

Bioinformatic approaches for  
identification of epigenetic profiles  
in neuropathology processes

Aproximaciones bioinformaticas  
para identificacion de perfiles  
epigeneticos en procesos  
neuropatologicos

By

GERARDO ALFONSO PÉREZ

DIRECTOR: DR. JAVIER CABALLERO VILLARRASO



UNIVERSIDAD DE CÓRDOBA

PhD Biomedicine

UNIVERSIDAD DE CORDOBA

JUNE 2022

TITULO: *Bioinformatic approaches for identification of epigenetic profiles in neuropathology processes*

AUTOR: *Gerardo Alfonso Perez*

---

© Edita: UCOPress. 2022  
Campus de Rabanales  
Ctra. Nacional IV, Km. 396 A  
14071 Córdoba

<https://www.uco.es/ucopress/index.php/es/>  
[ucopress@uco.es](mailto:ucopress@uco.es)

---



**TÍTULO DE LA TESIS:** Bioinformatics approaches for identification of epigenetic profiles in neuropathology processes

**DOCTORANDO:** Gerardo Alfonso Pérez

#### **INFORME RAZONADO DEL/DE LOS DIRECTOR/ES DE LA TESIS**

Durante el desarrollo de la presente Tesis Doctoral, en el periodo comprendido entre noviembre de 2020 y junio de 2022, el doctorando Gerardo Alfonso Pérez además de superar sobradamente los objetivos inicialmente demarcados ha ido mucho más allá y ha diseñado, desarrollado y probado herramientas computacionales destinadas a la identificación de perfiles moleculares identificativos de enfermedades neurodegenerativas. Ha creado algoritmos no lineales basados en técnicas de inteligencia artificial para la identificación de pacientes con entidades neurodegenerativas tales como la enfermedad de Alzheimer y el corea de Huntington, combinando datos de metilación de DNA (CpG) con técnicas de inteligencia artificial. También ha propuesto novedosos algoritmos bioinformáticos, como la utilización de la entropía de Shannon en la selección de CpGs aplicada a la detección de perfiles epigenéticos característicos de la esclerosis múltiple. Esto ha dado lugar a tres artículos en revistas indexadas en el primer cuartil del 'Journal Citation Reports', uno de ellos en primer decil. De esto partió el planteamiento de presentar el trabajo de Tesis Doctoral como compendio de artículos.

Además, ha presentado ocho comunicaciones en congresos internacionales y dos ponencias en jornadas nacionales. En dichas comunicaciones ha demostrado la proyección de estos nuevos algoritmos computacionales a otras enfermedades neurológicas como el Parkinson o la enfermedad de Creutzfeldt-Jakob, así como a trastornos psiquiátricos (como esquizofrenia, trastorno bipolar y depresión).

Cabe reseñar que la idea, temática y metodología del presente trabajo de Tesis Doctoral fueron iniciativa del doctorando. Tal hecho, unido al tiempo de ejecución inicialmente reseñado, hacen inferir la iniciativa, creatividad y gran capacidad de trabajo del aludido doctorando.

Por todo ello, se autoriza la presentación de la tesis doctoral.

Córdoba, 7 de junio de 2022

Firma del director

CABALLERO VILLARRASO JAVIER - 30541937F	Firmado digitalmente por CABALLERO VILLARRASO JAVIER - 30541937F Fecha: 2022.06.07 19:31:18 +02'00'
---	---



## ABSTRACT

**D**egenerative neurological diseases, such as Alzheimer, Multiple Sclerosis or Huntington Disease, are illnesses that are not well-known while at the same time having a significant impact on the quality of life of the patients and their survival. The focus of this dissertation is finding biomarkers for the identification of these diseases, ideally in a rapid a reliable manner. The analysis was carried out using DNA CpG methylation data. In recent years there has been very significant technological improvements. It is currently possible to obtain the methylation levels for hundreds of thousands of CpG in a patient in a fast and reliable manner. It is however challenging to analyze these amounts of new data. A reasonable approach to tackle this issue is using machine learning techniques that have proven useful in many other fields. In this dissertation I developed a nonlinear approach to identifying combinations of CpGs DNA methylation data, as biomarkers for Alzheimer (AD) disease. It will be shown that this approach increases the accuracy of the detection on patients with AD when compared to directly using all the data available. I also analyzed the case of Huntington Disease (HD). Using nonlinear techniques I was 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. Additionally, in this dissertation I present a technique, based on the concept of Shannon Entropy, to select CpGs as inputs for non-linear classification algorithms. It will be shown that this approach generates accurate classifications that are a statistically significant improvement over using all the data available or randomly selecting the same number of CpGs. The results seems to clearly illustrate that the analysis of the DNA methylation data, for the identification of patients suffering from the degenerative neurological diseases above mentioned, needs to be carefully carry out. Having the possibility of analyzing hundreds of thousands of CpGs level does not necessarily trans-

---

late into better results as some of these levels might be unrelated and only adding noise to the analysis. It will be shown that the proposed algorithms generate accurate results while at the same time decreasing the number of CpGs used. For instance, in the case of Alzheimer the results obtained with the proposed algorithm generate a sensitivity of 0.9007 and a specificity of 0.9485. One of the underlying expectations is that in the future there will be curative treatments for these illnesses, which do not currently exist. It is also assumed that early detection, similarly to many other diseases, might be important when such treatments appear. Using the current technology it is relatively simple to analyze DNA methylation data and hence it can become an interesting biomarker in the context of these illnesses.

## RESUMEN

Las enfermedades neurológicas degenerativas, como el Alzheimer, la Esclerosis Múltiple o la Enfermedad de Huntington son enfermedades que aún no son del todo conocidas y, al mismo tiempo, tienen un gran impacto en la calidad de vida del paciente y en su supervivencia. El enfoque de esta tesis es encontrar biomarcadores para la identificación de estas enfermedades, idealmente de una manera rápida y precisa. El análisis se llevó a cabo utilizando datos de metilación de ADN CpG. En los últimos años se han producido mejoras tecnológicas muy significativas. Actualmente es posible obtener los niveles de metilación para cientos de miles de CpG en un paciente de una manera rápida y confiable. Sin embargo, es difícil analizar estas cantidades de nuevos datos. Un enfoque razonable para abordar este problema es el uso de técnicas de aprendizaje automático que han demostrado ser útiles en muchos otros campos. En esta tesis doctoral desarrollé un enfoque no lineal para identificar combinaciones de datos de metilación del ADN (CpGs), como biomarcadores para la enfermedad de Alzheimer (EA). Se demostrará que este algoritmo aumenta la precisión de la detección en pacientes con EA en comparación con el uso directo de todos los datos disponibles. También analicé el caso de la enfermedad de Huntington (EH). Usando técnicas no lineales pude reducir el número de CpG considerados de cientos de miles a 237 utilizando también un enfoque no lineal. Se demostrará que el uso de solo estos 237 CpG y técnicas no lineales como las redes neuronales artificiales permite diferenciar con precisión entre pacientes de control y EH. Adicionalmente, en esta tesis presento una técnica, basada en el concepto de Entropía de Shannon, para seleccionar CpGs como entradas para algoritmos de clasificación no lineal. Se demostrará que este enfoque genera clasificaciones precisas con una mejora estadísticamente significativa sobre el uso de todos los datos disponibles o la selección aleatoria del mismo número de CpG.

Los resultados parecen ilustrar claramente que el análisis de los datos de metilación del ADN, para la identificación de pacientes que sufren de la enfermedad neurológica degenerativa antes mencionada, debe llevarse a cabo cuidadosamente. Tener la posibilidad de analizar cientos de miles de

---

niveles de CpG no necesariamente se traduce en mejores resultados, ya que algunos de estos niveles pueden no estar relacionados y solo agregar ruido al análisis. Se demostrará que los algoritmos propuestos generan resultados precisos y, al mismo tiempo, disminuyen el número de CpG utilizados. Por ejemplo, en el caso del Alzheimer los resultados obtenidos con el algoritmo propuesto generan una sensibilidad de 0,9007 y una especificidad de 0,9485. Una de las expectativas subyacentes es que en el futuro habrá tratamientos curativos para estas enfermedades, que actualmente no existen. También se supone que la detección temprana, de manera similar a muchas otras enfermedades, podría ser importante cuando aparecen tales tratamientos. Utilizando la tecnología actual, es relativamente simple analizar los datos de metilación del ADN y, por lo tanto, puede convertirse en un biomarcador interesante en el contexto de estas enfermedades.



## DEDICATION

I would like to thank my family for the patient and support provided during the process of doing my PhD. Particularly two my two kids, Leo and Alyssa, which can always make me smile, Wendy for been so supportive and my parents, Antonio y Pilar, for their encouragement and nice words of support.



## ACKNOWLEDGEMENTS

I would like to sincerely thank my PhD supervisor Prof. Javier Caballero Villarraso for the help and support provided during the difficult task of doing a PhD. I would not have been possible to do it without his encouragement. It has been a true pleasure working with him.



# TABLE OF CONTENTS

	<b>Page</b>
<b>List of Figures</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Degenerative neurological diseases . . . . .	1
1.1.1 Alzheimer Disease (AD) . . . . .	2
1.1.2 Huntington Disease (HD) . . . . .	5
1.1.3 Multiple Sclerosis (MS) . . . . .	6
1.2 Machine Learning . . . . .	8
<b>2 Hypothesis and objectives</b>	<b>11</b>
2.1 Hypothesis . . . . .	11
2.2 Conceptual hypothesis . . . . .	11
2.3 Operating hypothesis . . . . .	12
2.4 Objectives . . . . .	13
<b>3 Contributions</b>	<b>15</b>
3.1 An Entropy Approach to Multiple Sclerosis Identification . .	16

TABLE OF CONTENTS

---

3.2 Alzheimer Identification through DNA Methylation and Artificial Intelligence Techniques . . . . .	29
3.3 Neural Network Aided Detection of Huntington Disease . . .	44
<b>4 Discussion</b>	<b>59</b>
<b>5 Future investigations</b>	<b>63</b>
<b>6 Conclusions</b>	<b>65</b>
<b>7 List of contributions</b>	<b>69</b>
<b>Bibliography</b>	<b>83</b>

## LIST OF FIGURES

<b>FIGURE</b>	<b>Page</b>
1.1 Representation of a healthy brain (left) and a brain with AD (right).	3
1.2 Representation MS neurological damage. Top (healthy), Bottom (MS). . . . .	7
1.3 Representation of an artificial neuron. . . . .	9





## INTRODUCTION

**1.1 Degenerative neurological diseases**

**D**egenerative neurological diseases have a significant impact on the quality of life of patients as well as on their survival. Several of this type of disease have currently no curative therapy [59, 68, 77] but there is a very significant amount of research currently carried out. Furthermore, as the population, due to advances in medicine and sanitation among others, tends to live longer some of these illnesses are likely to appear more frequently [21, 58]. A good example of this is Alzheimer Disease, which is a paradigm of age-related disease [16, 69]. There is clearly a wide range of different degenerative neurological diseases with different causes and prognosis. For several of these disease there is a genetic as well as environmental factors, like in the case of AD, causing the illness. In this

dissertation I focus on three different degenerative neurological diseases:

- Alzheimer Disease (AD)
- Huntington Disease (HD)
- Multiple Sclerosis (MS)

These three illnesses have a high mortality rates and typically cause a significant impact on the quality of life of the patient (depending on the stage of the illness).

### **1.1.1 Alzheimer Disease (AD)**

Alzheimer disease is the most common type of dementia [64] with some estimates suggesting that it represent from 60% to 80% of all dementia [41]. The illness receives its name from the doctor Alois Alzheimer, which first described the illness in 1906 [22, 63, 97]. AD currently has no cure [44, 46] and has a significant impact on the quality of life of the patient [64]. Most of the prescribed medications to AD patients are to manage symptoms[55]. AD is likely caused by a combination of genetic and environmental factors [13, 78, 90]. Tanzi et al. concluded that genetic factors to be significant in 80% of the cases [85].

Some of the most visible symptoms of AD is the memory deterioration [39, 47, 94], particularly short-term memory, as well as cognitive impairment. AD causes deterioration on brain cells. More specifically, Serrano-Pozo et al. [74] mentioned the following lessons as some of the most frequently

## 1.1. DEGENERATIVE NEUROLOGICAL DISEASES

---

associated with AD: 1) amyloid plaques, 2) cerebral amyloid angiopathy, 3) Neurofibrillary tangles and 4) Neuronal and synaptic loss.

Amyloid deposits have been long associated with AD [52]. The actual mechanics of the relationship between deposits and AD remain not fully understood [73]. Frauschy et al. [36] realized in vivo experiments in rats injecting amyloids into the cortex and hippocampus. These experiments show a clear neuronal response to the amyloid. In figure 1.1 it can be seen a graphical representation of brain deterioration due to AD. The left part represents a healthy brain while the right side represents the brain of a patient with AD. It can be seen in the figure that the brain undergoes physical changes, such as shrinking.

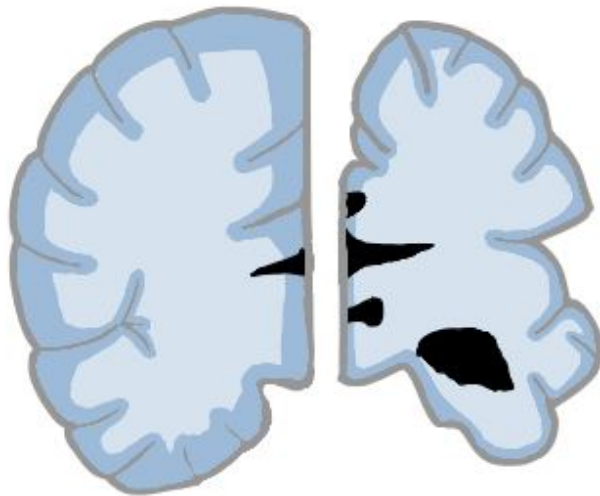


Figure 1.1: Representation of a healthy brain (left) and a brain with AD (right).

It is common to make the distinction between early onset AD, for patients

## CHAPTER 1. INTRODUCTION

---

that are <65 years old, and the more common late onset AD for patients over 65 years old [62]. The symptoms in early onset and late onset AD might be different. For instance, Koedam et al. [53] analyzed 270 patients with early onset AD and 90 patients with late onset AD. The authors classified the patients in two categories, one defined as memory presentation and the other as non-memory presentation. They concluded that 33% of the patients in the early onset group were classified as non-memory presentation while only 6% of the late onset AD patients were classified in the same group.

### **1.1.2 Huntington Disease (HD)**

Huntington Disease (HD) is a fatal [49, 61, 72] neurodegenerative disease [83] which causes cognitive deterioration [3, 67], movement disorders as well as psychiatric disorders. Perhaps some of the most recognizable symptoms are movement disorders such as involuntary movements [50, 79] as well as muscular rigidity [17, 18]. There are significant amounts of research on the illness but currently there is no curative therapy [71]. It is understood that the protein huntingtin [37, 45] plays an important role. A mutation consisting of the repetition of a CAG trinucleotide seems to be the cause of the illness [8, 84]. HD causes the progressive degeneration on brain cells which eventually leads to death. Some of the regions most affected by HD are the basal ganglia, hypothalamus and brain stem cells [23].

Craufurd et al. [24] studied behavioral changes in HD patients establishing that some behavioral changes are much more frequent than others. The authors mentioned that low energy, poor quality of work and impaired judgment were among the most common behavioral changes. They also found that depression and irritability occurred in approximately half of the cases. Psychotic episodes were rare. HD is more frequent among individuals of European descent [27].

### 1.1.3 Multiple Sclerosis (MS)

Multiple Sclerosis (MS) is a severe neurodegenerative, chronic, autoimmune disease [40] that is characterized by having a deterioration in myelin (secondary to damage to Schwann cells), see figure 1.2. It is among the most common non-traumatic reasons of disability among young adults [29]. The illness is due to genetic [31] and environmental factors[32]. Sospedra et al. mentioned that the illness likely requires an environmental insult [80]. The symptoms and evolution of MS differs greatly from patient to patient with a vast array of manifestations [51]. Some of the most common symptoms include weakness, tremors and poor coordination. Typically symptoms [10] are described as:

- Primary – Directly related to the illness, such as weakness.
- Secondary – Related to the primary symptoms, such as Infections.
- Tertiary – due to social and psychological factors, such as depression.

Even though the illness has attracted a significant amount of research there is no curative therapy for MS [43, 60, 93]. As with other neurodegenerative disease prescribed treatments typically focus on managing the symptoms or stop its progression [25, 38]. It is common for some patients to experienced periods of remission, which is some cases might last long periods of time. Cycles of relapse-remission are also very common [70]. Steinman [82] describes how 80% of the patients experience periodic relapses-remissions (at the early stages). It is also not uncommon that patients that experience a relapse-remission type of MS evolve to more debilitating type of MS [35, 81].

