SVU

الجامعة الافتراضية السورية

SYRIAN VIRTUAL UNIVERSITY

Syrian Arab Republic

Ministry of Higher Education

Syrian Virtual University

BIS ماجستير التأهيل و التخصص في المعلوماتية الحيوية

Deep learning in clinical epigenetics: shedding new light on pathological processes of Alzheimer's disease in the perspective of therapeutic approaches

A thesis submitted in partial fulfillment of the requirements for the degree of Master in Bioinformatics

Chahed Nahal

chahed_187103

Supervisor:

A.Prof. Raouf Hamdan

2023-2024

Dedication

This work is dedicated to the memory of my grandparents for their kindness, principles, and constant love of science, which made them an example of all that is genius and special in life.

I also dedicate this work to my family:

– my father and mother Nabil and Yola, for their goodness, kindness, and love,

– to my brother Ghadi, for his love, support and calming presence,

– my aunt Raghida, and uncles Fouad, Nawaf, Fawaz, Fadi and Samer, the family I cannot exist outside of.

I also dedicate this to the great friends that I met along the way – Shaza, Kenana, Raghad, and Rita. Thank you for always being here for me. May we always strive for success and science together.

Acknowledgements

Deepest regard and thanks are in order to Professor M. di Reggi for his insight in this work, kindness and support, and all my gratitude to engineers A. Zein, O. Mohammad, and K. Hejazi, for their aid, assistance, and support in the obstacles that came with this project.

Special thanks to Professor R. Hamdan for his supervision.

Abstract

This thesis discusses Alzheimer’s disease, its impact and pathogenesis, and delves into DNA methylation of Alzheimer’s disease: the most prominent epigenetic mechanism in the disease. Several analytic methods are present for the detection of DNA methylation on cytosine-guanine dinucleotides, but none can fully grasp all loci, they also have limitations due to computational ability along with their high cost. Artificial intelligence, therefore, can be of better benefit in this case by using the results of analytic methods such as epigenome-wide association studies, and whole-genome bisulfite sequencing, and extracting the data from them to help train and test models for the prediction of new previously undetected loci in the genome, for instance. First, however, it must be noted that artificial intelligence in epigenetics is very recently new and only a handful of studies have been employed for the identification of loci undergoing epigenetic tags in Alzheimer’s disease. Also, no deep learning models have been reported so far that are targeted towards the identification of methylated CpGs in an AD context. Our reference study EWASplus, for example, utilizes data from large resources on a super-computer-scale and uses it to predict new methylated loci of CpGs related to the disease.

Our goal was to take inspiration from this reference study in the aim of trying a new model – deep learning – in the hopes of coming closer to finding new therapeutic approaches. In this thesis, we used whole-genome bisulfite sequencing data and EWAS data to train two models: `RandomForestRegressor`, a machine learning model, and `KerasRegressor`, a deep learning model. Both were applied in this thesis to predict previously undetected methylated CpGs on chromosome 19, which contains the most important risk gene in Alzheimer’s disease: APOE.

The models resulted in the prediction of four CpGs on chromosome 19 that present with higher correlation than the rest in terms of methylation and Alzheimer's disease.

However, many technical and computational limitations were present in the application of the models, leading to low performance. This attempt at applying a deep learning model in this epigenetic context still remains promising, due to its higher efficacy in comparison with machine learning in general.

Therefore, it is immensely important that studies such as the one presented in this thesis have broader horizons in terms of resources to fully reach the potential of the models and datasets, leading to higher precision, and closer steps towards Alzheimer's disease therapy.

Summary

Background

Alzheimer's disease is the number one leading dementia-causing disease worldwide, with an estimated 60-70% percentage of dementia patients having Alzheimer's disease. It is classified as an elderly disease, since aging is a common risk factor, although rare cases do occur that classify as early-onset Alzheimer's disease. It is characterized by gradual memory loss, with short-term memory being affected at first, and it reaches a period of decline in cognitive performance and normal daily functions which leads the patient to constant need of care from others. The Alzheimer's brain is affected by the constant accumulation of amyloid beta, the hyperphosphorylation of tau protein, and neurofibrillary tangles. The causes are many and still obscure. Genetics play a role, with the gene APOE (located on chromosome 19) being the most prominent, and so do epigenetics, such as DNA methylation. As of now, this disease still has no cure – only some treatments are available that slow down the symptoms but no therapy has been found to halt disease progression, leading more and more research to help solve this complex disease.

Aim of study

The aim of this study is to apply a machine learning model and a deep learning model in the context of the epigenetics of Alzheimer's disease, and then attempt to identify novel unexplored areas of DNA (CpG loci) affected by DNA methylation that show positive correlation with the presence of Alzheimer's disease. The identification of these CpG loci will be of benefit for future clinical trials to target their methylation for finding better treatment of AD.

Methods

This study used artificial intelligence (machine learning and deep learning) for the identification of new CpG loci by training a model to distinguish AD-related CpG loci from epigenome-wide association studies (EWAS), selecting the top features that are of importance from the data, and then predicting other (novel) loci not present in this study (EWAS) on a chromosome-wide scale (chromosome 19) from genome data.

Results

This work applied machine learning and deep learning for selecting the most important features and for predicting new CpG loci of chromosome 19. Four loci were predicted on sites 474445, 18768688, 1580576, 939810, but low performance was found with models due to lack of regional resources and limitations along with the need for supercomputers for proper performance. These resulting loci can provide insight into the DNA methylation mechanisms present in Alzheimer's disease, and with the help of super computers, the models performed here can reach higher and more accurate results leading to accurate therapeutic approaches towards Alzheimer's disease.

Conclusion

In conclusion, the main aim was to utilize deep learning in clinical epigenetics and discuss therapeutic approaches. Both a machine learning model and a deep learning model were used in this thesis to identify methylated CpG loci on chromosome 19 in Alzheimer’s disease. Due to computational limitations, short time, and low internet speed, only a portion of the data use in this thesis’s reference study were used, which resulted in low scores and inaccurate predictions. Even though the data may be inaccurate, this still offers potential for these models to be used with larger data and better resources in the aim of extracting beneficial information regarding DNA methylation of CpG sites of the genome, paving the way for higher accuracy in terms of DNA-methylation-targeted therapy

Therefore, with higher computational power, better funding, and time, serious steps can be taken towards epigenetics-targeted drugs to help reverse the reversable sides of epigenetics in early stages of the disease before it is too late.

