308–310. [CrossRef]
7. Evers, M.; Evers, M.; Pepers, B.; Atalar, M.; Van Belzen, M.; Faull, R.; Roos, R.; Van Roon-Mom, W. Making (anti-) sense out of huntingtin levels in Huntington disease. *Mol. Neurodegener.* **2015**, *58*, 1–11. [CrossRef]
8. Thompson, L.; Plummer, S.; Schalling, M.; Altherr, M.; Gusella, J.; Housman, D.; Wasmuth, J. A gene encoding a fibroblast growth factor receptor isolated from the Huntington disease gene region of human chromosome 4. *Genomics* **1991**, *11*, 1133–1142. [CrossRef]
9. Cox, D.R.; Pritchard, C.A.; Uglum, E.; Casher, D.; Kobori, J.; Myers, R.M. Segregation of the Huntington disease region of human chromosome 4 in a somatic cell hybrid. *Genomics* **1989**, *4*, 397–407. [CrossRef]
10. Zuo, J.; Robblns, C.; Bahariloo, S.; Cox, D.; Myers, R. Construction of cosmid contigs and high-resolution restriction mapping of the Huntington disease region of human chromosome 4. *Hum. Mol. Genet.* **1993**, *2*, 889–899. [CrossRef]
11. Cattaneo, E.; Zuccato, C.; Tartari, M. Normal huntingtin function: An alternative approach to Huntington's disease. *Nat. Rev. Neurosci.* **2005**, *6*, 919–930. [CrossRef]
12. Saudou, F.; Humbert, S. The biology of huntingtin. *Neuron* **2016**, *89*, 910–926. [CrossRef] [PubMed]
13. Li, X.; Li, S.; Sharp, A.; Nucifora, F.; Schilling, G.; Lanahan, A.; Worley, P.; Snyder, S.; Ross, C. A huntingtin-associated protein enriched in brain with implications for pathology. *Nature* **1995**, *89*, 398–402. [CrossRef]
14. Vonsattel, J.P.; DiFiglia, M. Huntington disease. *J. Neuropathol. Exp. Neurol.* **1998**, *57*, 369. [CrossRef] [PubMed]
15. Dayalu, P.; Albin, R.L. Huntington disease: Pathogenesis and treatment. *Neurol. Clin.* **2015**, *33*, 101–114. [CrossRef] [PubMed]
16. Ghosh, R.; Tabrizi, S.J. Huntington disease. *Handb. Clin. Neurol.* **2018**, *147*, 255–278.
17. Ross, C.A.; Margolis, R. Huntington disease. *Medicine* **2020**, *76*, 305–338. [CrossRef]
18. Sobel, S.; Cowan, D. Impact of genetic testing for Huntington disease on the family system. *Am. J. Med. Genet.* **2000**, *90*, 49–59. [CrossRef]
19. Nance, M. Genetic testing of children at risk for Huntington's disease. *Neurology* **1997**, *49*, 1048–1053. [CrossRef]
20. Kalman, L.; Johnson, M.; Beck, J.; Berry-Kravis, E.; Buller, A.; Casey, B.; Feldman, G.; Handsfield, J.; Jakupciak, J.; Maragh, S. Development of genomic reference materials for Huntington disease genetic testing. *Genet. Med.* **2007**, *9*, 719–723. [CrossRef]
21. Singal, R.; Ginder, G. DNA methylation. *Blood J. Am. Soc. Hematol.* **1999**, *12*, 4059–4070.
22. Moore, L.; Le, T.; Fan, G. DNA methylation. *Neuropsychopharmacology* **2013**, *38*, 23–38. [CrossRef] [PubMed]
23. Robertson, K. DNA methylation and human disease. *Nat. Rev. Genet.* **2005**, *6*, 597–610. [CrossRef] [PubMed]
24. Bender, J. DNA methylation and epigenetics. *Annu. Rev. Plant Biol.* **2004**, *55*, 41–68. [CrossRef] [PubMed]
25. Holliday, R. DNA methylation and epigenetic inheritance. *Philos. Trans. R. Soc. Lond. Biol. Sci.* **1990**, *326*, 329–338. [CrossRef]
26. Lim, D.; Maher, E. DNA methylation: a form of epigenetic control of gene expression. *Obstet. Gynaecol.* **2010**, *12*, 37–42. [CrossRef]
27. Richardson, B. Impact of aging on DNA methylation. *Ageing Res. Rev.* **2003**, *2*, 245–261. [CrossRef]
28. Jung, M.; Pfeifer, G.P. Aging and DNA methylation. *BMC Biol.* **2015**, *13*, 1–8. [CrossRef]
29. Bell, C.G.; Lowe, R.; Adams, P.D.; Baccarelli, A.; Beck, S.; Bell, J.; Christensen, B.; Gladyshev, V.; Heijmans, B.; Horvath, S. DNA methylation aging clocks: Challenges and recommendations. *Genome Biol.* **2019**, *20*, 1–24. [CrossRef]