## 1.1. DEGENERATIVE NEUROLOGICAL DISEASES

---

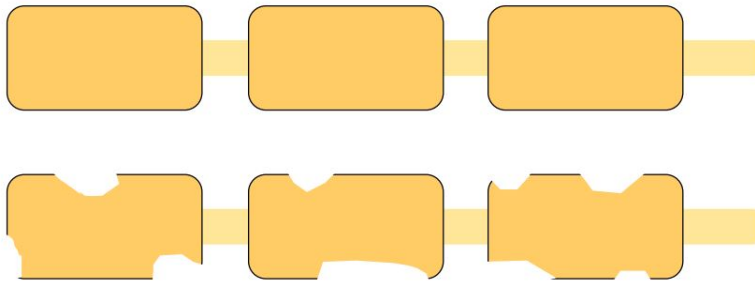


Figure 1.2: Representation MS neurological damage. Top (healthy), Bottom (MS).

The evolution of the illness, such as for instance what triggers periods or remission of how long will it last, are not well understood.

## 1.2 Machine Learning

Machine learning techniques [33, 48] are an increasingly popular set of tools with applications in many fields [11, 19, 56]. They can be used for several different purposes, including time series forecasting and classification purposes [1, 42, 65, 75]. Artificial neural networks (ANN) [5, 14] are a subset of these techniques. ANN are a biologically inspired algorithm [15, 34]. The basic component of an ANN is an artificial neuron [7]. Artificial neurons can be understood as a mathematical function that generate an output when provided with an input. Artificial neurons also have a related weight, as seen in figure 1.3. In the common practice of supervised learning [26, 57] this weight is iteratively modify in order to make the generated output as close as possible to the target output. Normally, an artificial neural networks has many artificial neurons and they are usually grouped in layers [54, 92]. The above mentioned supervised learning approach is the approach followed in this dissertation when trying to ascertain if a certain individual is healthy or present a certain neurodegenerative disease. In this approach the data is divided into two groups a training and a testing dataset [6, 28]. The training dataset is sued to train the neural network. During this process the weights of the individual neurons are iteratively changed to try to generate a binary output for the overall ANN (either “0” for healthy individual or “1” for patients with the disease) as accurate as possible. The training phase is done iteratively until either a certain acceptable level for the error is reached or a predefined maximum number of iterations is reached [86, 91]. After the training phase is complete i.e., the weights of the



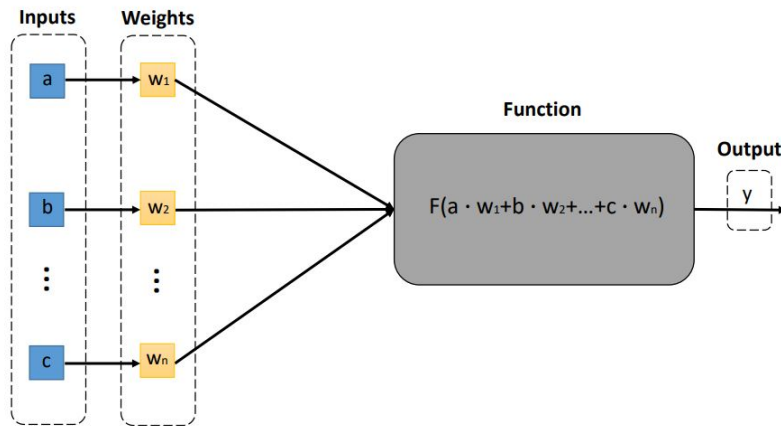


Figure 1.3: Representation of an artificial neuron.

ANN are already obtained. Then the ANN is tested with the testing dataset. It is important to mention that the testing dataset is not used during the training phase. Some measure of the error rate is then obtained for the testing dataset. This error rate in the testing dataset is a more accurate description of the real accuracy of the network as there could be overfitting [12, 66] in the training dataset.

As in any other techniques there are advantages and disadvantages [20, 30, 87] on using machine learning techniques, such as neural networks. One of the main advantages is that they are flexible [4, 76] and can be applied to a large number of different problems with accurate result [2, 9, 88]. In basic terms the only requirements, in principle, is having an underlying process with an input and an output signal. ANN do not require in depth knowledge of the underlying process [95]. The most frequently mentioned disadvantage is that these techniques can be difficult to interpret [28, 89, 96]. The forecast and classification estimations obtained by the

## CHAPTER 1. INTRODUCTION

---

ANN are relatively straightforward to be understood. However, the actual model i.e. the artificial neural network with the optimized weights, is likely going to be a rather complex mathematical model that might not be easy to be understood by the researcher.

## HYPOTHESIS AND OBJECTIVES

**I**n this chapter I present the hypothesis underlying this dissertation as well as the objectives.

### **2.1 Hypothesis**

### **2.2 Conceptual hypothesis**

Biomarkers for neurodegenerative diseases (such as Alzheimer disease, Huntington disease and multiple sclerosis) can be built using DNA CpG methylation data. Given the large amount of DNA CpG methylation data available a machine learning approach is suitable to analyze the data. Neural networks are a viable machine learning technique to analyze DNA CpG methylation data. These machine learning techniques can differentiate

between patients suffering from a neurodegenerative disease and controls patients using DNA CpG data input.

### **2.3 Operating hypothesis**

Reducing the dimensionality of the input data (DNA Methylation data) i.e., reducing the number of CpGs used in the analysis can increase the accuracy of the classification. The underlying process that enables the identification of patients vs. control individuals is not necessarily linear. Data needs to be divided into training and testing dataset when using neural networks in order to avoid issues such as overfitting. An excessively large amount of DNA CpG methylation data can introduce noise in the analysis, as not all CpG methylation levels will be relevant for illness identification.

## 2.4 Objectives

There are several objectives in this dissertation:

- Objective 1. One of the main aims of this dissertation is to provide alternative approaches to detect neurodegenerative diseases, such as Alzheimer Disease, Huntington Disease and Multiple Sclerosis using DNA methylation data as input and machine learning techniques as the data processing tool. This objective was achieved as illustrated in Chapter 3 (sections 3.1, 3.2 and 3.3).
- Objective 2. Another important objective in this dissertation is to show the importance of reducing the dimensionality of the input data i.e., using less CpGs. The objective is to show that appropriate reduction of the dimensionality of the data can increase the accuracy of the classification forecasts, differentiating between patients and control individuals. This objective was also achieved and illustrated in Chapter 3 (sections 3.1, 3.2 and 3.3).
- Objective 3. Developing algorithms for the selection of CpGs using non-linear techniques that can generate classification forecasts more accurate than using all the available CpGs. This result was also achieved and it is illustrated in Chapter 3 (sections 3.1, 3.2 and 3.3).
- Objective 4. Apply the concept of Shannon Entropy for CpG selection purposes, showing that it generates better forecasts than a direct approach using all information available. This aim was also achieved in Chapter 3 (section 3.1).



CHAPTER



## CONTRIBUTIONS

In this chapter it can be seen three (Q1) papers already published in peer review journals.

## **3.1 An Entropy Approach to Multiple Sclerosis Identification**

Authors: Gerardo Alfonso Perez, Javier Caballero Villarraso

Journal of Personalized Medicine. 2022, 12(3), 398.

<https://doi.org/10.3390/jpm12030398>

Current Impact Factor: 4.945

5-year Impact Factor: 4.994

JCR category rank:

Q1: Medicine, General & Internal

Q1: Health Care Sciences & Services



## Article

# An Entropy Approach to Multiple Sclerosis Identification

Gerardo Alfonso Perez <sup>1,\*</sup> and Javier Caballero Villarraso <sup>1,2</sup><sup>1</sup> Department of Biochemistry and Molecular Biology, University of Cordoba, 14071 Cordoba, Spain; bc2cavij@uco.es<sup>2</sup> Biochemical Laboratory, Reina Sofia University Hospital, 14004 Cordoba, Spain

\* Correspondence: ga284@cantab.net

**Abstract:** Multiple sclerosis (MS) is a relatively common neurodegenerative illness that frequently causes a large level of disability in patients. While its cause is not fully understood, it is likely due to a combination of genetic and environmental factors. Diagnosis of multiple sclerosis through a simple clinical examination might be challenging as the evolution of the illness varies significantly from patient to patient, with some patients experiencing long periods of remission. In this regard, having a quick and inexpensive tool to help identify the illness, such as DNA CpG (cytosine-phosphate-guanine) methylation, might be useful. In this paper, a technique is presented, based on the concept of Shannon Entropy, to select CpGs as inputs for non-linear classification algorithms. It will be shown that this approach generates accurate classifications that are a statistically significant improvement over using all the data available or randomly selecting the same number of CpGs. The analysis controlled for factors such as age, gender and smoking status of the patient. This approach managed to reduce the number of CpGs used while at the same time significantly increasing the accuracy.

**Keywords:** multiple sclerosis; DNA methylation; entropy



**Citation:** Alfonso Perez, G.; Caballero Villarraso, J. An Entropy Approach to Multiple Sclerosis Identification. *J. Pers. Med.* **2022**, *12*, 398. <https://doi.org/10.3390/jpm12030398>

Academic Editors: Youxin Wang and Ming Feng

Received: 21 January 2022

Accepted: 3 March 2022

Published: 4 March 2022

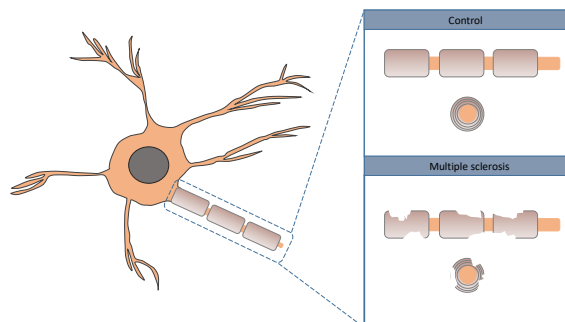
**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Multiple sclerosis (MS) is a chronic autoimmune illness affecting the brain and spinal cord associated with various degrees of disability. In MS, the immune system of the patient attacks the axons, more specifically, the myelin cover; see Figure 1 for a graphical illustration [1]. Inflammation is highlighted by some researchers as one of the drivers of neurodegeneration in MS [2–4]. The evolution of the illness varies greatly from patient to patient, with some individuals experiencing long periods of remissions due to mechanisms that are not yet well understood. The usual manifestation age of the illness is from 20 to 45 years old, but it can occasionally manifest at younger ages, even in children [5]. The causes of MS remain unclear, with a complex underlying combination of genetic and environmental factors the most likely cause [6–10].

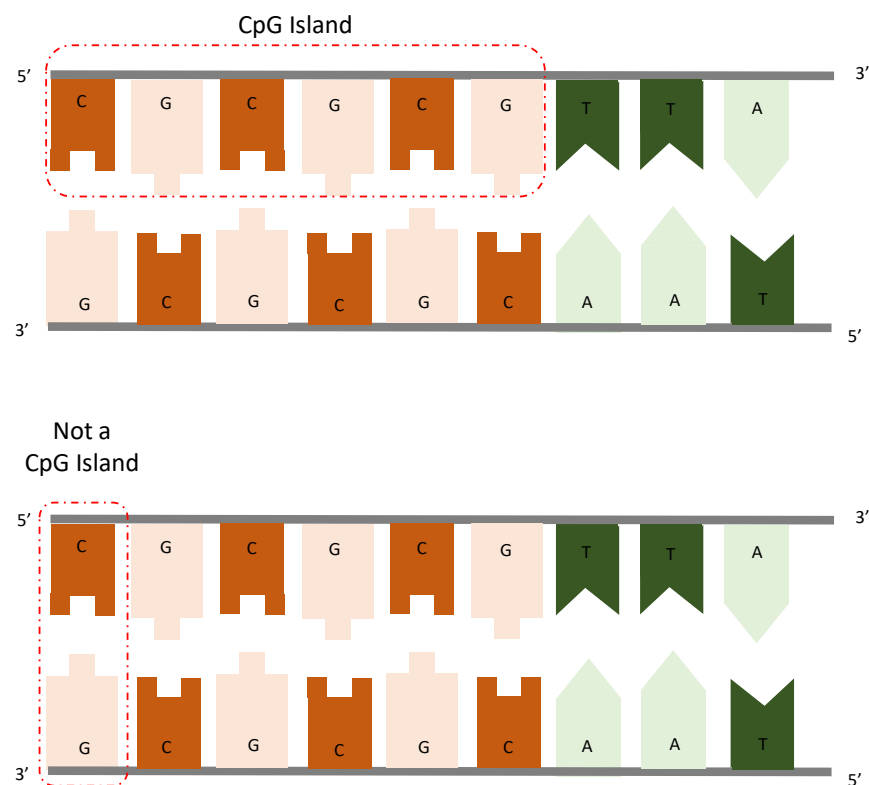


**Figure 1.** Graphical illustration of neurological damage in MS.

There are some gender considerations to take into account, as the illness is more common in women than men in a 3:1 ratio (and in some countries like Sweden even 5:1).

Some of the common symptoms of the illness include fatigue and numbness, typically in one side of the body [11,12]. Behavioral and cognition abnormalities are also common [13–15]. Currently there are many therapeutic approaches to control or stop the progression of the disease, but no curative treatment is available. However, a large amount of research has been generated regarding this disease. MS has a particularly high prevalence in some areas of Europe and the United States, particularly in northern regions [16].

CpG DNA methylation data has been used to analyze neurodegenerative diseases such as Alzheimer's [17–20] and Parkinson [21–23]. As can be seen in Figure 2, in the context of DNA methylation, CpG dinucleotide (or CpG) refers to cytosine followed by a guanine in the same DNA strand (typically 5' to 3'), not to be confused with cytosine and guanine paired in two complementary strands.

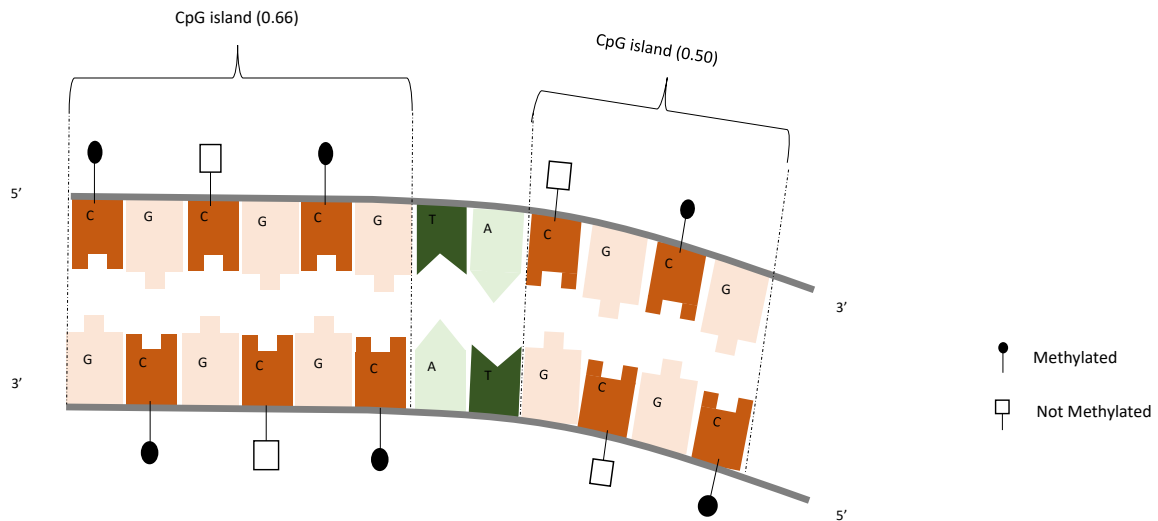


**Figure 2.** Illustration of CpG islands.

Methylation is simply the addition of a methyl group at the 5-carbon (see Figure 3). DNA methylation has been extensively studied in the context of aging, with several biological clocks built using such types of data. Technological advances in recent years have made possible the analysis of DNA methylation levels on thousands of CpGs in a fast and reliable way. In practice, what is obtained is the percentage level of methylation with a value ranging from 0 to 1 (100% methylated). DNA methylation for cancer diagnostics has made significant progress in the last decades, including many seminal papers [24–27]. There is also a significant body of research covering diabetes [28–32].

DNA methylation has also been used in the context of multiple sclerosis [33,34]. Most of the existing literature on the topic tends to use linear approaches. In this paper, we have followed a non-linear approach, which is in principle more generic and encompassing than a linear approach. Machine learning techniques have been successfully used in multiple applications of different types of diseases [35–38]. More specifically, neural networks have been used as an algorithm for the identification of neurodegenerative illnesses, such as Alzheimer's, using DNA methylation data as the input [39–41].

We applied the concept of Shannon Entropy in the context of DNA methylation applied to multiple sclerosis identification. As far as we are aware, this approach has not been followed before. Shannon Entropy is a concept initially developed in information theory, which attempts to quantify the amount of information contained in a certain set of data [42]. The precise mathematical definition of this concept will be introduced in the materials and methods section. It will be shown that using the concept of Shannon Entropy for CpG selection can generate accurate results.