Keywords: *Alzheimer’s Disease, Epigenomics, CpG, Deep Learning, APOE, DNA methylation, Chromosome 19*

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Name	Abbreviation
Alzheimer’s disease	AD
Late onset Alzheimer’s disease	LOAD
Magnetic resonance imaging	MRI
Positron emission tomography	PET
Amyloid beta	Aβ
Amyloid precursor protein	APP
Silencing RNA	siRNA
Micro-RNA	miRNA
Small nuclear RNA	snRNA
Cytosine-phospate-guanine	CpG
Epigenome-Wide Association Studies	EWAS
Paired helical filaments	PHF
DNA methyltransferases	DNMT
Whole-genome bisulfite sequencing	WGBS
Amyloid precursor protein	APP
Apolipoprotein E	APOE
Presenilin 1	PSEN1
Presenilin 2	PSEN2
Neurofibrillary tangle density	NFTs
The Consortium to Establish a Registry for Alzheimer's Disease	CERAD

Single nucleotide polymorphism	SNP
Random Forest Regressor	RFR
Keras Regressor	KR

Chapter 1: Theoretical Review of Study Problem

Introduction

In 1906, the famous psychiatrist and neuroanatomist, Alois Alzheimer, reported his findings in the 37th Meeting of South-West German Psychiatrists in Tubingen by describing a female patient that was admitted to a psychiatric hospital for paranoia, confusion, and memory disturbance, and he kept following her condition until her death 5 years later. The breakthrough of his report (*Uber einen eigenartigen, schweren Erkrankungsprozeß der Hirnrinde*) [1] was the findings he described in the brain, characterized by neurofibrillary tangles and distinctive plaques in the histology of the brain.

Thanks to Alois Alzheimer, and more extensive research in the past 60 years, we now know that Alzheimer's disease is a neurodegenerative disease (a disease caused by the degeneration and corruption of the nerves) that causes memory loss and atrophy of the hippocampus and medial temporal lobe. It is mostly associated with elderly patients since aging is the most common risk factor (classified as late onset Alzheimer's disease, 'LOAD') [2], even though early-onset AD can also be present in rare cases (5%) in which patients are younger than 65 years of age [3].

According to the World Health Organization, it is estimated that 60-70% of people living with dementia (55 million worldwide) have Alzheimer's disease [4], which is the most common form of dementia; and, according to Li, Feng, et al., from 1990 to 2019, the incidence and prevalence of Alzheimer's disease and other dementias increased by 147.95 and 160.84% [5].

As of today, the main causes, risk factors, and the actual pathological mechanisms behind the disease are still immensely obscure, due to many limitations, including technology, the subtlety of the first stages that may be obscure or simply ignored, along with the fact that the primary methods that allow precise understanding and analysis of the disease are through *post mortem* diagnosis of neurofibrillary tangles and abnormal plaque deposits on the brain, which hinder any accurate studies during the patient’s life. Common risk factors that are believed to play a role in the occurrence of AD are many and are yet to be further analyzed, but as of now, one can summarize them in the following figure:

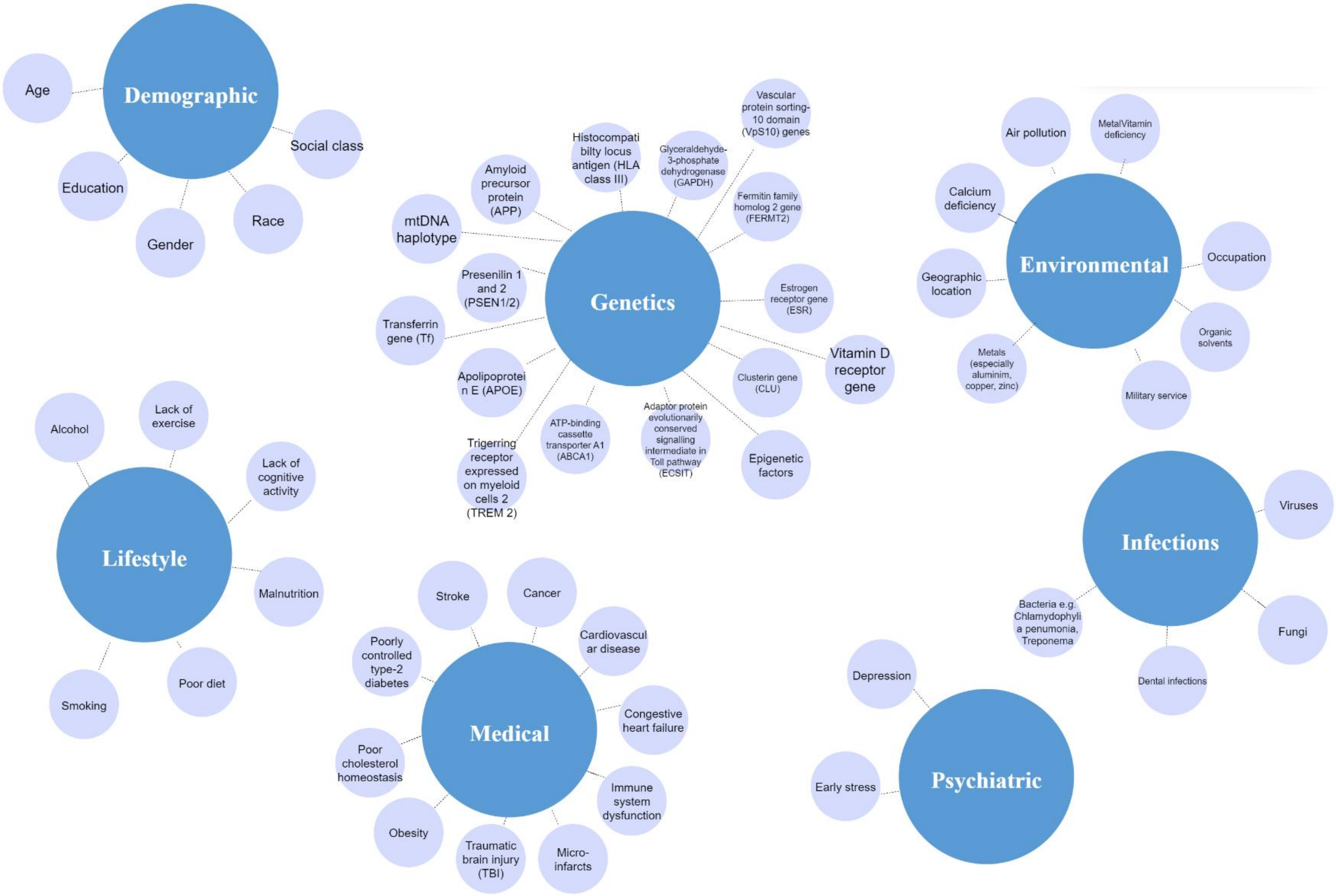

Demographic
Age
Social class
Education
Race
Gender
Genetics
Histocompati bilty locus antigen (HLA class III)
Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)
Vascular protein sorting-10 domain (VpS10) genes
Amyloid precursor protein (APP)
Fermitin family homolog 2 gene (FERMT2)
mtDNA haplotype
Presenilin 1 and 2 (PSEN1/2)
Estrogen receptor gene (ESR)
Transferrin gene (Tf)
Vitamin D receptor gene
Apolipoprotei n E (APOE)
Clusterin gene (CLU)
ATP-binding cassette transporter A1 (ABCA1)
Adaptor protein evolutionarily conserved signalling intermediate in Toll pathway (ECSIT)
Trigerring receptor expressed on myeloid cells 2 (TREM 2)
Epigenetic factors
Environmental
Air pollution
MetalVitamin deficiency
Calcium deficiency
Occupation
Geographic location
Organic solvents
Metals (especially aluminim, copper, zinc)
Military service
Lifestyle
Alcohol
Lack of exercise
Lack of cognitive activity
Malnutrition
Smoking
Poor diet
Infections
Viruses
Bacteria e.g. Chlamydophylia penumonia, Treponema
Fungi
Dental infections
Medical
Stroke
Cancer
Poorly controlled type-2 diabetes
Cardiovascular disease
Congestive heart failure
Poor cholesterol homeostasis
Immune system dysfunction
Obesity
Traumatic brain injury (TBI)
Micro-infarcts
Psychiatric
Depression
Early stress

Figure 1 The most common risk factors playing major roles in Alzheimer’s disease, including many environmental factors and epigenetics [6].

It is possible in most cases to diagnose a patient through brain scanning such as magnetic resonance imaging (MRI) and positron emission tomography (PET) and genetic testing [7], and recent evidence suggests that there may be biomarkers that can help identify the presence of the disease.

The most common characteristics that are believed to play a key role in the disease are amyloid beta and tau proteins, each of which has a different role in neurodegeneration. Amyloid beta (Aβ) is a product of the proteolysis (the breakdown of proteins) of a transmembrane protein, amyloid precursor protein (APP) [8]. When Aβ is not disposed of, it forms extracellular senile plaque accumulation which is believed to cause neuronal toxicity through the induction of mitochondrial dysfunction and oxidative stress in AD neurons. Therefore, through the gradual accumulation of aging cell residue with no way of disposal. As for tau protein, it is a microtubule-associated protein, which forms insoluble filaments that cause accumulation of neurofibrillary

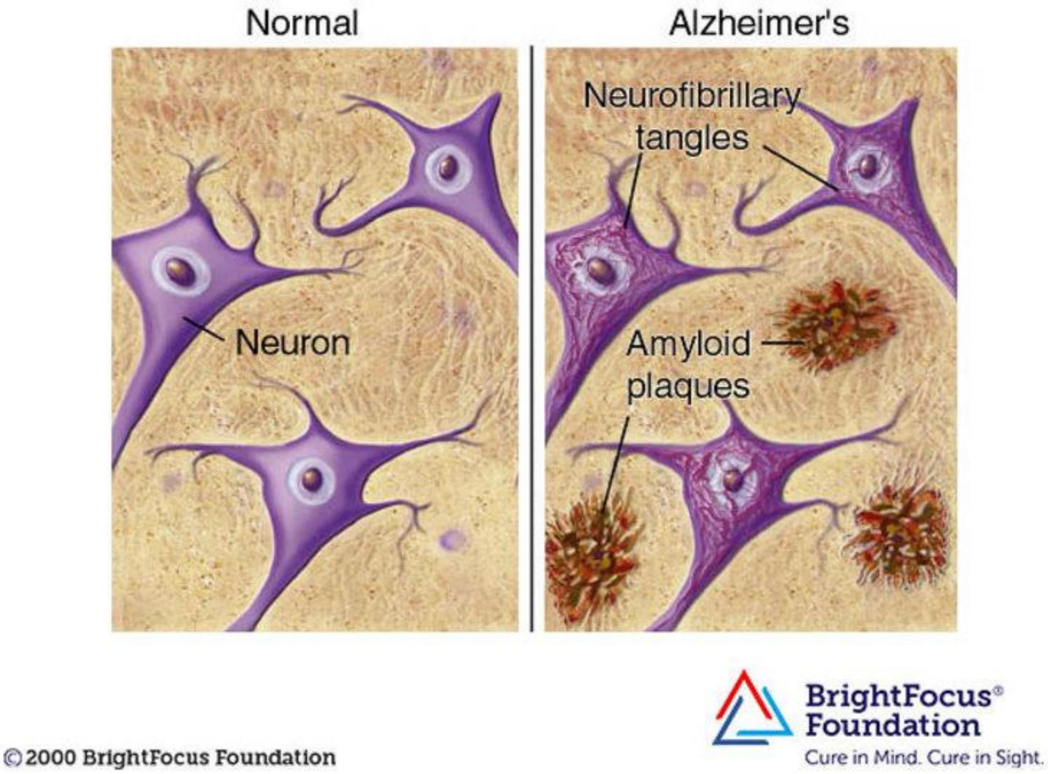


Figure 2 Comparison between normal neurons, and Alzheimer's neurons surrounded by amyloid plaques and cluttered inside by neurofibrillary tangles [10].

tangles in AD. It normally regulates the structure stability of microtubules, so when issues arise in the accumulation, tau forms neurofibrillary tangles, and the microtubules, and by cause, the neurons themselves, begin to collapse [9].

The reason why Aβ or tau malfunction or accumulate, is yet to be known, but recently, large evidence has had an inclination towards epigenetics and its ability to change the condition of a patient away from or closer to AD. Epigenetics is the effect the environment has on the genetic code without directly affecting the DNA or changing it, but in the meantime can also be heritable [11]. It can be either DNA methylation, histone modification, and noncoding RNA (siRNA, miRNA, snRNA). The most commonly studied epigenetic tag is DNA methylation [12], which will be the main epigenetic tag of this study. Through DNA methylation, a methyl group is transferred from adenosylmethionine to DNA [13], thus, an elevation occurs in plasma homocysteine which increases the risk for developing dementia and AD [14].