30. Das, P.; Singal, R. DNA methylation and cancer. *J. Clin. Oncol.* **2004**, *22*, 4632–4642. [CrossRef]
31. Kulis, M.; Esteller, M. DNA methylation and cancer. *Adv. Genet.* **2010**, *70*, 27–56.
32. Baylin, S. DNA methylation and gene silencing in cancer. *Nat. Clin. Pract. Oncol.* **2005**, *2*, s4–s11. [CrossRef] [PubMed]
33. Mastroeni, D.; Grover, A.; Delvaux, E.; Whiteside, C.; Coleman, P.; Rogers, J. Epigenetic changes in Alzheimer's disease: Decrements in DNA methylation. *Neurobiol. Aging* **2010**, *31*, 2025–2037. [CrossRef] [PubMed]
34. Bollati, V.; Galimberti, D.; Pergoli, L.; Dalla Valle, E.; Barretta, F.; Cortini, F.; Scarpini, E.; Bertazzi, P.A.; Baccarelli, A. DNA methylation in repetitive elements and Alzheimer disease. *Brain Behav. Immun.* **2011**, *25*, 1078–1083. [CrossRef] [PubMed]
35. De Jager, P.; Srivastava, G.; Lunnon, K.; Burgess, J.; Schalkwyk, L.; Yu, L.; Eaton, M.; Keenan, B.; Ernst, J.; McCabe, C. Alzheimer's disease: Early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nat. Neurosci.* **2014**, *17*, 1156–1163. [CrossRef] [PubMed]
36. Alfonso Perez, G.; Caballero Villarraso, J. An Entropy Approach to Multiple Sclerosis Identification. *J. Pers. Med.* **2022**, *12*, 398. [CrossRef]
37. Vidaki, A.; Ballard, D.; Lunnon, K.; Aliferi, A.; Miller, T.; Barron, L.; Court, D.; Keenan, B.; Ernst, J. DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. *Forensic. Sci. Int. Genet.* **2017**, *28*, 225–236. [CrossRef]
38. Marchevsky, A.; Tsou, J.; Laird-Offringa, I. Classification of individual lung cancer cell lines based on DNA methylation markers: use of linear discriminant analysis and artificial neural networks. *J. Mol. Diagn.* **2004**, *6*, 28–36. [CrossRef]
39. Zheng, C.; Xu, R. Predicting cancer origins with a DNA methylation-based deep neural network model. *PLoS ONE* **2020**, *5*, e0226461. [CrossRef]
40. Paluszczak, J.; Baer-Dubowska, W. Epigenetic diagnostics of cancer—The application of DNA methylation markers. *J. Appl. Genet.* **2006**, *47*, 365–376. [CrossRef]
41. Liu, B.; Liu, Y.; Pan, X.; Li, M.; Yang, S.; Li, S. DNA methylation markers for pan-cancer prediction by deep learning. *Genes* **2019**, *10*, 778. [CrossRef]

42. Macias-Garcia, L.; Martinez-Ballesteros, M.; Luna-Romera, J.; Garcia-Heredia, J.; Garcia-Gutierrez, J.; Riquelme-Santos, J. Autoencoded DNA methylation data to predict breast cancer recurrence: Machine learning models and gene-weight significance. *Artif. Intell. Med.* **2020**, *110*, 101976. [CrossRef] [PubMed]
43. Jurmeister, P.; Bockmayr, M.; Seegerer, P.; Bockmayr, T.; Treue, D.; Montavon, G.; Vollbrecht, C.; Arnold, A.; Teichmann, D.; Bressem, K. Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. *Sci. Transl. Med.* **2019**, *11*, eaaw8513. [CrossRef] [PubMed]
44. Paluszczak, J.; Baer-Dubowska, W. Characterizing DNA methylation alterations from the cancer genome atlas. *J. Clin. Investig.* **2014**, *124*, 17–23.