**Figure 3.** DNA methylation illustration.

*Motivation and Aims*

Biomarkers are an increasingly important field, particularly when they can be analyzed using non or minimally invasive techniques. In this regard, blood is a particularly interesting tissue as it can be cheaply and quickly obtained from a patient causing only minimal discomfort. Blood has a significant advantage over other tissues such as brain matter, which is much harder to obtain. DNA methylation data can be accurately and rapidly analyzed using technologies such as the Illumina machines. Shannon Entropy is a concept frequently used in machine learning. The motivation to use this approach for data selection is in trying to find techniques that might reduce the dimensionality of the data. Shannon Entropy is one of the few concepts in the existing literature directly related to the amount of information contained in the data, which seems to be a reasonable starting point when trying to reduce the dimensionality of the data while maintaining as much information as possible.

The aim of this article is to develop techniques to identify DNA methylation signatures applicable for the identification of multiple sclerosis patients.

**2. Materials and Methods**

The DNA methylation data for each individual was stored in a vector  $X^i$ .

$$X^i = \left\{ \begin{matrix} X_1^i \\ X_2^i \\ \vdots \\ X_m^i \end{matrix} \right\} \tag{1}$$

where  $m$  is the number of CpGs analyzed per patient. A numerical example would be:

$$X^2 = \begin{pmatrix} 0.211 \\ 0.723 \\ \vdots \\ 0.983 \end{pmatrix} \tag{2}$$

Which represents all the CpG information available for patient number 2. In this example, the methylation level in the first and second CpGs are 21.1% and 72.3%, respectively. As there is a large number of cases analyzed it is more convenient to group the data in a matrix form.

$$X = \begin{pmatrix} X_1^1 & X_1^2 & \dots & X_1^n \\ X_2^1 & X_2^2 & \dots & X_2^n \\ \vdots & \vdots & & \vdots \\ X_m^1 & X_m^2 & \dots & X_m^n \end{pmatrix} \tag{3}$$

In this notation, there are  $n$  cases (including both patients and controls) with  $m$  CpGs associated with each case. The status of the individual analyzed (multiple sclerosis or control) was defined with a binary variable  $\{0,1\}$  stored in a target vector  $T$ , with the value 0 indicating a healthy control case and the value 1 indicating a patient with multiple sclerosis.

$$T = \{0, 1, 0, \dots, 1\} \tag{4}$$

As there are  $n$  cases, there will be  $n$  entries for this vector. In this example, the first and third cases are control cases, and the second one a patient with MS. As a preliminary step, each CpG was individually linearly modeled against the classification vector  $T$  and only those with a  $p$ -value below 5% were included. The rest of the CpGs were discarded. The dimension of  $X$  was reduced from  $(n \cdot m)$  to  $(n \cdot l)$ , where  $l$  is the number of CpGs with a  $p$ -value below 5%.  $p$ -value prefiltering was carried out in all the data. The Shannon Entropy ( $H$ ) concept was then used to further filter the number of CpGs used. The Shannon Entropy approach step was carried out only for the training dataset. Shannon Entropy can be intuitively understood as the amount of information contained in some data and it is a concept borrowed from information theory. The mathematical expression for Shannon Entropy is as follows:

$$H = - \sum_i P_i \log_2(P_i) \tag{5}$$

This concept is typically applied in discrete mathematics. The probabilities can be estimated empirically. In simple terms, more entropy translates into more information contained. After the initial filtering, the absolute value of the Shannon Entropy was estimated for each CpG.

$$H = \begin{pmatrix} H_1 \\ H_2 \\ \cdot \\ \cdot \\ H_l \end{pmatrix} \tag{6}$$

Only CpGs with an entropy value ( $H_i$ ) bigger than certain predefined value ( $H_i^f$ ) were considered. All the other CpGs were excluded from the analysis. In this way we obtained  $H^*$ .

$$H^* = \begin{pmatrix} H_1^* \\ H_2^* \\ \cdot \\ \cdot \\ H_q^* \end{pmatrix} \quad (7)$$

In this notation  $q \leq l$ . After selecting the CpGs, it is necessary to choose the classification algorithm that is used. A neural network with a hidden layer and an output layer was used. The hidden layer contained 50 artificial neurons, while the output layer contained a single artificial neuron. The 50 neurons in the hidden layer are of the sigmoid symmetric transfer function type. The neuron in the output layer is of the type sigmoid positive transfer function (both of these transfer functions are built-in in Matlab). All the neurons include a bias factor. The neural network was trained with the scaled conjugate backpropagation algorithm. Another four learning algorithms were tested (Levenberg–Marquardt, resilient backpropagation, one-step secant and gradient descent). As in the case of the transfer functions in the artificial neural networks, the learning algorithms are also built-in options in Matlab. Among all the learning algorithms, the best results were obtained using the scaled conjugate backpropagation approach. The data was divided into a training and a testing dataset. The testing dataset accounted for approximately 15% of the data. All the calculations were carried out in Matlab. Neural networks have been extensively used for modeling purposes and can accurately describe many complex underlying dynamics. An important step is to check that the classification error obtained using the above mentioned Shannon Entropy approach for CpG selection is more accurate than the one obtained when using the same number of randomly selected CpGs; in other words, controlling that the improvement in accuracy is not simply due to the reduction in the dimensionality of the data.

All the calculations were done in Matlab, the Shannon Entropy value was calculated using an existing Matlab function. The methylation data was analyzed using two decimals of precision in percentage terms. The analysis did not appear to be very sensitive to an increase to the third decimal place, but it started to have more impact thereafter (four or five decimal places in percentage terms). We believe that using two decimal places is a reasonable precision considering the likely accuracy of the experimental data.

A sensitivity analysis was also carried out. The underlying assumption was that CpGs with very little data variation would be less useful for classification purposes. In an extreme case, if the DNA methylation level for a given CpG was the same for all patients, then this information would not be useful for classification purposes. We did not assume that the CpGs with the most data variation (measured as the standard deviation) were necessarily the best choices, as other factors such as experimental noise (and potentially many others) can increase the variation of the data. However, it seemed reasonable to carry out a sensitivity analysis over reasonable values of the volatility of the DNA methylation data.

### Data

DNA methylation data for 279 individuals were obtained from the GEO database (publicly available data) with the accession code GSE 106648 [43]. The database contained both individuals with multiple sclerosis (140) as well as control individuals (139). The age range was from 16 to 66 years old, and there were 77 male individuals. There were more females than male patients. This is consistent with the observation that MS tends to be more common among females than males; 138 of the individuals in the dataset were

smokers. Age, gender and smoking status (Table 1) were used as inputs in the model. As in the case of DNA methylation, these factors were allocated to their corresponding training or testing dataset.

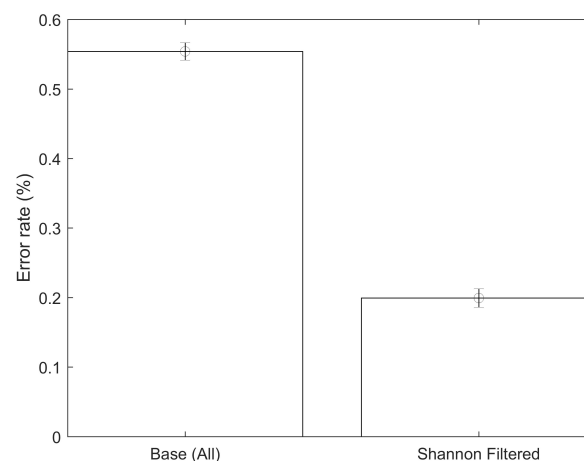
**Table 1.** Basic descriptive information of the patients.

Description	Amount
Male	77
Female	202
Smokers	138
Non-smokers	141
Age	16, 77

The DNA methylation data [43] was obtained from peripheral blood tissue using the Illumina Human Methylation 450 Beach Chip. There were 485,512 CpG DNA methylation data per patient.

### 3. Results

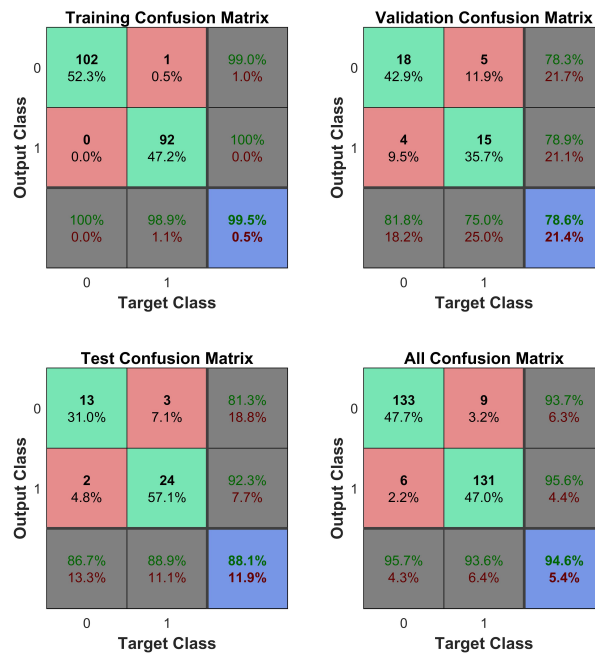
As can be seen in Figure 4, the average classification error using all the available data with a  $p$ -value below 5% was 55.4%, while the error obtained when using only the CpGs with the top 10% Shannon Entropy values (9499 CpGs) was 19.93%, which is a statistically significant improvement. Equivalently, the proposed approach (using Shannon Entropy as a filter) generated a successful classification rate of approximately 80.07%, while the direct approach (using all the data) generated a successful classification rate of approximately 44.6%. The direct approach likely generates poor classifications due to the issue of local minima, which is likely improved by the introduced Shannon Entropy filtering. The model accuracy was substantially improved while at the same time reducing the amount of input data required in the mode. After the two steps ( $p$ -value filtering and Shannon Entropy filtering), the amount of CpGs was reduced by approximately 98% compared to the total initial data available. These results were obtained by dividing the data into training and testing datasets, with the testing dataset not used during the training phase. The testing dataset contained approximately 15% of the total data. Unless explicitly mentioned, all the results shown below refer to the testing dataset results. All the models controlled for age, gender and smoking status of the patients. As it can be seen in Table 2, the average sensitivity and specificity obtained were 78.3% and 81.8%, respectively. An example showing a confusion matrix and ROC can be seen in Figures 5 and 6.



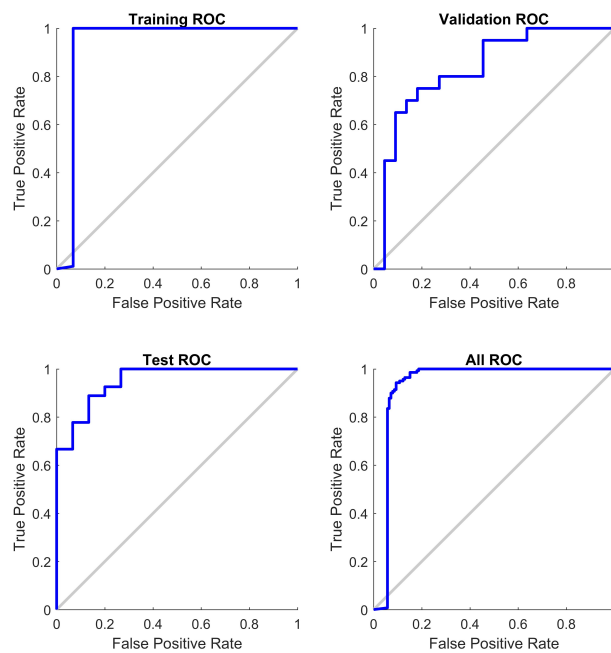
**Figure 4.** Error rate comparison between direct approach and Shannon Entropy filtered approach.

**Table 2.** Average classification forecasting accuracy.

Accuracy Measure	Percentage
Average successful classification	80.1%
Sensitivity	78.3%
Specificity	81.8%



**Figure 5.** A sample confusion matrix (after *p*-value prefiltering and Shannon Entropy filtering).

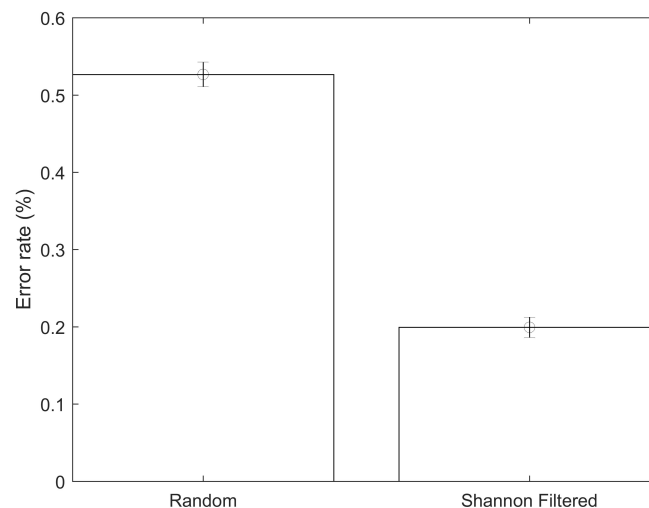


**Figure 6.** ROC (after *p*-value prefiltering and Shannon Entropy filtering).

In order to compare the results, two baseline values were obtained using the volatility (standard deviation) as an indicator. In the first baseline case, the top 2% most volatile CpGs were selected without any prefiltering (such as  $p$ -value). This was done in order to have a dimensionality comparable to the results obtained using the proposed approach ( $p$ -value prefiltering plus Shannon Entropy filtering). The classification success ratio using this technique was approximately 51.6%. A second baseline level was obtained. In this case,  $p$ -value prefiltering was carried out followed by a selection of the most volatile CpGs. The threshold value for the volatility was selected in order to make the final dimension of the data, i.e., number of CpGs selected, approximately the same as the one obtained in the proposed approach ( $p$ -value plus Shannon filtering). The successful classification rate was 56.1%.

An important test to carry out is comparing the performance of the obtained CpGs by the Shannon Entropy approach (as inputs for the classification algorithm) to the results using a matrix of randomly selected CpGs. In this way, we account for the reduction in dimensionality of the data. Ten randomly selected sets of CpGs of the same size as the one obtained using the Shannon Entropy approach (9499) were selected. All the included CpGs in this random approach had  $p$ -values of less than 5%, i.e., this analysis was carried out after the initial linear filtering. Ten simulations were carried out for each of the ten different randomly selected sets of CpGs. The average value and the confidence interval can be seen in Figure 7. The Shannon Entropy approach generates classifications that are statistically significantly more accurate than a random selection of the same size.

As mentioned in the methods and materials section, a sensitivity analysis using the standard deviation of the DNA methylation data for each CpG was also carried out. In Figure 8, the results of selecting the CpGs with the highest volatility are shown. The range selected encompassed the top 5% to the top 50%, in 5% increments. For example, the first column shows the error rate (misclassifications) when using the top 5% of CpGs according to their standard deviation from the initial pool containing 9499 CpGs (after the initial filtering using Shannon Entropy filtering).

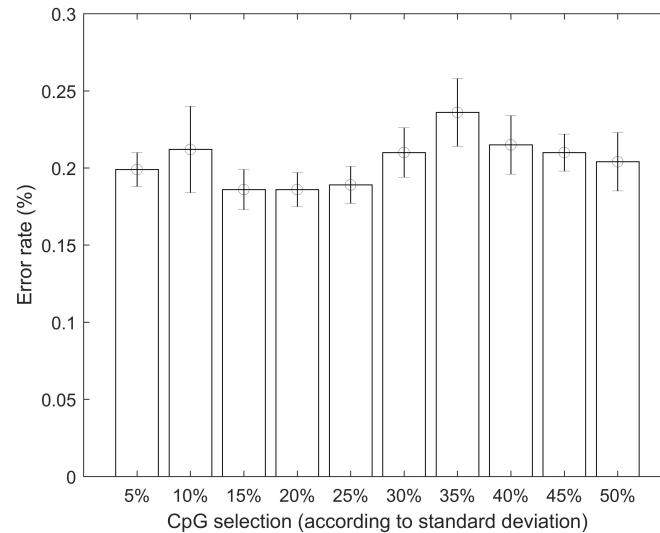


**Figure 7.** Error rate comparison between the Shannon Entropy filtered approach and random selection of the same size.

The intuition behind this approach is selecting CpGs with variation in the methylation values. As an extreme example, completely flat data (with standard deviation equal to zero) will arguably contain no value from a classification point of view. It is also acknowledged that some of that volatility might be caused by experimental and other sources of noise. The best results were obtained when using the top 15% most volatile CpGs with an average



correct classification rate of 81.42%. However, the results were not statistically different (at a 5% significance) when compared with the results obtained by filtering for Shannon Entropy only (no filtering according to the standard deviation of the CpGs).