Even though it is believed that epigenetics might be the key to understanding the pathophysiology of AD, it has much to be studied in order to fully grasp all the mechanisms involved. Recently, new technologies and bioinformatics approaches have emerged to cover this subject, such as Epigenome-Wide Association Studies (EWASs) which provide detailed insight into the methylation sites of the brain of AD patients. However, this technology only provides information about a mere 2% of all CpG sites in the genome affected by DNA methylation [15]. More robust and affordable solutions became possible with the introduction of machine learning into the field, which has helped in detecting many epigenetic marks in the disease through computational methods using supervised learning or unsupervised learning. The main attraction of machine learning is its ability to process large amounts of clinical and biological data to extract meaningful patterns and results. It has its downside, however, due to the limited data size that can be used, its need for structured data, and traditional algorithms.

Many scientists are now inclined towards deep learning, which is a subset of machine learning that uses complex learning methods which mimic the learning processes of the human brain, therefore enabling more detailed and precise outputs can also avoid some limitations of the process.

It is possible, then, that a more thorough understanding of the mechanism of the epigenetics of AD can be achieved, since deep learning can be a more comprehensive option to fully extract meaningful data from AD to process all DNA methylation tags of AD and how the mechanism of the disease occurs in order to achieve a thorough understanding of diseases which will aid in the discovery and design of novel and effective therapies.

This leads to the final gap of AD, which is the therapeutic approach. Unfortunately, no cure has been discovered yet, and the available treatments are those that may change the progression of the disease by slowing the decline of memory and thought, like anti-amyloid antibody intravenous infusion therapy (Aducanumab, Lecanemab) [16] or simply treat the symptoms like behavior changes, cognitive symptoms, and psychological symptoms (Donepezil, Memantine, Pantethine, Suvorexant) [17]. It is still necessary to find all the pathological processes and mechanisms of how AD truly emerges and progresses. Which is why deep learning along with machine learning was suggested here since it has never been applied before in this context to attempt a new method of CpG identification.

Alzheimer's Disease and its Known Pathological Processes:

Alzheimer's disease is a very complicated neurodegenerative disease, in which a small percentage of cases is caused by genetic mutations or are familial [18], and in most cases, the disease is multifactorial and is subject to many risk factors, both environmental and genetic [19]. Scientists have discovered many genes that are responsible for the disease, and many hypotheses have been suggested for the explanation of the mechanisms in which AD emerges.

These hypotheses include the cholinergic hypothesis, amyloid hypothesis, tau propagation hypothesis, mitochondrial cascade hypothesis, calcium homeostasis hypothesis, neurovascular hypothesis, inflammatory hypothesis, metal ion hypothesis, and lymphatic system hypothesis [20].

Even though these hypotheses have been suggested (some even partially proven) for the explanation of how AD occurs precisely, they have yet to prove results when used for the treatment of the disease in clinical trials [21].

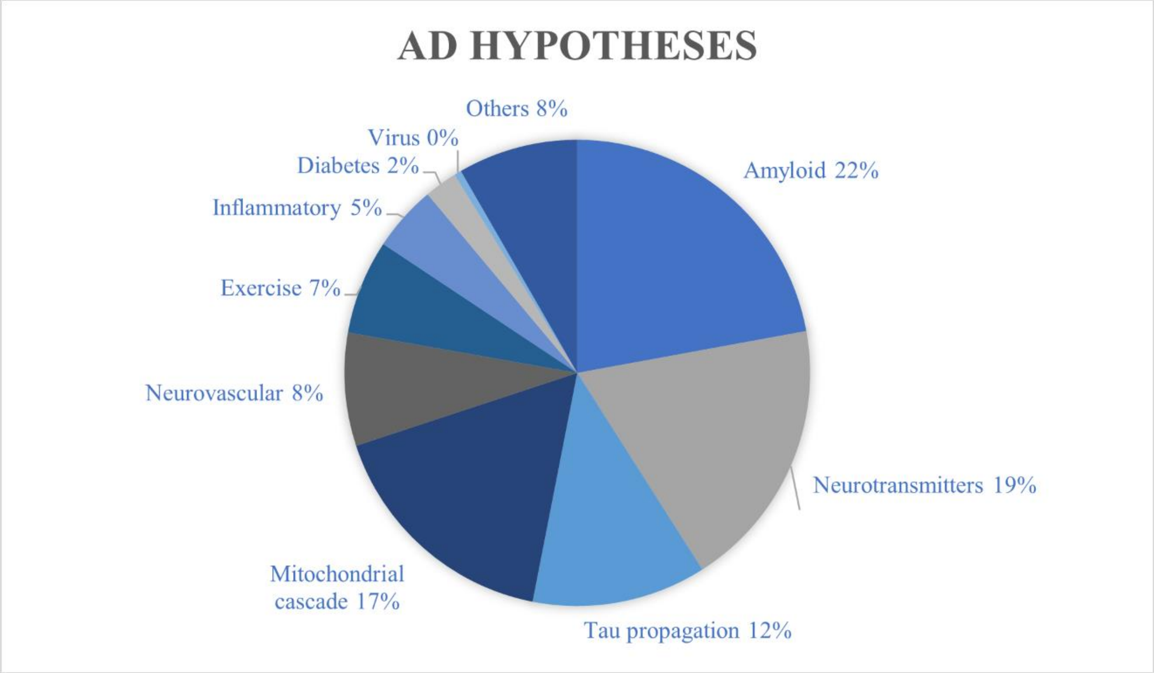


Figure 3 Percentage of the clinical trials tested for each hypothesis up to 2019 [22].

The two most common hypotheses regarding Alzheimer’s disease are the amyloid hypothesis and the tau hypothesis. The amyloid hypothesis, which was first postulated by J Hardy and D Allsop in 1991 [23] stated that the central mechanism by which AD occurs is thought to be an amyloid precursor protein (APP) gene mutation, and it was explained that the finding of the APP gene mutation in familial AD makes it clear that the deficiencies occurring in AD from neurotransmitters, to enzymes and receptor, are caused by *“amyloid disposition or abnormal APP processing”[23]*.