45. 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. [CrossRef]
46. Alfonso Perez, G.; Caballero Villarraso, J. Alzheimer Identification through DNA Methylation and Artificial Intelligence Techniques. *Mathematics* **2021**, *9*, 2482. [CrossRef]
47. Horvath, S.; Langfelder, P.; Kwak, S.; Aaronson, J.; Rosinski, J.; Vogt, T.; Eszes, M.; Faull, R.; Curtis, M.; Waldvogel, H. Huntington's disease accelerates epigenetic aging of human brain and disrupts DNA methylation levels. *Aging* **2016**, *8*, 1485. [CrossRef]
48. De Souza, R.; Islam, S.; McEwen, L.; Mathelier, A.; Mathelier, A.; Mah, S.; Wasserman, W.; Kobor, M.; Leavitt, B. DNA methylation profiling in human Huntington's disease brain. *Hum. Mol. Genet.* **2016**, *10*, 2013–2030. [CrossRef]
49. Lu, A.; Narayan, P.; Grant, M.; Langfelder, P.; Wang, N.; Kwak, S.; Wilkinson, H.; Chen, R.; Chen, J.; Bawden, C. DNA methylation study of Huntington's disease and motor progression in patients and in animal models. *Nat. Commun.* **2020**, *11*, 1–15. [CrossRef]
50. Lones, M.A.; Alty, J.E.; Cosgrove, J.; Duggan-Carter, P.; Jamieson, S.; Naylor, RF.; Turner, A.J.; Smith, S.L. A New Evolutionary Algorithm-Based Home Monitoring Device for Parkinson's Dyskinesia. *J. Med. Syst.* **2017**, *41*, 1–8. [CrossRef]
51. Elahifasae, F.; Li, F.; Yang, M. A Classification Algorithm by Combination of Feature Decomposition and Kernel Discriminant Analysis (KDA) for Automatic MR Brain Image Classification and AD Diagnosis. *Comput. Math. Methods Med.* **2019**, *2019*, *1437123*. [CrossRef]
52. Tautan, A.M.; Ionescu, B.; Santarnecchi, E. Artificial intelligence in neurodegenerative diseases: A review of available tools with a focus on machine learning techniques. *Artif. Intell. Med.* **2021**, *117*, 102081. [CrossRef] [PubMed]
53. Vitale, A.; Villa, R.; Ugga, L.; Romeo, V.; Stanzione, A.; Cuocolo, R. Artificial intelligence applied to neuroimaging data in Parkinsonian syndromes: Actuality and expectations. *Math. Biosci. Eng.* **2021**, *18*, 1753–1773. [CrossRef] [PubMed]
54. Bahado-Sing, R.O.; Vishweswaraiah, S.; Aydas, B.; Mishra, N.K.; Guda, C.; Radhakrishna, U. Deep Learning/Artificial Intelligence and Blood-Based DNA Epigenomic Prediction of Cerebral Palsy. *Int. J. Mol. Sci.* **2019**, *20*, 2075. [CrossRef]
55. Sh, Y.; Liu, B.; Zhang, J.; Zhou, Y.; Hu, Z.; Zhang, X. Application of Artificial Intelligence Modeling Technology Based on Fluid Biopsy to Diagnose Alzheimer's Disease. *Front. Aging Neurosci.* **2021**, *13*, 768229. [CrossRef] [PubMed]
56. Tabrizi, S.J.; Fox, N.C. Automated quantification of caudate atrophy by local registration of serial MRI: Evaluation and application in Huntington's disease. *Neuroimage* **2009**, *47*, 1659–1665.
57. Rizk-Jackson, A.; Stoffers, D.; Sheldon, S.; Kuperman, J.; Dale, A.; Goldstein, J.; Corey-Bloom, J.; Poldrack, R.A.; Aron, A.R. Evaluating imaging biomarkers for neurodegeneration in pre-symptomatic Huntington's disease using machine learning techniques. *Neuroimage* **2011**, *56*, 788–796. [CrossRef]
58. Lovrecic, L.; Kastrin, A.; Kobal, J.; Pirtosek, Z.; Krainc, D.; Peterlin, B. Gene expression changes in blood as a putative biomarker for Huntington's disease. *Mov. Disord.* **2009**, *25*, 2277–2281. [CrossRef]
59. Lovrecic, L.; Slavkov, I.; Dzeroski, S.; Peterlin, B. ADP-ribosylation factor guanine nucleotide-exchange factor 2 (ARFGEF2): A new potential biomarker in Huntington's disease. *J. Int. Med. Res.* **2010**, *38*, 1653–1662. [CrossRef]