**Figure 8.** Sensitivity analysis according to the standard deviation of the value of the CpGs. Error rate as a function of the amount of CpGs selected according to their standard deviation.

#### 4. Discussion

An innovative approach is shown for the selection of DNA methylation CpGs to be used in non-linear classification models. This approach is based on the concept of Shannon Entropy, which is an idea borrowed from the information theory field. Shannon Entropy, in simple terms, can be understood as a measure of the amount of information contained in a set of data. The overall data was first filtered, discarding the CpG with  $p$ -values above 5%. A quality pre-check of the data was also carried out, excluding CpGs with missing data. The analyzed dataset appeared to be of good quality with no major data issues. Using the two steps approach of  $p$ -value prefiltering followed by the proposed Shannon Entropy filtering, the dataset was reduced from an original size of approximately 485,512 to a final size of 9499 CpGs, which represents a 98% reduction. The classification analysis, distinguishing between control and multiple sclerosis patients, using the entire dataset, did not generate accurate results. The error rate when using the Shannon Entropy approach was 19.93% (80.07% correct classification), which is a statistically significant improvement over the base case. These error rates were obtained using artificial neural networks as the classification algorithm. All the analyses were carried out controlling for age, gender and smoking status of the patients. It was also tested if the increase in accuracy was due simply to the reduction in the dimensionality of the data. In order to do this, several random CpG configurations of the same size (9499 CpGs) as the one obtained using the Shannon Entropy approach were tested. Their average error rate was 52.66%, which is statistically significantly higher than the results obtained using the Shannon Entropy. This suggests that the Shannon Entropy approach might be a reasonable approach to select potential CpGs relevant for the classification analysis. This type of tool might become rather useful in the future, as the amount of CpGs analyzed per person increases and the computational costs increase accordingly. Another interesting analysis is controlling for the volatility, i.e., the standard deviation, of the CpGs. A sensitivity analysis was carried out in this regard by selecting CpGs according to their standard deviation (in buckets of 5%), i.e., top 5%, top 10%, and so on. When carrying out this type of analysis, there were some improvements in the average accuracy, but these improvements were not statistically significant.

These results were consistent with other articles that found a relationship between DNA methylation in other tissues such as the hippocampus [44]. Using blood as the selected tissue [43] is better suited for clinical purposes. Having a simple test, such as one based on DNA methylation data, which can be applied to many different diseases in a rapid and inexpensive way, can be useful. Multiple sclerosis is a relatively difficult illness to diagnose. Using only clinical symptoms and imaging, such as MRI, is frequently requested when the presence of illness is suspected. From a clinical point of view, it might be practical to have techniques, such as DNA methylation levels in the blood, which can be identified, with a reasonable level of accuracy, the presence of MS with a simple blood test. The physician can use the results from the blood-based biomarker combined with the clinical assessment to decide if it is necessary to carry out further tests, such as imaging.

A very interesting area of future research is the temporal evolution of the DNA methylation in multiple sclerosis, given the diverse evolution of the illness, particularly the long periods of remission experienced by some patients. Further research is necessary to determine feasibility, but it might be possible to use this type of approach for early detection. As more data becomes available, it might be possible to distinguish between different types of illness progression using DNA methylation data. It is possible that differentiating between the different types of evolution might help in targeting therapies in a more precise way.

## 5. Conclusions

Technical improvements are making possible the generation of large amounts of epigenetic data, such as DNA CpG methylation data, that can be used for the detection of several different types of illnesses, such as multiple sclerosis (MS). Multiple sclerosis is a complex illness with genetic and environmental factors, and importantly, an uncertain evolution with some patients experiencing long periods of remission. In this paper, we present a technique based on the Shannon Entropy concept for the selection of CpGs as inputs for MS identification using non-linear techniques such as artificial neural networks. It was shown that using the proposed approach, the number of CpGs used decreased while the accuracy of the classifications significantly improved. As more DNA methylation data becomes available, it is important to have techniques to efficiently filter these large amounts of information. In this regard, borrowing concepts like Shannon Entropy from other disciplines, such as information theory, might be an interesting approach. Having more data is likely beneficial but not all the new data will be helpful for analysis with a large percentage potentially adding noise. Therefore, it is important to develop techniques to further facilitate quantitative data analysis.

In the future, as more DNA CpG methylation data becomes available, it might be possible to extend this type of analysis in order to identify patients with different types of MS evolution. Currently, MS has no cure, but it is a field of intense research. It is possible that differentiating between the different types of evolution might help in targeting therapies in a more precise way, and this is a very appealing area of future research.

**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:** The data is available in the GEO Database with accession code GSE106648.

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

## References

- Sospedra, M.; Martin, R. Immunology of multiple sclerosis. *Annu. Rev. Immunol.* **2005**, *23*, 683–747. [[CrossRef](#)]
- Dendrou, C.; Fugger, L.; Friese, M. Immunopathology of multiple sclerosis. *Nat. Rev. Immunol.* **2015**, *15*, 545–558. [[CrossRef](#)] [[PubMed](#)]
- Lassmann, H. Multiple sclerosis pathology. *Cold Spring Harb. Perspect. Med.* **2018**, *18*. [[CrossRef](#)]
- Frohman, E.; Racke, M.; Raine, C. Multiple sclerosis—The plaque and its pathogenesis. *N. Engl. J. Med.* **2006**, *9*, 942–955. [[CrossRef](#)] [[PubMed](#)]
- Goldenberg, M. Multiple sclerosis review. *Pharm. Ther.* **2012**, *37*, 175–184.
- Dobson, R.; Giovannoni, G. Multiple sclerosis a review. *Eur. J. Neurol.* **2019**, *26*, 27–40. [[CrossRef](#)] [[PubMed](#)]
- Ebers, G. Environmental factors and multiple sclerosis. *Lancet Neurol.* **2008**, *7*, 268–277. [[CrossRef](#)]
- Dyment, D.; Ebers, G.; Sadovnick, D. Genetics of multiple sclerosis. *Lancet Neurol.* **2004**, *3*, 104–110. [[CrossRef](#)]
- Rudick, R.; Cohen, J.; Weinstock-Guttman, B.; Kinkel, R.; Ransohoff, R. Management of multiple sclerosis. *N. Engl. J. Med.* **1997**, *22*, 1604–1611. [[CrossRef](#)]
- Wu, G.; Alvarez, E. The immunopathophysiology of multiple sclerosis. *Neurol. Clin.* **2011**, *29*, 257–278. [[CrossRef](#)]
- Krupp, L. Fatigue in multiple sclerosis. *Arch. Neurol.* **1988**, *45*, 435–437. [[CrossRef](#)]
- Rudick, R.; Schiffer, R.; Schwetz, K.; Herdon, R. Multiple sclerosis: The problem of incorrect diagnosis. *Arch. Neurol.* **1986**, *43*, 578–583. [[CrossRef](#)]
- Feinstein, A. The neuropsychiatry of multiple sclerosis. *Can. J. Psychiatry* **2004**, *49*, 157–163. [[CrossRef](#)]
- Chiaravalloti, N.; DeLuca, J. Cognitive impairment in multiple sclerosis. *Lancet Neurol.* **2008**, *7*, 1139–1151. [[CrossRef](#)]
- Heldner, M.; Kaufmann-Ezra, S.; Gutbrod, K.; Bernasconi, C.; Bigi, S.; Blatter, V.; Kamm, C. Behavioral changes in patients with multiple sclerosis. *Front. Neurol.* **2017**, *8*, 437. [[CrossRef](#)]
- McFarlin, D.; McFarland, H. Multiple sclerosis. *N. Engl. J. Med.* **1982**, *307*, 1246–1251. [[CrossRef](#)]
- Liu, D.; Wang, Y.; Jing, H.; Meng, Q.; Yang, J. Mendelian randomization integrating GWAS and mQTL data identified novel pleiotropic DNA methylation loci for neuropathology of Alzheimer’s disease. *Neurobiol. Aging* **2021**, *97*, 18–27. [[CrossRef](#)]
- Mastroeni, D.; Grover, A.; Whiteside, C.; Coleman, P. Epigenetic changes in Alzheimer’s disease: decrements in DNA methylation. *Neurobiol. Aging* **2010**, *31*, 2025–2037. [[CrossRef](#)]
- Bollati, V.; Galimberti, D.; Pergoli, L.; Dalla Valle, E.; Barretta, F.; Cortini, F. DNA methylation in repetitive elements and Alzheimer disease. *Brain Behav. Immunity* **2011**, *25*, 1078–1083. [[CrossRef](#)]
- Blanch, M.; Mosquera, J.; Ansoleaga, B.; Ferrer, I.; Barrachina, M. Altered mitochondrial DNA methylation pattern in Alzheimer disease-related pathology and in Parkinson disease. *Am. J. Pathol.* **2016**, *186*, 385–397. [[CrossRef](#)]
- Masliyah, E.; Duamop, W.; Galasko, D.; Desplats, P. Distinctive patterns of DNA methylation associated with Parkinson disease: Identification of concordant epigenetic changes in brain and peripheral blood leukocytes. *Epigenetics* **2013**, *8*, 1030–1038. [[CrossRef](#)]
- Miranda-Morales, E.; Meier, K.; Sandoval-Carrillo, A.; Salas-Pacheco, J.; Vazquez-Cardenas, P.; Arias-Carrion, O. Implications of DNA methylation in Parkinson’s disease. *Front. Mol. Neurosci.* **2017**, *10*, 225. [[CrossRef](#)]
- Wulner, U.; Kaut, O.; Piston, D.; Schmitt, I. DNA methylation in Parkinson’s disease. *J. Neurochem.* **2016**, *139*, 108–120. [[CrossRef](#)]
- Chan, K.A.; Jiang, P.; Chan, C.W.; Sun, K.; Wong, J.; Hui, E.P.; Ng, S.S.; Chan, H.L. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 18761–18768. [[CrossRef](#)]
- Lehmann-Werman, R.; Neiman, D.; Zemmour, H.; Moss, J.; Magenheimer, J.; Vaknin-Dembinsky, A.; Rubertsson, S.; Nellgard, B.; Blennow, K.; Zetterberg, H.; et al. Identification of tissue specific cell death using methylation patterns of circulating DNA. *Proc. Natl. Acad. Sci. USA* **2016**, *29*, 1826–1834. [[CrossRef](#)]
- Guo, S.; Diep, D.; Plongthongkum, N.; Fung, H.L.; Zhang, K.; Zhang, K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat. Genet.* **2017**, *49*, 635–642. [[CrossRef](#)]
- Chen, L.H.; Pan, C.; Diplas, B.H.; Xu, C.; Hansen, L.J.; Wu, Y.; Chen, X.; Geng, Y.; Sun, T.; Sun, Y.; et al. The integrated genomic and epigenomic landscape of brainstem glioma. *Nat. Genet.* **2020**, *11*, 3077.
- Bell, C.; Christopher, G. Genome-wide DNA methylation analysis for diabetic nephropathy in type 1 diabetes mellitus. *BMC Med. Genet.* **2010**, *3*, 33.
- Bansal, A.; Pinney, S. DNA methylation and its role in the pathogenesis of diabetes. *Pediatr. Diabetes* **2017**, *18*, 167–177. [[CrossRef](#)]
- Davegardh, C.; Garcia-Calzon, S.; Bacos, K.; Ling, C. DNA methylation in the pathogenesis of type 2 diabetes in humans. *Mol. Metab.* **2018**, *14*, 12–25. [[CrossRef](#)]
- Rakyan, V.; Beyan, H.; Down, T.; Hawa, M.; Maslau, S.; Anden, D.; Leslie, R. Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis. *Epigenomics* **2015**, *7*, 451–460. [[CrossRef](#)] [[PubMed](#)]
- Ronn, T.; Ling, C. DNA methylation as a diagnostic and therapeutic target in the battle against Type 2 diabetes. *PLoS Genet.* **2011**, *7*, 451–460.

33. Bos, S.; Page, C.; Andreassen, B.; Elboudwarej, E.; Gustavsen, M.; Briggs, F.; Barcellos, L. Genome-wide DNA methylation profiles indicate CD8+ T cell hypermethylation in multiple sclerosis. *PLoS ONE* **2015**, *10*, e0117403. [[CrossRef](#)] [[PubMed](#)]
34. Kukulova, O.; Kabilov, M.; Danilova, L.; Popova, E.; Baturina, O.; Tsareva, E. Whole-Genome DNA methylation analysis of peripheral blood mononuclear cells in multiple sclerosis patients with different disease courses. *Acta Nat.* **2016**, *8*, 103–110.
35. Cruz, J.; Wishart, D. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform.* **2006**, *2*, 59–77. [[CrossRef](#)]
36. Fan, Y.; Li, Y.; Bao, X.; Zhu, H.; Lu, L.; Yao, Y.; Li, Y.; Su, M.; Feng, F.; Feng, S.; et al. Development of Machine Learning Models for Predicting Postoperative Delayed Remission in Patients With Cushing's Disease. *J. Clin. Endocrinol. Metab.* **2021**, *106*, 217–231. [[CrossRef](#)]
37. Kourou, K.; Exarchos, T.; Exarchos, K.; Karamouzis, M.; Fotiadis, D. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol.* **2015**, *13*, 8–17. [[CrossRef](#)]
38. Li, Y.; Chen, Z. Performance evaluation of machine learning methods for breast cancer prediction. *Appl. Comput. Math.* **2018**, *7*, 212–216. [[CrossRef](#)]
39. 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](#)]
40. Alfonso Perez, G.; Caballero Villarraso, J. Alzheimer Identification through DNA Methylation and Artificial Intelligence Techniques. *Mathematics* **2021**, *9*, 2482. [[CrossRef](#)]
41. Spolnicka, M.; Pospiech, E.; Peplonska, B. DNA methylation in EVOVL2 and Clorf132 correctly predicted chronological age of individuals from three disease groups. *Int. J. Leg. Med.* **2018**, *132*, 1–11. [[CrossRef](#)]
42. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [[CrossRef](#)]
43. Kular, L.; Liu, Y.; Ruhrmann, S.; Zheleznyakova, G.; Marabita, F.; Gomez-Cabrero, D.; James, T.; Ewing, E.; Linden, M.; Gornikiewicz, B.; et al. DNA methylation as a mediator of HLA-DRB1\*15:01 and a protective variant in multiple sclerosis. *Nat. Commun.* **2018**, *9*, 2397. [[CrossRef](#)]
44. Chomyk, A.M.; Volsko, C.; Tripathi, A.; Deckard, S.A.; Trapp, B.D.; Fox, R.J.; Dutta, R. DNA methylation in demyelinated multiple sclerosis hippocampus. *Bell Syst. Tech. J. Sci. Rep.* **2017**, *18*. [[CrossRef](#)]

## **3.2 Alzheimer Identification through DNA Methylation and Artificial Intelligence Techniques**

Authors: Gerardo Alfonso Perez, Javier Caballero Villarraso

Mathematics. 2021, 9(19), 2482

<https://doi.org/10.3390/math9192482>

Current Impact Factor: 2.258

5-year Impact Factor: 2.1

JCR category rank:

Q1: Mathematics

Article

# Alzheimer Identification through DNA Methylation and Artificial Intelligence Techniques

Gerardo Alfonso Perez <sup>1,\*</sup>  and Javier Caballero Villarraso <sup>1,2</sup>

<sup>1</sup> Department of Biochemistry and Molecular Biology, University of Cordoba, 14071 Cordoba, Spain; bc2cavij@uco.es

<sup>2</sup> 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

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 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, ref. [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, ref. [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.

$$y_i = \begin{cases} 0 & \text{if Control} \\ 1 & \text{if AD} \end{cases} \tag{1}$$

In this way, a vector  $Y = \{Y_1, Y_2, \dots, Y_{nc}\}$  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.

$$X^j = \begin{pmatrix} X^1 \\ X^2 \\ \cdot \\ \cdot \\ \cdot \\ X^{mn} \end{pmatrix} \tag{2}$$

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.

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

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

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



For clarity purposes it is perhaps convenient showing 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:

$$X^1 = \begin{pmatrix} 0.9832 \\ 0.6145 \\ 0.1254 \\ 0.7845 \\ 0.6548 \end{pmatrix} \tag{6}$$

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.

### 3.1. 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  $(p\text{-value}) < 0.005$ .

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

with  $m < mn$ .

### 3.2. 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.

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

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

The values inside the above grids represent the  $X_i$  selected. As an example,  $n_1$  represents  $X_1$ .  $\Delta n_l$  and  $\Delta n_s$  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  $\Delta n_l$ . 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 ( $D$ ).

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

For clarity purposes, let simplify the notation:

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

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  $x_i = \{X_1, \dots, X_m\} = \{X_i\}$ , 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.