APP is a single-pass transmembrane protein – a protein that passes through the bilipid layer of the cell membrane and spans it only once – which plays a role in the health and growth of neuronal cells [24]. When it is properly metabolized and cleaved and then subjected to hydrolysis in healthy conditions, it is through the α pathway, in which APP is hydrolyzed by α-secretase and then by γ-secretase (*Figure 4*). This process results in soluble Aβ [25]. The second process (hypothesized cause of AD) is the β pathway, in which the hydrolysis occurs through β-secretase then γ-secretase to produce insoluble Amyloid beta (Aβ) [26] which is the root of AD neurodegeneration. In normal cases, a small amount of APP is hydrolyzed though the β pathway, but the Aβ protein resulting is eliminated by the immune system. Amyloid beta can be defined as

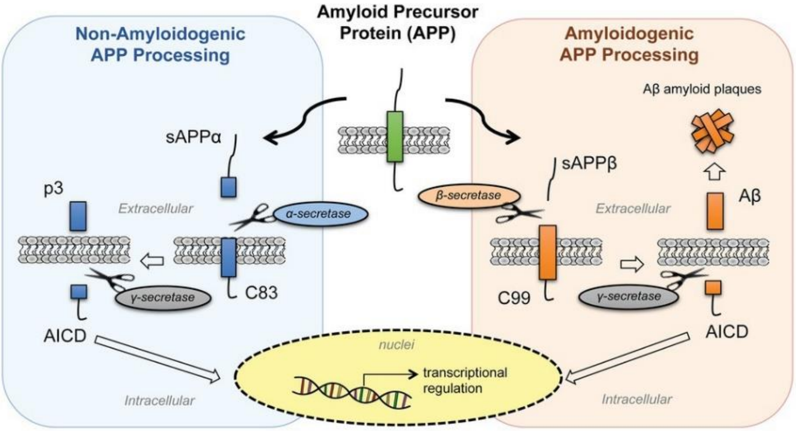


Figure 4 A comparison of α and γ pathways in which APP is disposed of, and β pathway which leads to Aβ accumulation [28].

a protein-formed fibril characterized by cross-β organization, and amyloids can be divided into either functional (like in bacteria) [27] or disease-associated (like in AD).

When mutations occur and the disease is present, the β pathway becomes dominant over α and γ which causes the excessive extracellular accumulation of Aβ protein that is insoluble and which the immune system cannot eliminate.

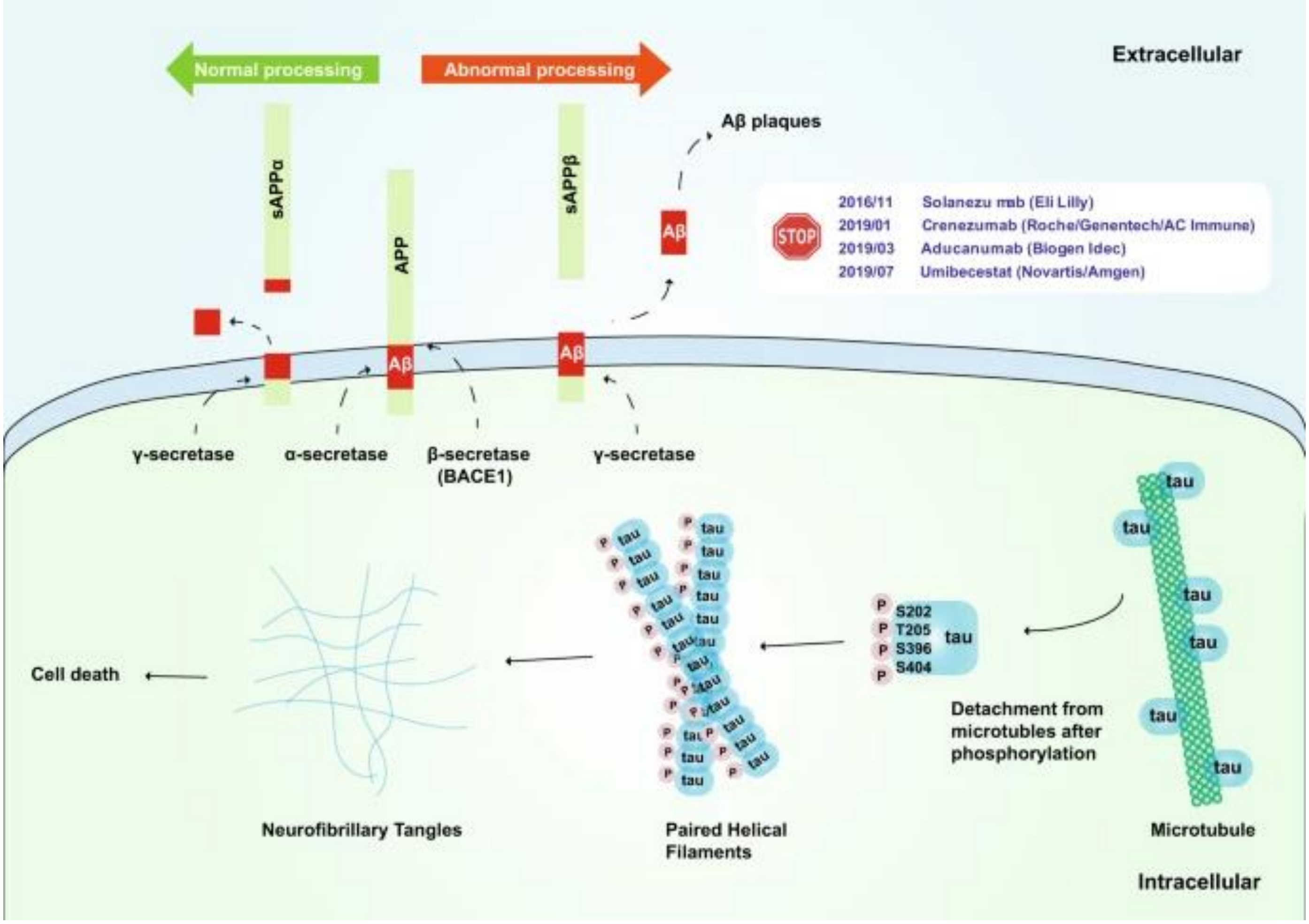


Figure 5 Describing both the amyloid (upper) and tau hypothesis (lower) along with the two pathways found regarding APP [22].

As to the mechanism in which Aβ protein is toxic to the neurons, it can be explained by the high concentrations that cause neurotoxicity to mature neurons due to the dendritic and axonal atrophy that is caused, and then followed by inflammation and then neuronal death [29].