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

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  $\forall X_i, i=1, \dots, m$ , 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  $X_i$ s. Otherwise it is included. In this way a filtering is carried out.

$$\{X_1, \dots, X_m\} \rightarrow \{X_1, \dots, X_{m^*}\} \forall S_j \tag{15}$$

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:

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

**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:

$$\max\{HR(S_1), \dots, HR(m)\} \rightarrow S_{max}^1 \tag{18}$$

$$\max\{HR(S_1), \dots, HR(m)\} < S_{max}^1 \rightarrow S_{max-1}^1 \tag{19}$$

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

$$\{S_{max}^1, S_{max-1}^1\} \tag{20}$$

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

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

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

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

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

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

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  $S_{max}^t$ .

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.

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

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 from  $2^{10q}$  to  $102^q$  combinations. It can be proven that:

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

**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:

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

induction hypothesis

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

$$2^{10(k+1)} = 2^{10k}2^{10} > 10 \cdot 2^k2^{10} = 10 \cdot 2^k22^9 = 10 \cdot 2^{k+1}2^9 > 10 \cdot 2^{k+1} \tag{31}$$

which completes the proof by induction.  $\square$

### 3.3. 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 or 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 and 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).

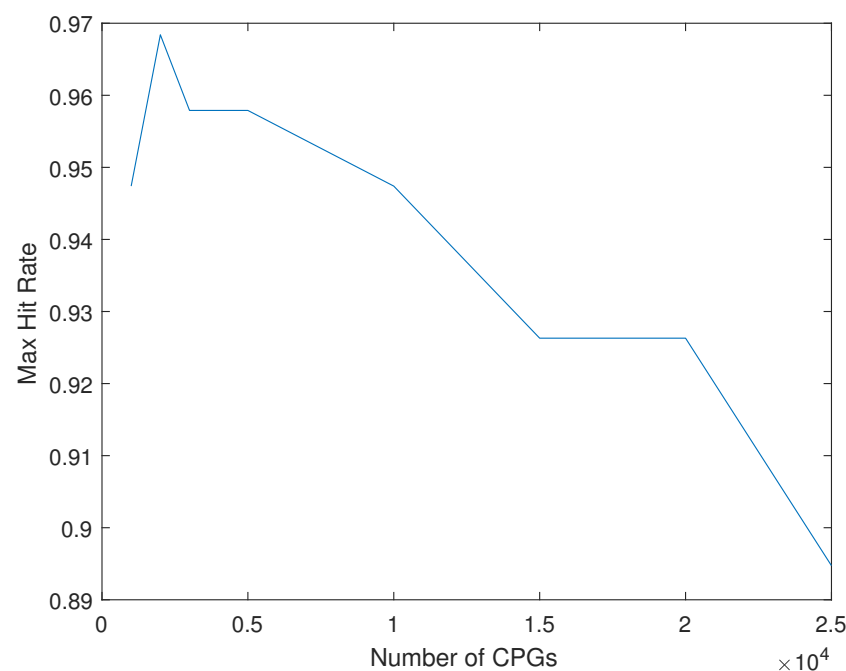
Controls	HR (Linear)	HR (Gaussian)	HR (Polynomial)	CpGs
None	0.8211	0.7921	0.8167	All
Age	0.8947	0.8142	0.8391	All
Gender	0.8211	0.7921	0.8167	All
Cell type	0.8211	0.7921	0.8167	All
Brain Region	0.8211	0.7921	0.8167	All

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).

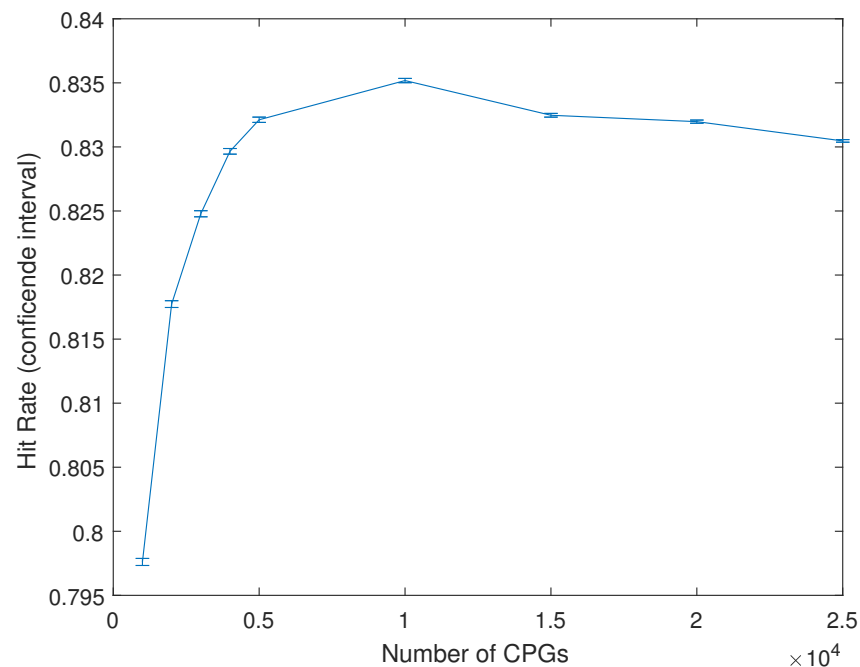
Controls	Hit Rate	CpGs
None	0.7263	41,784
Age	0.8424	41,784
Gender	0.7263	41,784
Cell type	0.7263	41,784
Brain Region	0.7263	41,784

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 increase 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 differences 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.

**Supplementary Materials:** The following are available online at <https://www.mdpi.com/2227-7390/9/19/2482/s1>.

**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 1 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. [[CrossRef](#)]
3. Morris, R.G.; Kopelman, M.D. The memory deficits in Alzheimer-type dementia: A review. *Q. J. Exp. Psychol.* **1986**, *38*, 575–602. [[CrossRef](#)]
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. [[CrossRef](#)]
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. [[CrossRef](#)] [[PubMed](#)]
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. [[CrossRef](#)] [[PubMed](#)]
7. Blennow, K.; Hampel, H.; Weiner, M.; Zetterberg, H. Cerebrospinal fluid and plasma biomarkers in Alzheimer disease. *Nat. Rev. Neurol.* **2010**, *6*, 131–144. [[CrossRef](#)]
8. Hsieh, C.L. Dependence of transcriptional repression on CpG methylation density. *Mol. Cell. Biol.* **1994**, *14*, 5487–5494. [[CrossRef](#)]
9. Cooper, D.N.; Krawczak, M. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Hum. Genet.* **1989**, *83*, 181–188. [[CrossRef](#)]
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. [[CrossRef](#)]
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. [[CrossRef](#)] [[PubMed](#)]
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. [[CrossRef](#)] [[PubMed](#)]
13. Mikeska, T.; Craig, J.M. DNA methylation biomarkers: Cancer and beyond. *Genes* **2014**, *5*, 821–864. [[CrossRef](#)]
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*, 293. [[CrossRef](#)]
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. [[CrossRef](#)] [[PubMed](#)]

16. Kuan, P.F.; Wang, S.; Zhou, X.; Chu, H. A statistical framework for Illumina DNA methylation arrays. *Bioinformatics* **2010**, *26*, 2849–2855. [[CrossRef](#)]
17. You, L.; Han, Q.; Zhu, L.; Zhu, Y.; Bao, C.; Yang, C.; Lei, W.; Qian, W. Decitabine-mediated epigenetic reprogramming enhances anti-leukemia efficacy of CD123-targeted chimeric antigen receptor T-cells. *Front. Immunol.* **2020**, *11*, 1787. [[CrossRef](#)]
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. [[CrossRef](#)]
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. [[CrossRef](#)]
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. [[CrossRef](#)]
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**. [[CrossRef](#)]
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. [[CrossRef](#)]
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. [[CrossRef](#)]
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. [[CrossRef](#)]
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. [[CrossRef](#)]
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. [[CrossRef](#)]
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.; Péligrini-Issac, M.; Colliot, O.; Sarazin, M.; Dubois, B.; Lehericy, S.; Benali, H. Support vector machine-based classification of Alzheimer’s disease from whole-brain anatomical MRI. *Neuroradiology* **2009**, *51*, 73–83. [[CrossRef](#)] [[PubMed](#)]
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. [[CrossRef](#)]
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. [[CrossRef](#)] [[PubMed](#)]
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. [[CrossRef](#)] [[PubMed](#)]
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. [[CrossRef](#)] [[PubMed](#)]
37. Horvath, S.; Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **2018**, *19*, 371–384. [[CrossRef](#)] [[PubMed](#)]
38. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **2013**, *14*, 1–20. [[CrossRef](#)]
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. [[CrossRef](#)] [[PubMed](#)]
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. [[CrossRef](#)] [[PubMed](#)]
41. Grossi, E.; Stocco, 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. [[CrossRef](#)]
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. [[CrossRef](#)]
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. [[CrossRef](#)]
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. [[CrossRef](#)]
45. Lalkhen, A.G.; McCluskey, A. Clinical tests: Sensitivity and specificity. *Contin. Educ. Anaesth. Crit. Care Pain* **2008**, *8*, 221–223. [[CrossRef](#)]

46. Tanzi, R.E. FDA Approval of Aduhelm Paves a New Path for Alzheimer's Disease. *ACS Chem. Neurosci.* **2021**, *12*, 2714–2715. [[CrossRef](#)] [[PubMed](#)]
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. [[CrossRef](#)]
48. Ayton, S. Brain volume loss due to donanemab. *Eur. J. Neurol.* **2021**, *28*, e67–e68. [[CrossRef](#)]
49. Vellas, B.J. The Geriatrician, the Primary Care Physician, Aducanumab and the FDA Decision: From Frustration to New Hope. *J. Nutr. Health Aging* **2021**, *25*, 821–823. [[CrossRef](#)] [[PubMed](#)]

### **3.3 Neural Network Aided Detection of Huntington Disease**

Authors: Gerardo Alfonso Perez, Javier Caballero Villarraso

Journal of Clinical Medicine. 2022, 11(8), 2110

<https://doi.org/10.3390/jcm11082110>

Current Impact Factor: 4.242

5-year Impact Factor: 4.567

JCR category rank:

Q1: Medicine, General & Internal

## Article

# Neural Network Aided Detection of Huntington Disease

Gerardo Alfonso Perez <sup>1,\*</sup>  and Javier Caballero Villarraso <sup>1,2</sup> 

<sup>1</sup> Department of Biochemistry and Molecular Biology, University of Cordoba, 14071 Cordoba, Spain; bc2cavij@uco.es

<sup>2</sup> 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.



**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

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** Huntington disease; DNA methylation; neural networks

## 1. Introduction

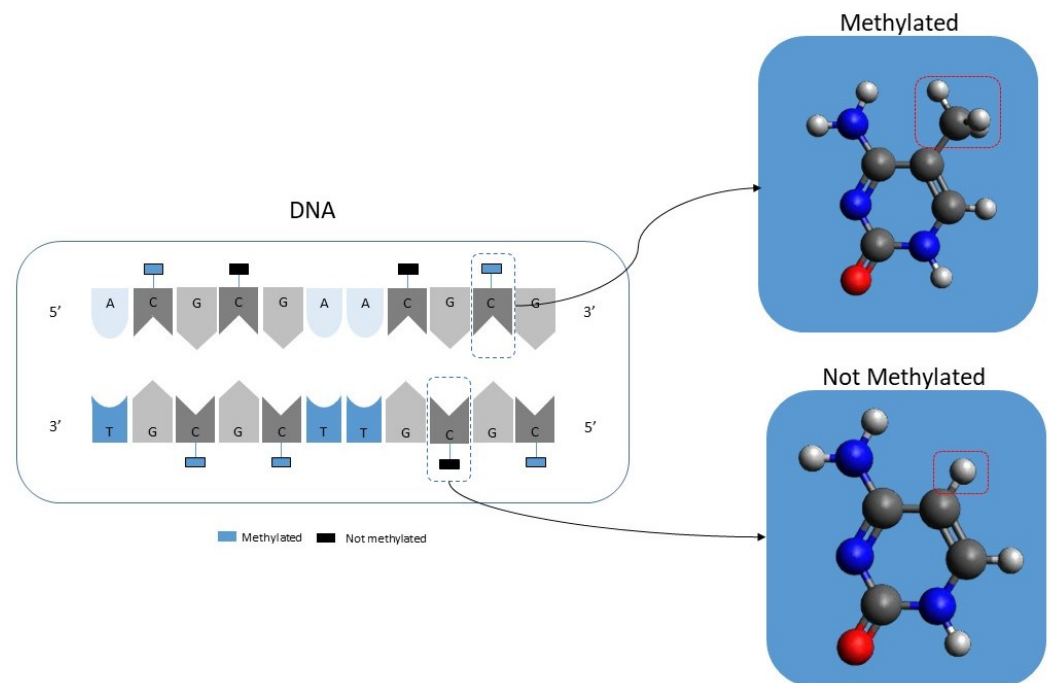
Huntington disease (HD) is a neurological progressive disorder [1–4]. The typical onset of the illness is in mid-adult life [5–7] causing uncontrolled movements as well as declining cognitive and reasoning skills. The disease is associated with a mutation of a gene in Chromosome 4 [8–10] related to the 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.

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

Therefore for  $m$  cases analyzed there is a vector  $Y$

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

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

$$X_p = \begin{pmatrix} X_p^1 \\ X_p^2 \\ \vdots \\ X_p^n \end{pmatrix} \tag{3}$$

Hence, the dataset can be visualized as follows:

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

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  $\varphi$  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.

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

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

$$A = \begin{pmatrix} y_1 & y_2 & \dots & y_k \\ X_1^1 & X_2^1 & \dots & X_k^1 \\ X_1^2 & X_2^2 & \dots & X_k^2 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ X_1^n & X_2^n & \dots & X_k^n \end{pmatrix} \tag{6}$$

$$B = \begin{pmatrix} y_{k+1} & y_{k+2} & \dots & y_m \\ X_{k+1}^1 & X_{k+2}^1 & \dots & X_m^1 \\ X_{k+1}^2 & X_{k+2}^2 & \dots & X_m^2 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ X_{k+1}^n & X_{k+2}^n & \dots & X_m^n \end{pmatrix} \tag{7}$$

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

$$YP = \{YP_{k+1}, \dots, YP_m\} \tag{8}$$

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

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

$$if \begin{cases} YP_l = Y_l \text{ then } a_l = 1 \\ else a_l = 0 \end{cases} \tag{9}$$

- Estimate the accuracy

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

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

$$MF^i = \frac{1}{k} \sum F^i \tag{11}$$

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

$$MF = \{MF^1, \dots, MF^n\} \tag{12}$$

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

$$MF^{new} = \{MF_*^1, \dots, MF_*^{nn}\} \text{ with } nn \leq n. \tag{13}$$



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.

$$\left\{ \begin{matrix} X_p^{*1} \\ X_p^{*2} \\ \cdot \\ \cdot \\ X_p^{*nn} \end{matrix} \right\} \tag{14}$$

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  $B^*$ .

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

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

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

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

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  $\{Y_{k+1}, Y_{k+2}, \dots, Y_m\}$

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

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

- (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  $\{F_1^*, \dots, F_k^*\}$ .

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

8. Reduce the number 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:

$$\left\{ \begin{matrix} X_p^{*1} \\ X_p^{*2} \\ \cdot \\ \cdot \\ X_p^{*nn} \end{matrix} \right\} \tag{21}$$

After this step the new CpG list is:

$$\left\{ \begin{matrix} X_p^{**1} \\ X_p^{**2} \\ \cdot \\ \cdot \\ X_p^{**(nn-1)} \end{matrix} \right\} \tag{22}$$

9. Repeat steps 2 to 5 with the new CpG list (of dimensionality  $nn-1$ ).
10. Estimate the average ( $MF^{**}$ ) of the accuracy  $\{F_1^*, \dots, F_k^*\}$ .
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.

$$A = \begin{pmatrix} 0 & 0 & \dots & 1 \\ 0.651094 & 0.650451 & \dots & 0.634303 \\ 0.960434 & 0.954877 & \dots & 0.957124 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0.077337 & 0.063247 & \dots & 0.090948 \end{pmatrix} \tag{23}$$

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.

$$\begin{pmatrix} cg00000029 & 0.651094 \\ cg00000108 & 0.960434 \\ cg00000109 & 0.899284 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \end{pmatrix} \tag{24}$$

### 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.

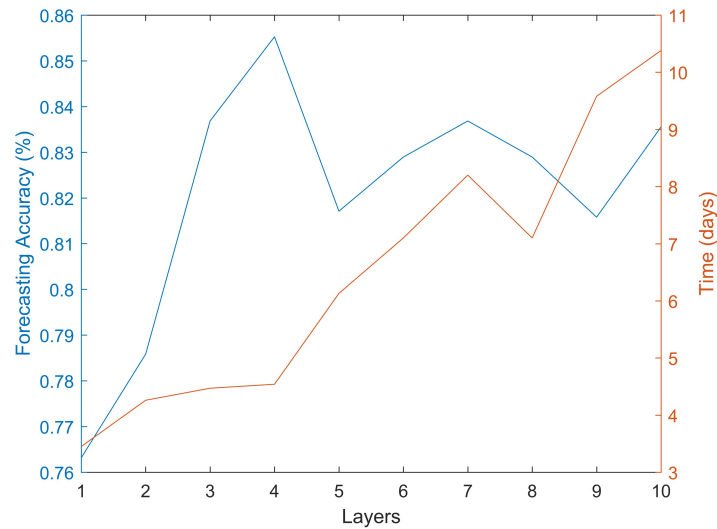
### 3.4. Similarities and Differences with Previously Published Research

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 groscope [50]. Other authors have also carried out machine-learning approaches based on diagnostic imaging information, such as Elahifasae 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 on 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.
3. Sanberg, P.R.; Coyle, J.T. Scientific approaches to Huntington's disease. *CRC Crit. Rev. Clin. Neurobiol.* **1984**, *1*, 1–44.
4. Sturrock, A.; Leavitt, B. The clinical and genetic features of Huntington disease. *J. Geriatr. Psychiatry Neurol.* **2010**, *1*, 243–259.
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.

6. Siemers, E. Huntington disease. *Arch. Neurol.* **2001**, *58*, 308–310.
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.
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.
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.
10. Zuo, J.; Robblins, 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.
11. Cattaneo, E.; Zuccato, C.; Tartari, M. Normal huntingtin function: An alternative approach to Huntington’s disease. *Nat. Rev. Neurosci.* **2005**, *6*, 919–930.
12. Saudou, F.; Humbert, S. The biology of huntingtin. *Neuron* **2016**, *89*, 910–926.
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.
14. Vonsattel, J.P.; DiFiglia, M. Huntington disease. *J. Neuropathol. Exp. Neurol.* **1998**, *57*, 369.
15. Dayalu, P.; Albin, R.L. Huntington disease: Pathogenesis and treatment. *Neurol. Clin.* **2015**, *33*, 101–114.
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.
18. Sobel, S.; Cowan, D. Impact of genetic testing for Huntington disease on the family system. *Am. J. Med Genet.* **2000**, *90*, 49–59.
19. Nance, M. Genetic testing of children at risk for Huntington’s disease. *Neurology* **1997**, *49*, 1048–1053.
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.
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.
23. Robertson, K. DNA methylation and human disease. *Nat. Rev. Genet.* **2005**, *6*, 597–610.
24. Bender, J. DNA methylation and epigenetics. *Annu. Rev. Plant Biol.* **2004**, *55*, 41–68.
25. Holliday, R. DNA methylation and epigenetic inheritance. *Philos. Trans. R. Soc. Lond. Biol. Sci.* **1990**, *326*, 329–338.
26. Lim, D.; Maher, E. DNA methylation: a form of epigenetic control of gene expression. *Obstet. Gynaecol.* **2010**, *12*, 37–42.
27. Richardson, B. Impact of aging on DNA methylation. *Ageing Res. Rev.* **2003**, *2*, 245–261.
28. Jung, M.; Pfeifer, G.P. Aging and DNA methylation. *BMC Biol.* **2015**, *13*, 1–8.
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.
30. Das, P.; Singal, R. DNA methylation and cancer. *J. Clin. Oncol.* **2004**, *22*, 4632–4642.
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.
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.
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.
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.
36. Alfonso Perez, G.; Caballero Villarraso, J. An Entropy Approach to Multiple Sclerosis Identification. *J. Pers. Med.* **2022**, *12*, 398.
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.
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.
39. Zheng, C.; Xu, R. Predicting cancer origins with a DNA methylation-based deep neural network model. *PLoS ONE* **2020**, *5*, e0226461.
40. Paluszczak, J.; Baer-Dubowska, W. Epigenetic diagnostics of cancer—The application of DNA methylation markers. *J. Appl. Genet.* **2006**, *47*, 365–376.
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.
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.
43. Jurmeister, P.; Bockmayr, M.; Seegerer, P.; Bockmayr, T.; Treue, D.; Montavon, G.; Vollbrecht, C.; Arnold, A.; Teichmann, D.; Bressan, 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.



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.
46. Alfonso Perez, G.; Caballero Villarraso, J. Alzheimer Identification through DNA Methylation and Artificial Intelligence Techniques. *Mathematics* **2021**, *9*, 2482.
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.
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.
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.
50. Lones, M.A.; Alty, J.E.; Cosgrove, J.; Duggan-Carter, P.; Jamieson, S.; Naylor, R.F.; 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.
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.
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.
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.
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.
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.
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.
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.
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.



## DISCUSSION

**N**eurodegenerative diseases are likely going to become more frequent as life expectancy increase in many countries as a consequence of these diseases being typically age related. These neurodegenerative illnesses, including Alzheimer Disease, Multiple Sclerosis and Huntington Disease do not have yet a curative therapy. However, there are research teams across the world tackling these illnesses and it is possible that in the not too distant future we start seeing some treatments breakthrough. As with other illnesses it is likely that early detection plays an important role when a treatment is available. In this regard it is important to develop biomarkers that can, with a reasonable level of precision, identify the illness. Ideally these biomarkers should be easily obtained in a way as minimally invasive as possible. For example through a blood test rather than a brain tissue sample. In this regard DNA CpG methylation can

be an interesting option. DNA methylation levels can currently be obtained relatively easily, although it is only available in research settings. In this dissertation it is shown how mathematical models, using as an input these methylation levels can be successfully apply to the identification of patients with Alzheimer Disease, Multiple Sclerosis and Huntington Disease. It is important to mention that the underlying process relating methylation levels and the presence of any of these neurodegenerative illness is not necessarily linear and hence it seems reasonable to use non-linear modeling techniques such as machine learning techniques. More specifically, neural networks were extensively used in this dissertation as a forecasting classification tool.

In the case of Alzheimer Disease a model for the selection of CpGs to be included was presented. This approach managed to reduce the number of CpGs used for several hundred thousands to only 4,300 while increasing the accuracy of the classification. The sensitivity and specificity of this approach differentiating patients with AD from control individuals were 0.9007 and 0.9485 respectively. The proposed selection algorithm used a nonlinear combinatorial approach. It should also be noticed that it is not possible to tests all the potential combinations of more 400,000 CpGs. Reducing the dimensionality of the data i.e., reducing the number of CpGs seems a reasonable step as not all the CpGs will have information that it is relevant for the task of differentiating between AD and control patients. The results shows that the algorithms generates more accurate results than using all the CpG available directly.

Multiple sclerosis was another of the neurodegenerative illness analyzed

---

in this dissertation. A CpG selection approach based on the concept of Shannon Entropy was used. Shannon Entropy is a concept borrowed from the field of information theory and can be understood as the amount of information present in a given set of data. For instance, if the set of data consists only of a vector composed of four identical number i.e., the number 1, the information contained in this vector is a priori potentially less than the information contained in a vector of the same size but composed of integer number ranging from 1 to 9. I present an algorithm, based on this idea, which is able to reduce the number of CpGs substantially while increasing the accuracy of the classification. The number of CpGs was decreased by 98% (only 3% of the initial CpGs were selected). In total number the amount of CpGs were reduced from the original 485,512 to the final 9,499. The correct classification rate obtained using this approach as 80.07%, which is a statistically significant improvement over the base case. It was also tested the hypothesis that the improvement in accuracy was due to the benefit of a (random) decrease in dimensionality. It was shown that this was not the case. In fact random combinations of CpGs of the same total number than the one selected by the algorithm did not generate accurate results, further suggesting that the use of the Shannon Entropy approach is a valid one.

Another algorithm was presented, in this case for the identification of Huntington disease. This approach also reduced the number of CpGs used as inputs. Huntington Disease, similarly to some other neurodegenerative disease does not currently have a cure but it is an area of intense research. It was shown in this dissertation that it is possible to use machine learning techniques, such a neural networks, for the identification of patients with

HD, using as input DNA methylation levels from blood samples. Several different types of neural networks configurations were analyzed. The best results were obtained when using a neural network with four hidden layers. Increasing the complexity of the neural network after a certain threshold did not appear to increase accuracy. It also should be pointed out that the computational demands should be taken into account with some of the models requiring in excess of ten days to be completed.

This dissertation shows that it is possible to use, with reasonable levels of accuracy, machine learning techniques, such as neural networks, and nonlinear combinatorial algorithms, for the task of identifying patients with a neurodegenerative disease, using DNA CpG methylation data as an input. It was also illustrated the importance of the selection of the CpGs. Given the rapidly increasing number of CpGs available, with some machines now able to generate in excess of 800,000 of CpGs methylation levels per patient, it becomes clear that this CpGs selection process needs to be carried out in an automated fashion.

## FUTURE INVESTIGATIONS

**A**s technological advances enable increasingly large number of CpGs to be analyzed in a rapid way the importance of having appropriate algorithms to distinguish between patients with neurodegenerative illnesses and control cases will also increase. Technological advances have increase the number of CpGs by an order of magnitude in recent years with the first machines able to analyze only approximately 20,000 CpGs while the current ones reaching more than 800,000. It is likely that in the not to distant future the number of CpGs will increase again by another order of magnitude. It would be interesting to see, as a line of future research, if the algorithm proposed in this dissertation can generate accurate classification with these increased databases. Another interesting line of future research would be a comparison of DNA methylation profiles among different neurodegenerative diseases. There might be some

## CHAPTER 5. FUTURE INVESTIGATIONS

---

commonalities among these profiles.

It would also seem interesting to study the time evolution of methylation profiles. For instance, it might be possible to try to forecast the expected relapse-remission patterns of patients with Multiple Sclerosis. So far, to the best of my knowledge, it remains uncertain what type of evolution a patient with Multiple Sclerosis is going to have with very substantial differences in the relapse-remission patterns among different patients.

The amount of data in the field of methylation is very likely going to continue increasing. Big data techniques are therefore likely to play an important role as more traditional approaches are likely unfeasible when trying to handle hundred of thousands or even millions of variables.



## CONCLUSIONS

Having techniques that can determine if an individual has Alzheimer Disease, Huntington Disease or Multiple Sclerosis 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 curative therapy. In the case of Alzheimer the results suggest that the approach 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 that increasing the number of CpGs does not improve the forecast. This is very likely related to the issue of local minima.

## CHAPTER 6. CONCLUSIONS

---

I also proposed another algorithm applied to the case of Huntington Disease. In this case, the sensitivity and specificity obtained were 0.95 and 0.80 respectively. The algorithm was able to reduce the initial list of 485,512 CpGs to a total number of 237 CpGs, while at the same time increasing the accuracy of the classification forecasts distinguishing between control and HD cases. As in other neurodegenerative disease there is no curative therapy for HD. However, there are many research teams working on this illness and I am cautiously optimistic that there will be progress. When a curative therapy is available it is likely that early detection might play an important role.

Finally, I also introduced the concept of Shannon Entropy applied to the identification of patients with Multiple Sclerosis. To the best of my knowledge this concept has never been used for the selection of DNA CpG methylation as an input for a classification forecast for Multiple Sclerosis (or any other neurodegenerative disease). The proposed algorithm generated a correct classification rate of approximately 80.07%. This was a statistically significant improvement compared to using all the CpGs available as well as compared to randomly selected groups of CpGs of the same dimensions than the selected by the Shannon Entropy approach.

The nonlinear algorithms presented for all the neurodegenerative diseases analyzed, including Alzheimer Disease, Huntington Disease and Multiple Sclerosis, generated more accurate classification forecasts than the direct approach of using all the available data. They also generated more accurate forecasts than filtering only using linear criteria, suggesting that the underlying mechanism linking CpG methylation data and the identification

---

of a given illness might not be necessarily linear and that a combinatorial nonlinear approach might be a valid analysis technique.



## LIST OF CONTRIBUTIONS

Three articles published in journals (Q1):

- Alfonso Pérez G, Caballero Villarraso J. Alzheimer Identification through DNA Methylation and Artificial Intelligence Techniques. *Mathematics*. 2021; 9(19):2482. doi: 10.3390/math9192482. FACTOR DE IMPACTO (JCR) 2,258 - CUARTIL (JCR) 1º (nº Orden= 24º de 330; (PRIMER DECIL); categoría: 'Mathematics'). ISSN: 2227-7390. Edit. MDPI.
- Alfonso Pérez G, Caballero Villarraso J. An Entropy Approach to Multiple Sclerosis Identification. *J Pers Med*. 2022; 12:398. doi: 10.3390/jpm12030398. FACTOR DE IMPACTO (JCR) 4,945 - CUARTIL (JCR) 1º (nº Orden= 15º de 108; categoría: 'Health Care Sciences & Services'). ISSN: 2075-4426. Edit. MDPI.

## CHAPTER 7. LIST OF CONTRIBUTIONS

---

- Alfonso Perez G, Caballero Villarraso J. Neural Network Aided Detection of Huntington Disease. *J Clin Med.* 2022; 11(8):2110. doi: 10.3390/jcm11082110. FACTOR DE IMPACTO (JCR) 4,242 - CUARTIL (JCR) 1º (nº Orden= 39º de 169; categoría: 'Medicine, General & Internal'). ISSN: 2077-0383. Edit. MDPI.

---

10 abstracts reported in congresses/meetings:

- Alfonso Pérez G, Caballero Villarraso J. Alzheimer disease identification through DNA methylation and neural networks. 2nd targeting therapy for Alzheimer and related neurodegenerative diseases conference. London, (UK), 1-3 Noviembre 2021.
- Alfonso Pérez G, Caballero Villarraso J. Artificial intelligence techniques in Alzheimer detection. 24th IFCC-EFLM European Congress of Clinical Chemistry and Laboratory Medicine (EuroMedLab). Munich, (Germany), 10-14 April 2022.
- Alfonso Pérez G, Caballero Villarraso J. DNA methylation biomarkers for Alzheimer disease (AD). AP/DP 2022 International conference on Alzheimer's and Parkinson's Diseases and related neurological disorders. Barcelona, (Spain), 15-20 Marzo 2022.
- Alfonso Pérez G, Caballero Villarraso J. Parkinson disease (PD) biomarkers using artificial neural networks. AP/DP 2022 International conference on Alzheimer's and Parkinson's Diseases and related neurological disorders. Barcelona, (Spain), 15-20 Marzo 2022.
- Alfonso Pérez G, Caballero Villarraso J. Application of neural network in Alzheimer disease methylation biomarker. AP/DP 2022 International conference on Alzheimer's and Parkinson's Diseases and related neurological disorders. Barcelona, (Spain), 15-20 Marzo 2022.

## CHAPTER 7. LIST OF CONTRIBUTIONS

---

- Alfonso Pérez G, Caballero Villarraso J. Genetics and environment – an integrated approach. London 2022 Global conference of Alzheimer disease international. London, (UK), 6 Junio 2022.
- Alfonso Pérez G, Caballero Villarraso J. Discriminant analysis applied to psychiatric illnesses. 24th IFCC-EFLM European Congress of Clinical Chemistry and Laboratory Medicine (EuroMedLab). Munich, (Germany), 10-14 April 2022.
- Alfonso Pérez G, Caballero Villarraso J. A machine learning approach for the identification of mild cognitive impairment using DNA methylation. 24th IFCC-EFLM European Congress of Clinical Chemistry and Laboratory Medicine (EuroMedLab). Munich, (Germany), 10-14 April 2022.
- Alfonso Pérez G. Estudio De Técnicas De Big Data en Alzheimer. IX Congreso Científico de Investigadores en Formación. NUEVOS DESAFÍOS, NUEVAS OPORTUNIDADES. Córdoba, (Spain), 3-6 Mayo 2021.
- Alfonso Pérez G. Detección de la enfermedad de Creutzfeldt-Jakob (CJD) usando inteligencia artificial. X Congreso Científico de Investigadores en Formación. Córdoba, (Spain), 3-6 Mayo 2021.



## CERTIFICATE OF ATTENDANCE

PRESENTED TO

GERARDO ALFONSO PEREZ

This certificate is to confirm the attendance of GERARDO ALFONSO PEREZ at the 2nd Targeting Therapy of Alzheimer's and Related Neurodegenerative Diseases Virtual Conference, which took place from 1<sup>st</sup> – 3<sup>rd</sup> November 2021.

A poster titled 'ALZHEIMER DISEASE IDENTIFICATION THROUGH DNA METHYLATION AND NEURAL NETWORKS' was also presented at the conference poster sessions.

Thank you for participating.