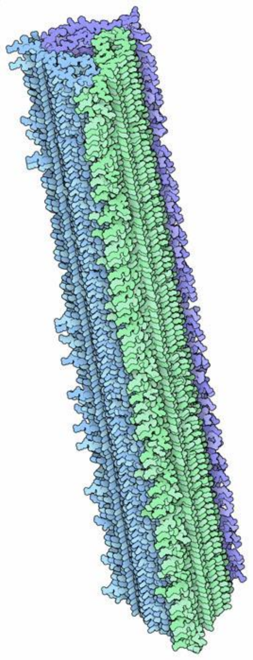


Figure 6 Amyloid-beta fibril from a patient with Alzheimer's disease, showing how the APP broken chain aggregates into long fibrils that stop the function of cells. Studied by x-ray diffraction, they have a 'cross-beta pattern' characterized by the accumulation of small peptides to form beta strands by stacking upon one another. The beta strands then use hydrogen bonds to form huge beta sheets with structures similar to spider silk in strength. [30]

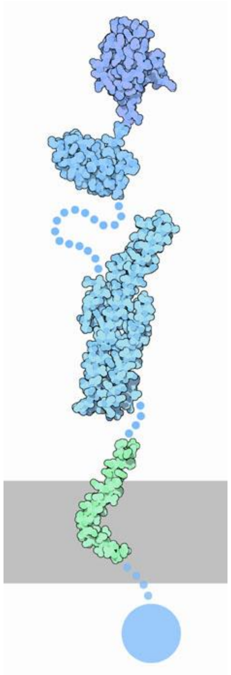


Figure 7 Amyloid-beta precursor protein. It can be fragmented into smaller pieces with the help of secretases. The larger portion (top) is released outside of the cell, normally, aiding in the control of neuronal growth, while the smaller (lower) piece remains inside the cell. [31]

Both the presence of neurotoxic Aβ (outside the cell) and tau (inside the cell) in the brain trigger immune responses from the microglia – the immune cells responsible for the protection of the central nervous system against infection and inflammation [32]- which attempt to dispose of dead cell debris, but when the microglia can't keep up with the increasing amounts of accumulation and dead cells, chronic inflammation begins. All of the processes mentioned earlier would therefore lead to: cell death, loss of neuronal communication, atrophy due to cell death, and chronic inflammation, which all lead to death in later stages [33].

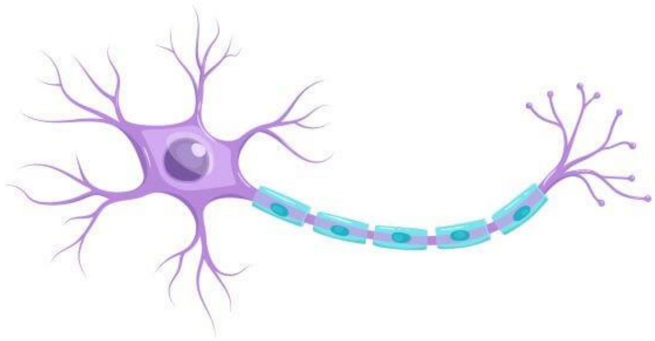


Figure 9 A simple healthy neuron, containing a cell body, axon, and dendrites [34]

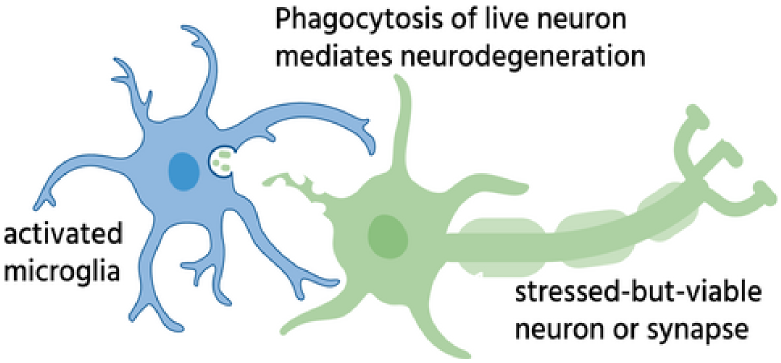


Figure 8 The main inflammation cause of AD, in which the microglia start attacking neuronal cells due to neurodegeneration, thereby increasing the problem at hand [35].

However, both amyloid and tau proteins have functional roles in the healthy normal brain. Amyloid for instance, has been found to be of use in protection against a wide range of cytotoxic factors such as UV and oxidative damage [36].

As for the tau hypothesis, in 1986, Kosik et al. found that the neurofibrillary tangles in the AD brain were made up of phosphorylated tau proteins [37]. Tau is a microtubule-associated protein in neurons. In the healthy brain, tau stabilizes the microtubules of the neuron. However, when it is phosphorylated, it dissociates from the microtubules and aggregates, forming paired helical filaments (PHF) which later transform into neurofibrillary tangles.

The tau hypothesis states that tau hyperphosphorylation precedes amyloid beta accumulation and that tau aggregation is the main cause of AD

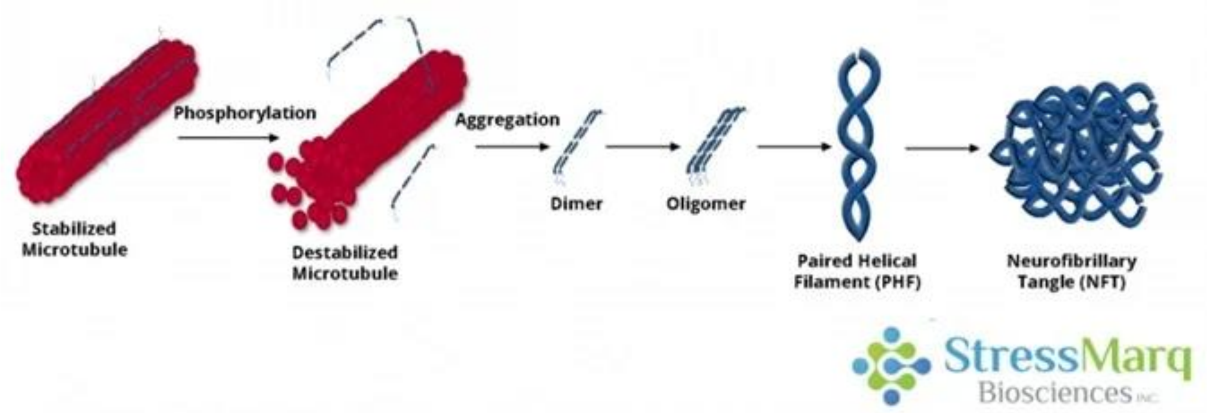


Figure 10 How microtubules transform in AD under phosphorylation into neurofibrillary tangles, characteristic of the disease [38].

Epigenetics: Definition and Mechanisms

The term epigenetics (epi “over” + genetics) means “What is above genetics.” It was coined by Waddington [39], and developed through many definitions to reach the definition it has today, which is the explanation of the interaction between the environment and genes.