*AMJOHNSON*

Amy Johnson, Conference Manager

## CHAPTER 7. LIST OF CONTRIBUTIONS

---



UNIVERSIDAD DE CÓRDOBA

**La Vicerrectora de Posgrado e Innovación Docente de la Universidad de Córdoba**


**ACREDITA que**

**Gerardo Alfonso Pérez, con DNI nº 78559838A ha asistido al IX Congreso Científico de Investigadores en Formación, con el título NUEVOS DESAFÍOS, NUEVAS OPORTUNIDADES, organizado por las Escuelas de Doctorado Educo y eidA3 (sede Córdoba) de la Universidad de Córdoba, celebrado en Córdoba los días 3 a 6 de mayo de 2021 y ha presentado la comunicación oral titulada “Estudio De Técnicas De Big Data En Alzheimer“**

**Fdo.: Julieta Mérida García**



Código Seguro de Verificación	VBPSMAKOR4QFU4YZEV7QBQA7SQ	Fecha y Hora	27/05/2021 13:04:17
Normativa	Este documento incorpora firma electrónica reconocida de acuerdo a la ley 59/2003, 19 de diciembre, de firma electrónica		
Firmado por	JULIETA MERIDA GARCIA		
Url de verificación	<a href="https://sede.uco.es/verifirma/">https://sede.uco.es/verifirma/</a>	Página	1/1





24th IFCC-EFLM European Congress of Clinical  
Chemistry and Laboratory Medicine  
**10 - 14 April 2022**

## ABSTRACT SUBMISSION

Dear Dr. **Alfonso Perez Gerardo**,

on behalf of the Scientific Committee, I am delighted to inform you that your abstract titled "**Artificial intelligence techniques in Alzheimer detection**" has been accepted for **poster presentation** at the forthcoming **IFCC-EFLM EuroMedLab Munich 2021 Congress** ([new dates](#): 10-14 April 2022).

A Poster Code has been assigned to your work: **M022**

This code indicates the poster panel and which day you have to set up your poster (i.e.M001 = panel 001 – day Monday, where the M stands for Monday; T001= panel 001 day Tuesday; W001= panel 001 day Wednesday).

### REGISTRATION

**Admission to the poster area and to the scientific sessions is allowed only to participants duly registered;**

Please make sure you have been registered to the Congress in order to present your poster.

**Without registration, your abstract will not be published or displayed during the Congress.**

We remind you that the deadline for the reduced registration is **21 March 2022**.

Registration rates are as follows:

Full Registration	€840,00 (vat included)
Young Registration (≤35 years)	€450,00 (vat included)
Day Registration	€360,00 (vat included)

### VENUE

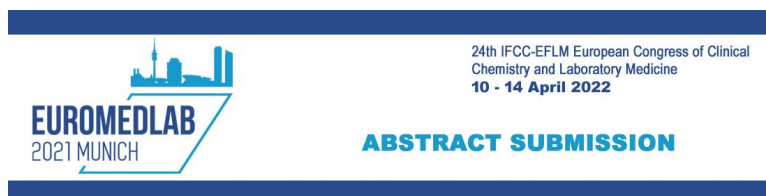
The Congress will be held at the ICM INTERNATIONALES CONGRESS CENTER MÜNCHEN  
(Messegelände 81823 Munich, Germany)

The Poster Area will be located inside the Exhibition Area at first floor and it will be properly sign posted.

### POSTER PRESENTATION

## CHAPTER 7. LIST OF CONTRIBUTIONS

---



Dear Dr. **Gerardo Alfonso Perez**,

on behalf of the Scientific Committee, I am delighted to inform you that your abstract titled "**A machine learning approach for the identification of mild cognitive impairment using DNA methylation**" has been accepted for **poster presentation** at the forthcoming **IFCC-EFLM EuroMedLab Munich 2021 Congress** ([new dates](#): 10-14 April 2022).

A Poster Code has been assigned to your work: **M080**

This code indicates the poster panel and which day you have to set up your poster (i.e. M001 = panel 001 – day Monday, where the M stands for Monday; T001= panel 001 day Tuesday; W001= panel 001 day Wednesday).

### **REGISTRATION**

**Admission to the poster area and to the scientific sessions is allowed only to participants duly registered;**

Please make sure you have been registered to the Congress in order to present your poster.

**Without registration, your abstract will not be published or displayed during the Congress.**

We remind you that the deadline for the reduced registration is **21 March 2022**.

Registration rates are as follows:

Full Registration	€840,00 (vat included)
Young Registration (≤35 years)	€450,00 (vat included)
Day Registration	€360,00 (vat included)

### **VENUE**

The Congress will be held at the ICM INTERNATIONALES CONGRESS CENTER MÜNCHEN (Messegelände 81823 Munich, Germany)

The Poster Area will be located inside the Exhibition Area at first floor and it will be properly sign posted.



24th IFCC-EFLM European Congress of Clinical  
Chemistry and Laboratory Medicine  
**10 - 14 April 2022**

## ABSTRACT SUBMISSION

Dear Dr. **Gerardo Alfonso Perez**,

on behalf of the Scientific Committee, I am delighted to inform you that your abstract titled "**Discriminant analysis applied to psychiatric illnesses**" has been accepted for **poster presentation** at the forthcoming **IFCC-EFLM EuroMedLab Munich 2021 Congress** ([new dates](#); 10-14 April 2022).

A Poster Code has been assigned to your work: **W160**

This code indicates the poster panel and which day you have to set up your poster (i.e.M001 = panel 001 – day Monday, where the M stands for Monday; T001= panel 001 day Tuesday; W001= panel 001 day Wednesday).

### REGISTRATION

**Admission to the poster area and to the scientific sessions is allowed only to participants duly registered;**

Please make sure you have been registered to the Congress in order to present your poster.

**Without registration, your abstract will not be published or displayed during the Congress.**

We remind you that the deadline for the reduced registration is **21 March 2022**.

Registration rates are as follows:

Full Registration	€840,00 vat included)
Young Registration (≤35 years)	€450,00 (vat included)
Day Registration	€360,00 (vat included)

### VENUE

The Congress will be held at the ICM INTERNATIONALES CONGRESS CENTER MÜNCHEN (Messegelände 81823 Munich, Germany)

The Poster Area will be located inside the Exhibition Area at first floor and it will be properly sign posted.

## CHAPTER 7. LIST OF CONTRIBUTIONS

---



**Abstract Number: 553**

**Abstract Title: PARKINSON'S DISEASE (PD) BIOMARKERS USING ARTIFICIAL NEURAL NETWORKS**

Dear Dr. Gerardo Alfonso Perez,

Thank you for your abstract submission for the upcoming **hybrid AD/PD22: 16<sup>th</sup> International Conference on Alzheimer's & Parkinson's Diseases** taking place **Online** and in **Barcelona, Spain** during **March 15-20, 2022**

On behalf of the Scientific Committee, we are pleased to inform you that your abstract has been accepted for both a **Paper (onsite) + ePoster presentation (virtual platform)**.

The ePoster will consist of a 1-page PDF file and a 5-minutes MP3 audio file. You will be sent a link to upload your ePoster in December 2021.

Instructions for preparing both the Paper and ePoster will be posted [here](#).

### **Registration for the Conference**

- The Poster is only secured in the program by registering for the Conference. Only abstracts of presenters who have registered and paid their fees by **Monday, 15<sup>th</sup> November 2021** will be included in the final program. You may register and pay associated fees online [here](#). *Please note that you are asked to register earlier than regular delegates as we need to finalise anyone who has a poster in the program.*
- Please let us know if you will not be the presenting author, or if you will be registered as part of a group/by a company.
- In order to prepare the correct number of boards for onsite presentation, you will be asked to confirm, by no later than January 15<sup>th</sup>, 2022, that you are able to travel to Barcelona and present the **Paper** Poster onsite. A questionnaire for confirmation will be sent to you in early January 2022.

For any questions regarding your abstract, please contact us at [adpd\\_abstracts@kenes.com](mailto:adpd_abstracts@kenes.com).

Yours sincerely,

**AD/PD 2022 Conference Secretariat**

---

Dear Dr Gerardo Alfonso Perez

As the presenter for abstract id 5, titled Genetics and environment – an integrated approach, at 35th Global Conference of Alzheimer's Disease International, Thursday, 9th June, 2022 to Saturday, 11th June, 2022, please ensure you register to attend the conference whether virtually or in-person. If you have not registered already, please be aware that the early bird registration deadline ends on **31 March**, so register soon to make the most of the reduced rates. There are further reduced rates for those from low- and middle-income countries, as defined by the World Bank, family carers, students and people living with dementia.

If your abstract has been selected for a virtual session or a pre-recorded session you must still register for the conference. If someone else is presenting on your behalf please ensure that the presenter is registered and let us know who the presenter will be.

Registration rates can be found on the ADI [conference website](#). Please follow the link to register. You will be able to login using the same email address that you created your abstract account with [ga284@cantab.net](mailto:ga284@cantab.net). Please ensure your account name is the name that you would like on your delegate badge.

We look forward to seeing you in London, UK and online from Thursday, 9th June, 2022.

With best wishes,

Alzheimer's Disease International

[events@alzint.org](mailto:events@alzint.org)

[www.adiconference.org](http://www.adiconference.org)

## CHAPTER 7. LIST OF CONTRIBUTIONS

---



**Abstract Number: 595**

**Abstract Title: DNA METHYLATION BIOMARKERS FOR ALZHEIMER DISEASE (AD)**

Dear Dr. Gerardo Alfonso Perez,

Thank you for your abstract submission for the upcoming **hybrid AD/PD22: 16<sup>th</sup> International Conference on Alzheimer's & Parkinson's Diseases** taking place **Online** and in **Barcelona, Spain** during **March 15-20, 2022**

On behalf of the Scientific Committee, we are pleased to inform you that your abstract has been accepted for both a **Paper (onsite) + ePoster presentation (virtual platform)**.

The ePoster will consist of a 1-page PDF file and a 5-minutes MP3 audio file. You will be sent a link to upload your ePoster in December 2021.

Instructions for preparing both the Paper and ePoster will be posted [here](#).

### **Registration for the Conference**

- The Poster is only secured in the program by registering for the Conference. Only abstracts of presenters who have registered and paid their fees by **Monday, 15<sup>th</sup> November 2021** will be included in the final program. You may register and pay associated fees online [here](#). *Please note that you are asked to register earlier than regular delegates as we need to finalise anyone who has a poster in the program.*
- Please let us know if you will not be the presenting author, or if you will be registered as part of a group/by a company.
- In order to prepare the correct number of boards for onsite presentation, you will be asked to confirm, by no later than January 15<sup>th</sup>, 2022, that you are able to travel to Barcelona and present the **Paper** Poster onsite. A questionnaire for confirmation will be sent to you in early January 2022.

For any questions regarding your abstract, please contact us at [adpd\\_abstracts@kenes.com](mailto:adpd_abstracts@kenes.com).

Yours sincerely,

**AD/PD 2022 Conference Secretariat**





UNIVERSIDAD DE CÓRDOBA

El Vicerrector de Cultura, Comunicación y Proyección Social. Por  
suplencia de la Vicerrectora de Posgrado e Innovación Docente de la  
Universidad de Córdoba

**ACREDITA que**

**Gerardo Alfonso Perez**, con DNI nº **78559838A** ha asistido al **X Congreso Científico de Investigadores en Formación**, con el título *“El arte de investigar”*, organizado por las Escuelas de Doctorado Educo y eidA3 (sede Córdoba) de la Universidad de Córdoba, celebrado en Córdoba los días 3 al 6 de mayo de 2022, y **ha presentado la comunicación en formato videoposter** titulada *“Detección De La Enfermedad De Creutzfeldt-Jakob (Cjd) Usando Inteligencia Artificial”*

**Fdo: Luis Manuel Medina Canalejo**



<b>Código Seguro De Verificación:</b>	c9FNaoiOvgKimgTW/gSfKQ==	<b>Fecha</b>	17/05/2022	
<b>Normativa</b>	Este documento incorpora firma electrónica reconocida de acuerdo a la Ley 59/2003, de 19 de diciembre, de firma electrónica.			
<b>Firmado Por</b>	Luis Manuel Medina Canalejo			
<b>Url De Verificación</b>	<a href="https://sede.uco.es/verifirma/code/c9FNaoiOvgKimgTW/gSfKQ==">https://sede.uco.es/verifirma/code/c9FNaoiOvgKimgTW/gSfKQ==</a>	<b>Página</b>	1/1	

## CHAPTER 7. LIST OF CONTRIBUTIONS

---



**Abstract Number: 605**

**Abstract Title: APPLICATIONS OF NEURAL NETWORKS IN ALZHEIMER DISEASE METHYLATION BIOMARKERS**

Dear Dr. Gerardo Alfonso Perez,

Thank you for your abstract submission for the upcoming **hybrid AD/PD22: 16<sup>th</sup> International Conference on Alzheimer's & Parkinson's Diseases** taking place **Online** and in **Barcelona, Spain** during **March 15-20, 2022**

On behalf of the Scientific Committee, we are pleased to inform you that your abstract has been accepted for both a **Paper (onsite) + ePoster presentation (virtual platform)**.

The ePoster will consist of a 1-page PDF file and a 5-minutes MP3 audio file. You will be sent a link to upload your ePoster in December 2021.

Instructions for preparing both the Paper and ePoster will be posted [here](#).

### **Registration for the Conference**

- The Poster is only secured in the program by registering for the Conference. Only abstracts of presenters who have registered and paid their fees by **Monday, 15<sup>th</sup> November 2021** will be included in the final program. You may register and pay associated fees online [here](#). *Please note that you are asked to register earlier than regular delegates as we need to finalise anyone who has a poster in the program.*
- Please let us know if you will not be the presenting author, or if you will be registered as part of a group/by a company.
- In order to prepare the correct number of boards for onsite presentation, you will be asked to confirm, by no later than January 15<sup>th</sup>, 2022, that you are able to travel to Barcelona and present the **Paper** Poster onsite. A questionnaire for confirmation will be sent to you in early January 2022.

For any questions regarding your abstract, please contact us at [adpd\\_abstracts@kenes.com](mailto:adpd_abstracts@kenes.com).

Yours sincerely,

**AD/PD 2022 Conference Secretariat**

## BIBLIOGRAPHY

- [1] K. ABHISHEK, M. SINGH, S. GHOSH, AND A. ANAND, *Weather forecasting model using artificial neural network*, *Procedia Technology*, 4 (2012), pp. 311–318.
- [2] O. I. ABIODUN, A. JANTAN, A. E. OMOLARA, K. V. DADA, N. A. MOHAMED, AND H. ARSHAD, *State-of-the-art in artificial neural network applications: A survey*, *Heliyon*, 4 (2018), p. e00938.
- [3] O. R. ADAM AND J. JANKOVIC, *Symptomatic treatment of huntington disease*, *Neurotherapeutics*, 5 (2008), pp. 181–197.
- [4] A. AIMAR, H. MOSTAFA, E. CALABRESE, A. RIOS-NAVARRO, R. TAPIADOR-MORALES, I.-A. LUNGU, M. B. MILDE, F. CORRADI, A. LINARES-BARRANCO, S.-C. LIU, ET AL., *Nullhop: A flexible convolutional neural network accelerator based on sparse representations of feature maps*, *IEEE transactions on neural networks and learning systems*, 30 (2018), pp. 644–656.
- [5] J. A. ANDERSON, *An introduction to neural networks*, MIT press, 1995.

## BIBLIOGRAPHY

---

- [6] J.-B. BARDIN, G. SPREEMANN, AND K. HESS, *Topological exploration of artificial neuronal network dynamics*, *Network Neuroscience*, 3 (2019), pp. 725–743.
- [7] K. BASTERRETXEA, J. M. TARELA, AND I. DEL CAMPO, *Approximation of sigmoid function and the derivative for hardware implementation of artificial neurons*, *IEE Proceedings-Circuits, Devices and Systems*, 151 (2004), pp. 18–24.
- [8] G. P. BATES, R. DORSEY, J. F. GUSELLA, M. R. HAYDEN, C. KAY, B. R. LEAVITT, M. NANCE, C. A. ROSS, R. I. SCAHILL, R. WETZEL, ET AL., *Huntington disease*, *Nature reviews Disease primers*, 1 (2015), pp. 1–21.
- [9] J. BEHLER, *Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations*, *Physical Chemistry Chemical Physics*, 13 (2011), pp. 17930–17955.
- [10] A. B. BEN-ZACHARIA, *Therapeutics for multiple sclerosis symptoms*, *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, 78 (2011), pp. 176–191.
- [11] Q. BI, K. E. GOODMAN, J. KAMINSKY, AND J. LESSLER, *What is machine learning? a primer for the epidemiologist*, *American journal of epidemiology*, 188 (2019), pp. 2222–2239.
- [12] I. BILBAO AND J. BILBAO, *Overfitting problem and the over-training in the era of data: Particularly for artificial neural networks*, in

- 2017 eighth international conference on intelligent computing and information systems (ICICIS), IEEE, 2017, pp. 173–177.
- [13] T. D. BIRD, *Genetic aspects of alzheimer disease*, Genetics in Medicine, 10 (2008), pp. 231–239.
- [14] C. M. BISHOP, *Neural networks and their applications*, Review of scientific instruments, 65 (1994), pp. 1803–1832.
- [15] E. J. BOERS AND H. KUIPER, *Biological metaphors and the design of modular artificial neural networks*, (1992).
- [16] H. BRAAK AND E. BRAAK, *Frequency of stages of alzheimer-related lesions in different age categories*, Neurobiology of aging, 18 (1997), pp. 351–357.
- [17] C. BRACKENRIDGE, *Factors influencing dementia and epilepsy in huntington’s disease of early onset*, Acta Neurologica Scandinavica, 62 (1980), pp. 305–311.
- [18] ———, *Parental factors associated with rigidity in huntington’s disease.*, Journal of Medical Genetics, 17 (1980), pp. 112–114.
- [19] G. CARLEO, I. CIRAC, K. CRANMER, L. DAUDET, M. SCHULD, N. TISHBY, L. VOGT-MARANTO, AND L. ZDEBOROVÁ, *Machine learning and the physical sciences*, Reviews of Modern Physics, 91 (2019), p. 045002.