Epigenetics can be described as the mechanism that turns the genes on or off (which leads to either producing a certain protein, or refraining from its production) without directly affecting the building blocks of the DNA. Changes in gene expression are passed on from one cell to its descendants [40]. Epigenetic modifications can be either DNA methylation, histone modification, or noncoding RNA (miRNA, siRNA, piRNA, lncRNA) [12].

The most studied epigenetic mechanism is DNA methylation, which is the addition of a methyl group to DNA, causing the turning “off” of genes, whereas the absence of methyl on the DNA can cause the turning “on” of the gene, and therefore enabling transcription [13]. It is controlled by the

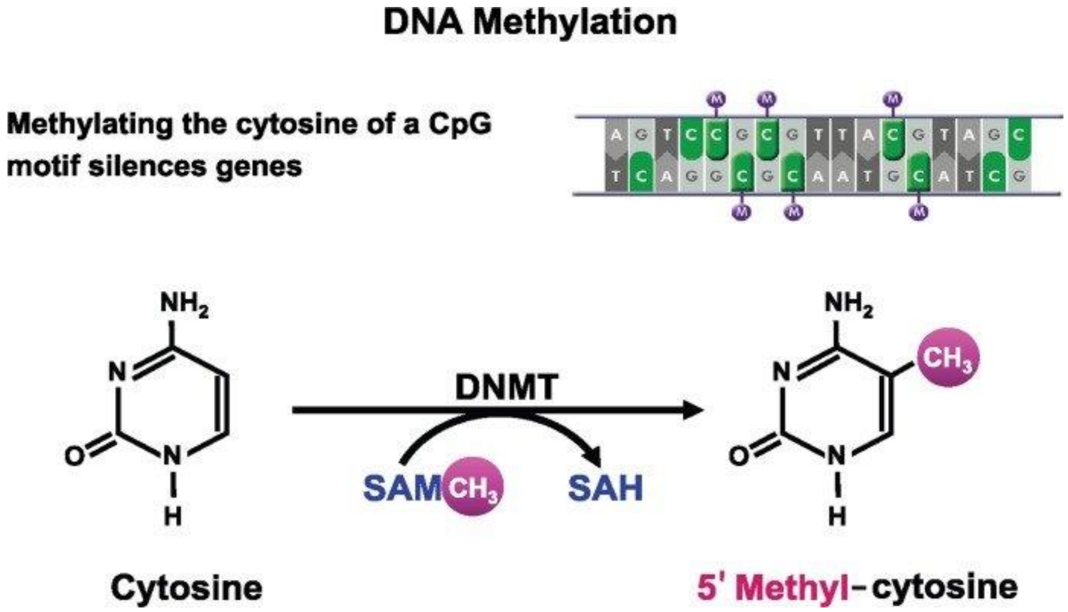

DNA Methylation
Methylating the cytosine of a CpG motif silences genes
A G T C C G C G T T A C G T A G C
T C A G G C G C A A T G C A T C G
M
NH$_2$
N
O
N
H
DNMT
SAM CH$_3$
SAH
CH$_3$
Cytosine
5' Methyl-cytosine

Figure 11 How cytosine is methylated into 5' methyl-cytosine [41]

DNA methyltransferases, DNMTs which are a group of enzymes. Even though DNA consists of a sequence of four bases adenine cytosine guanine and thymine, DNA methylation only concerns cytosine: The DNMTs usually modify cytosine in the DNA when it is present near guanine (referred to as cytosine-guanine dinucleotides or clusters), by the transfer of a methyl group from the universal methyl donor, S-adenosyl-L-methionine (SAM), to the 5-position of cytosine residues in DNA [42].

These cytosine-guanine dinucleotides or clusters (CpGs for short) are named CpG islands, and when these islands are located on the promoter region of a gene, the gene undergoes methylation, which results in the repression of the expression of that gene [43], and it is suggested that 70-80% of these CpG are methylated in somatic cells.

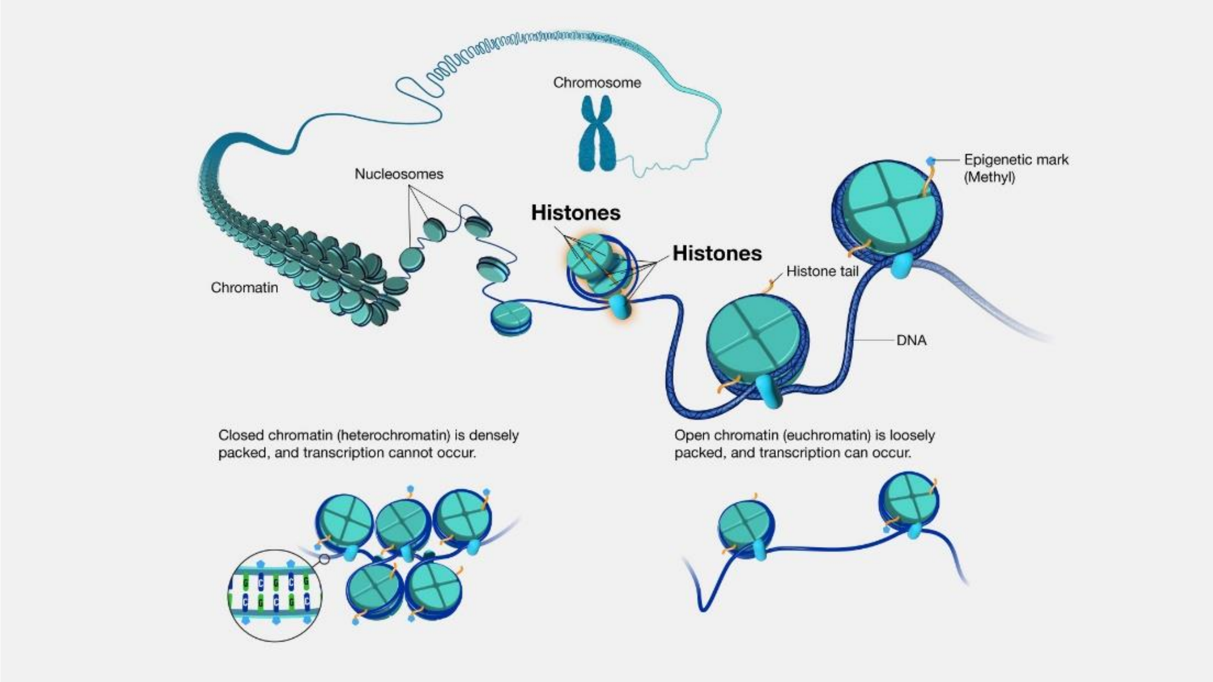

Chromosome
Nucleosomes
Histones
Histones
Epigenetic mark
(Methyl)
Histone tail
Chromatin
DNA
Closed chromatin (heterochromatin) is densely packed, and transcription cannot occur.
Open chromatin (euchromatin) is loosely packed, and transcription can occur.