## BIBLIOGRAPHY

---

- [20] S. CHEN AND S. A. BILLINGS, *Neural networks for nonlinear dynamic system modelling and identification*, International journal of control, 56 (1992), pp. 319–346.
- [21] Y. CHRISTEN, *Oxidative stress and alzheimer disease*, The American journal of clinical nutrition, 71 (2000), pp. 621S–629S.
- [22] G. CIPRIANI, C. DOLCIOTTI, L. PICCHI, AND U. BONUCCELLI, *Alzheimer and his disease: a brief history*, Neurological Sciences, 32 (2011), pp. 275–279.
- [23] P. M. CONNEALLY, *Huntington disease: genetics and epidemiology*, American journal of human genetics, 36 (1984), p. 506.
- [24] D. CRAUFURD, J. C. THOMPSON, AND J. S. SNOWDEN, *Behavioral changes in huntington disease*, Cognitive and Behavioral Neurology, 14 (2001), pp. 219–226.
- [25] H. J. CRAYTON AND H. S. ROSSMAN, *Managing the symptoms of multiple sclerosis: a multimodal approach*, Clinical therapeutics, 28 (2006), pp. 445–460.
- [26] P. CUNNINGHAM, M. CORD, AND S. J. DELANY, *Supervised learning*, in Machine learning techniques for multimedia, Springer, 2008, pp. 21–49.
- [27] P. DAYALU AND R. L. ALBIN, *Huntington disease: pathogenesis and treatment*, Neurologic clinics, 33 (2015), pp. 101–114.

- [28] J. E. DAYHOFF AND J. M. DELEO, *Artificial neural networks: opening the black box*, *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 91 (2001), pp. 1615–1635.
- [29] R. DOBSON AND G. GIOVANNONI, *Multiple sclerosis—a review*, *European journal of neurology*, 26 (2019), pp. 27–40.
- [30] C. DUMITRU AND V. MARIA, *Advantages and disadvantages of using neural networks for predictions.*, *Ovidius University Annals, Series Economic Sciences*, 13 (2013).
- [31] D. A. DYMENT, G. C. EBERS, AND A. D. SADOVNICK, *Genetics of multiple sclerosis*, *The Lancet Neurology*, 3 (2004), pp. 104–110.
- [32] G. C. EBERS, *Environmental factors and multiple sclerosis*, *The Lancet Neurology*, 7 (2008), pp. 268–277.
- [33] I. EL NAQA AND M. J. MURPHY, *What is machine learning?*, in *machine learning in radiation oncology*, Springer, 2015, pp. 3–11.
- [34] O. ELUYODE AND D. T. AKOMOLAFE, *Comparative study of biological and artificial neural networks*, *European Journal of Applied Engineering and Scientific Research*, 2 (2013), pp. 36–46.
- [35] G. FILIPPINI, L. MUNARI, B. INCORVAIA, G. C. EBERS, C. POLMAN, R. D’AMICO, AND G. P. RICE, *Interferons in relapsing remitting multiple sclerosis: a systematic review*, *The Lancet*, 361 (2003), pp. 545–552.

## BIBLIOGRAPHY

---

- [36] S. A. FRAUTSCHY, A. BAIRD, AND G. M. COLE, *Effects of injected alzheimer beta-amyloid cores in rat brain*, Proceedings of the National Academy of Sciences, 88 (1991), pp. 8362–8366.
- [37] A. GIUBELLINO, T. R. BURKE, AND D. P. BOTTARO, *Grb2 signaling in cell motility and cancer*, Expert opinion on therapeutic targets, 12 (2008), pp. 1021–1033.
- [38] K. GOHIL, *Multiple sclerosis: progress, but no cure*, Pharmacy and Therapeutics, 40 (2015), p. 604.
- [39] C. A. GOLD AND A. E. BUDSON, *Memory loss in alzheimer’s disease: implications for development of therapeutics*, Expert review of neurotherapeutics, 8 (2008), pp. 1879–1891.
- [40] M. M. GOLDENBERG, *Multiple sclerosis review*, Pharmacy and therapeutics, 37 (2012), p. 175.
- [41] L. GOLDMAN, D. A. AUSIELLO, AND A. I. SCHAFER, *Goldman-Cecil. Tratado de medicina interna*, Elsevier Health Sciences, 2021.
- [42] C. HAMZAÇEBI, *Improving artificial neural networks’ performance in seasonal time series forecasting*, Information Sciences, 178 (2008), pp. 4550–4559.
- [43] N. J. HOLLAND AND M. MADONNA, *Nursing grand rounds: multiple sclerosis*, Journal of Neuroscience nursing, 37 (2005), p. 15.



- [44] D. M. HOLTZMAN, E. MANDELKOW, AND D. J. SELKOE, *Alzheimer disease in 2020*, Cold Spring Harbor perspectives in medicine, 2 (2012), p. a011585.
- [45] A. T. HOOGEVEEN, R. WILLEMSSEN, N. MEYER, K. E. D. ROOLJ, R. A. ROOS, G.-J. B. V. OMMEN, AND H. GALJAARD, *Characterization and localization of the huntington disease gene product*, Human molecular genetics, 2 (1993), pp. 2069–2073.
- [46] Y. HUANG AND L. MUCKE, *Alzheimer mechanisms and therapeutic strategies*, Cell, 148 (2012), pp. 1204–1222.
- [47] H. JAHN, *Memory loss in alzheimer’s disease*, Dialogues in clinical neuroscience, (2022).
- [48] M. I. JORDAN AND T. M. MITCHELL, *Machine learning: Trends, perspectives, and prospects*, Science, 349 (2015), pp. 255–260.
- [49] J. M. KARASINSKA AND M. R. HAYDEN, *Cholesterol metabolism in huntington disease*, Nature Reviews Neurology, 7 (2011), pp. 561–572.
- [50] S. C. KIRKWOOD, J. L. SU, P. M. CONNEALLY, AND T. FOROUD, *Progression of symptoms in the early and middle stages of huntington disease*, Archives of neurology, 58 (2001), pp. 273–278.
- [51] I. KISTER, T. E. BACON, E. CHAMOT, A. R. SALTER, G. R. CUTTER, J. T. KALINA, AND J. HERBERT, *Natural history of multiple*

## BIBLIOGRAPHY

---

- sclerosis symptoms*, International journal of MS care, 15 (2013), pp. 146–156.
- [52] D. S. KNOPMAN, H. AMIEVA, R. C. PETERSEN, G. CHÉTELAT, D. M. HOLTZMAN, B. T. HYMAN, R. A. NIXON, AND D. T. JONES, *Alzheimer disease*, Nature reviews Disease primers, 7 (2021), pp. 1–21.
- [53] E. L. KOEDAM, V. LAUFFER, A. E. VAN DER VLIES, W. M. VAN DER FLIER, P. SCHELTENS, AND Y. A. PIJNENBURG, *Early-versus late-onset alzheimer’s disease: more than age alone*, Journal of Alzheimer’s Disease, 19 (2010), pp. 1401–1408.
- [54] H. KUKREJA, N. BHARATH, C. SIDDESH, AND S. KULDEEP, *An introduction to artificial neural network*, Int J Adv Res Innov Ideas Educ, 1 (2016), pp. 27–30.
- [55] A. KUMAR, J. SIDHU, A. GOYAL, AND J. W. TSAO, *Alzheimer disease*, (2018).
- [56] K. G. LIAKOS, P. BUSATO, D. MOSHOU, S. PEARSON, AND D. BOCHTIS, *Machine learning in agriculture: A review*, Sensors, 18 (2018), p. 2674.
- [57] B. LIU, *Supervised learning*, in Web data mining, Springer, 2011, pp. 63–132.
- [58] M. LOCK, *The alzheimer conundrum*, in The Alzheimer Conundrum, Princeton University Press, 2013.

- [59] I. LOMA AND R. HEYMAN, *Multiple sclerosis: pathogenesis and treatment*, *Current neuropharmacology*, 9 (2011), pp. 409–416.
- [60] H. J. MACLEAN AND M. S. FREEDMAN, *Multiple sclerosis: following clues from cause to cure*, *The Lancet Neurology*, 8 (2009), pp. 6–8.
- [61] R. L. MARGOLIS AND C. A. ROSS, *Diagnosis of huntington disease*, *Clinical chemistry*, 49 (2003), pp. 1726–1732.
- [62] M. F. MENDEZ, *Early-onset alzheimer disease*, *Neurologic clinics*, 35 (2017), pp. 263–281.
- [63] H.-J. MÖLLER AND M. B. GRAEBER, *The case described by alois alzheimer in 1911*, *European archives of psychiatry and clinical neuroscience*, 248 (1998), pp. 111–122.
- [64] C. MOUNT AND C. DOWNTON, *Alzheimer disease: progress or profit?*, *Nature medicine*, 12 (2006), pp. 780–784.
- [65] A. PALMER, J. J. MONTANO, AND A. SESÉ, *Designing an artificial neural network for forecasting tourism time series*, *Tourism management*, 27 (2006), pp. 781–790.
- [66] G. PANCHAL, A. GANATRA, P. SHAH, AND D. PANCHAL, *Determination of over-learning and over-fitting problem in back propagation neural network*, *International Journal on Soft Computing*, 2 (2011), pp. 40–51.

## BIBLIOGRAPHY

---

- [67] J. S. PAULSEN, *Cognitive impairment in huntington disease: diagnosis and treatment*, *Current neurology and neuroscience reports*, 11 (2011), pp. 474–483.
- [68] K. A. QUAID AND M. MORRIS, *Reluctance to undergo predictive testing: the case of huntington disease*, *American journal of medical genetics*, 45 (1993), pp. 41–45.
- [69] C. A. RAJI, O. LOPEZ, L. KULLER, O. CARMICHAEL, AND J. BECKER, *Age, alzheimer disease, and brain structure*, *Neurology*, 73 (2009), pp. 1899–1905.
- [70] J. RÍO, C. NOS, M. TINTORÉ, N. TÉLLEZ, I. GALÁN, R. PELAYO, M. COMABELLA, AND X. MONTALBAN, *Defining the response to interferon- $\beta$  in relapsing-remitting multiple sclerosis patients*, *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 59 (2006), pp. 344–352.
- [71] C. A. ROSS, E. H. AYLWARD, E. J. WILD, D. R. LANGBEHN, J. D. LONG, J. H. WARNER, R. I. SCAHILL, B. R. LEAVITT, J. C. STOUT, J. S. PAULSEN, ET AL., *Huntington disease: natural history, biomarkers and prospects for therapeutics*, *Nature Reviews Neurology*, 10 (2014), pp. 204–216.
- [72] C. A. ROSS AND S. J. TABRIZI, *Huntington’s disease: from molecular pathogenesis to clinical treatment*, *The Lancet Neurology*, 10 (2011), pp. 83–98.

- [73] F. J. SEPULVEDA, J. PARODI, R. W. PEOPLES, C. OPАЗO, AND L. G. AGUAYO, *Synaptotoxicity of alzheimer beta amyloid can be explained by its membrane perforating property*, PloS one, 5 (2010), p. e11820.
- [74] A. SERRANO-POZO, M. P. FROSC, E. MASLIAH, AND B. T. HYMAN, *Neuropathological alterations in alzheimer disease*, Cold Spring Harbor perspectives in medicine, 1 (2011), p. a006189.
- [75] P. SHARMA AND M. KAUR, *Classification in pattern recognition: A review*, International Journal of Advanced Research in Computer Science and Software Engineering, 3 (2013).
- [76] R. SIMUTIS, D. DILIJONAS, L. BASTINA, AND J. FRIMAN, *A flexible neural network for atm cash demand forecasting*, in Proceedings of the sixth WSEAS international conference on computational intelligence, man-machine systems and cybernetics (CIMMACS 07), vol. 162165, 2007.
- [77] A. K. SINGHAL, V. NAITHANI, O. P. BANGAR, ET AL., *Medicinal plants with a potential to treat alzheimer and associated symptoms*, International Journal of Nutrition, Pharmacology, Neurological Diseases, 2 (2012), p. 84.
- [78] I. SKOOG, R. N. KALARIA, AND M. BRETELER, *Vascular factors and alzheimer disease.*, Alzheimer disease and associated disorders, 13 (1999), pp. S106–14.

## BIBLIOGRAPHY

---

- [79] J. S. SNOWDEN, D. CRAUFURD, H. L. GRIFFITHS, AND D. NEARY, *Awareness of involuntary movements in huntington disease*, *Archives of neurology*, 55 (1998), pp. 801–805.
- [80] M. SOSPEDRA AND R. MARTIN, *Immunology of multiple sclerosis*, *Annu. Rev. Immunol.*, 23 (2005), pp. 683–747.
- [81] L. STEINMAN, *A molecular trio in relapse and remission in multiple sclerosis*, *Nature Reviews Immunology*, 9 (2009), pp. 440–447.
- [82] ———, *Immunology of relapse and remission in multiple sclerosis*, *Annual review of immunology*, 32 (2014), pp. 257–281.
- [83] A. STURROCK AND B. R. LEAVITT, *The clinical and genetic features of huntington disease*, *Journal of geriatric psychiatry and neurology*, 23 (2010), pp. 243–259.
- [84] K. L. SUGARS AND D. C. RUBINSZTEIN, *Transcriptional abnormalities in huntington disease*, *TRENDS in Genetics*, 19 (2003), pp. 233–238.
- [85] R. E. TANZI, *The genetics of alzheimer disease*, *Cold Spring Harbor perspectives in medicine*, 2 (2012), p. a006296.
- [86] G. TEZEL AND M. BUYUKYILDIZ, *Monthly evaporation forecasting using artificial neural networks and support vector machines*, *Theoretical and applied climatology*, 124 (2016), pp. 69–80.
- [87] J. V. TU, *Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes*, *Journal of clinical epidemiology*, 49 (1996), pp. 1225–1231.

- [88] F. J. TWEEDIE, S. SINGH, AND D. I. HOLMES, *Neural network applications in stylometry: The federalist papers*, *Computers and the Humanities*, 30 (1996), pp. 1–10.
- [89] F.-Y. TZENG AND K.-L. MA, *Opening the black box-data driven visualization of neural networks*, IEEE, 2005.
- [90] C. VAN CAUWENBERGHE, C. VAN BROECKHOVEN, AND K. SLEEGERS, *The genetic landscape of alzheimer disease: clinical implications and perspectives*, *Genetics in Medicine*, 18 (2016), pp. 421–430.
- [91] T. VUJICIC, T. MATIJEVIC, J. LJUCOVIC, A. BALOTA, AND Z. SEVARAC, *Comparative analysis of methods for determining number of hidden neurons in artificial neural network*, in *Central european conference on information and intelligent systems*, Faculty of Organization and Informatics Varazdin, 2016, p. 219.
- [92] S.-C. WANG, *Artificial neural network*, in *Interdisciplinary computing in java programming*, Springer, 2003, pp. 81–100.
- [93] H. L. WEINER, *The challenge of multiple sclerosis: how do we cure a chronic heterogeneous disease?*, *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 65 (2009), pp. 239–248.
- [94] B. T. WINSLOW, M. ONYSKO, C. M. STOB, AND K. A. HAZLEWOOD, *Treatment of alzheimer disease*, *American family physician*, 83 (2011), pp. 1403–1412.

## BIBLIOGRAPHY

---

- [95] D. WU, K. KIM, G. EL FAKHRI, AND Q. LI, *Iterative low-dose ct reconstruction with priors trained by artificial neural network*, IEEE transactions on medical imaging, 36 (2017), pp. 2479–2486.
- [96] Z. ZHANG, M. W. BECK, D. A. WINKLER, B. HUANG, W. SIBANDA, H. GOYAL, ET AL., *Opening the black box of neural networks: methods for interpreting neural network models in clinical applications*, 2018.
- [97] N. ZILKA AND M. NOVAK, *The tangled story of alois alzheimer*, Bratislavske lekarske listy, 107 (2006), p. 343.