Figure 12 the structure of histones of the chromosome and their relationship with DNA, along with a brief visual description of transcription regulation [44].

As for histone modifications, they occur by the adding of a chemical group on the histones themselves which can be described as the proteins that support the chromosomes by the tight wrapping of DNA on the histone complexes, and they also regulate the transcription of DNA (which is the copying of DNA to make more DNA). Their modifications can be characterized into histone acetylation, methylation, phosphorylation, ubiquitinylation, sumoylation, ADP ribosylation, deimination, proline isomerization, citrullination, formylation, succinylation, butyrylation, propionylation and crotonylation, and they occur at the posttranslational level. Of the above, for example, histone acetylation causes the chromatin to be exposed, thereby activating the gene (gene is “on”), whereas histone methylation would cause the chromatin to be closed, and consequently, the gene would be inactivated or “off” [45].

The last mechanism is noncoding RNA, which is all RNA transcripts that are not translated into protein. Some are functional such as rRNA and tRNA, and some have only recently been discovered to have an actual function as the role of "epigenetic factors" through the regulation of gene expression, which can be at the transcriptional, posttranscriptional, and translational levels [46]. It has been suggested recently that ncRNAs participate in DNA methylation and histone modifications along with gene silencing [47].

As of now, the "epigenetics" term continues to evolve and is now the reversible change in the effect the environment can have on genes, whether it be a positive or negative change.

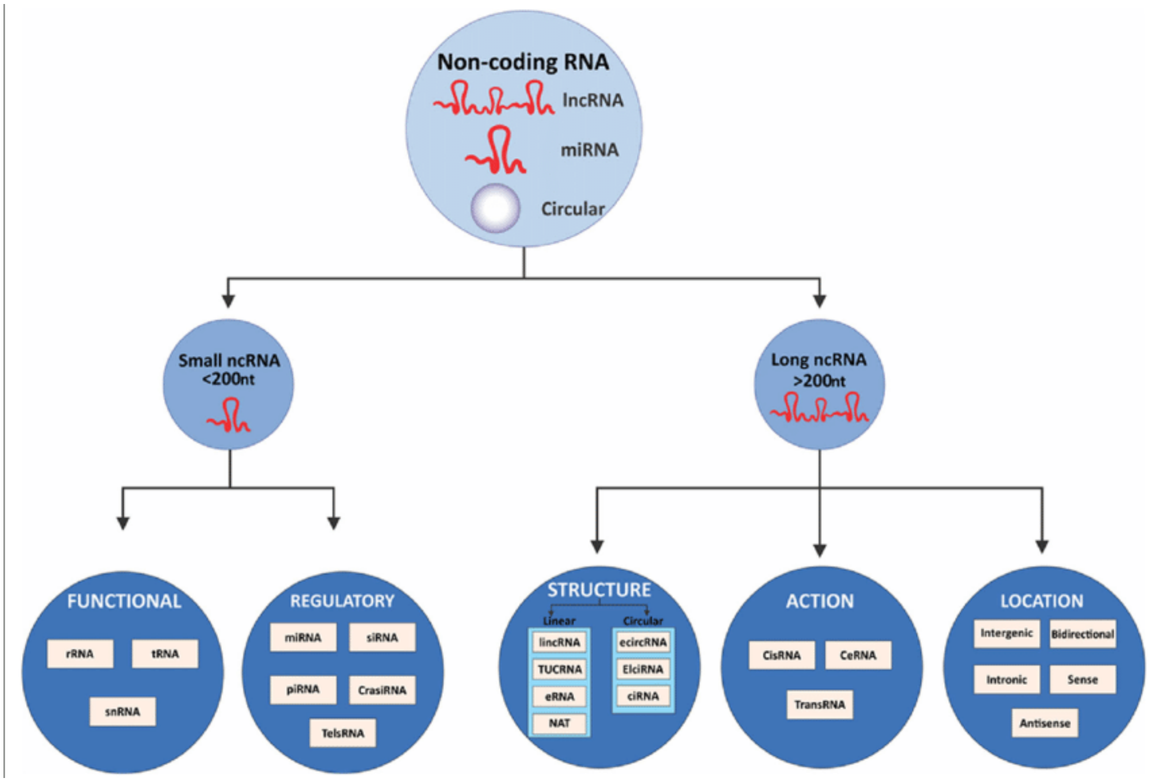


Figure 13 Different noncoding RNA [48]

Regarding the technologies responsible for the measurement and quantification of each epigenetic tag [49], DNA methylation can be studied using any of the following: PCR-based bisulfite sequencing, MSP, pyrosequencing, WGBS, HumanMethylation450, RRBS, MRE-Seq, MeDIP, ELISA-based assay, single-cell bisulfite sequencing, SMRT sequencing, nanopore sequencing, OxBS-seq.

Histone modifications can be measured with either, ChIP-PCR, ChIP-chip, ChIP-seq, or ELISA-based assay.

Whereas ncRNAs are detected using qRT-PCR, RNA-Seq, and HITS-CLIP.

Table 1 Most important genes in Alzheimer's disease [50].

Gene Symbol	**Gene name**	**Chromosome**
APP	Amyloid precursor protein	21q21
APOE	Apolipoprotein E	19q13.32
PSEN1	Presenilin 1	14q24.2
PSEN2	Presenilin 2	1q42.13

Genes and the epigenetic mechanisms of Alzheimer's disease

Even though the true mechanism of how AD occurs is still unclear, and genetic factors have been proven to be of cause in 70% of cases, the majority of genes believed to play a role in AD pathogenesis do not individually cause the disease through either an inflammatory pathway, cholesterol metabolic pathway, or others. These genes include amyloid precursor protein (APP) on chromosome 21, Presenilin 1 (PSEN1) on chromosome 14, and presenilin 2 (PSEN2) on chromosome 1 [50]. Change in these genes can result in abnormal protein production associated with AD. chromosome 14, and presenilin 2 (PSEN2) on chromosome 1 [50].

But another gene is also known to influence the risk of AD: the APOE (apolipoprotein E) gene [51], located on chromosome 19q13.32, remains the most prevalent and the strongest genetic risk factor, causing more than half of all AD cases.

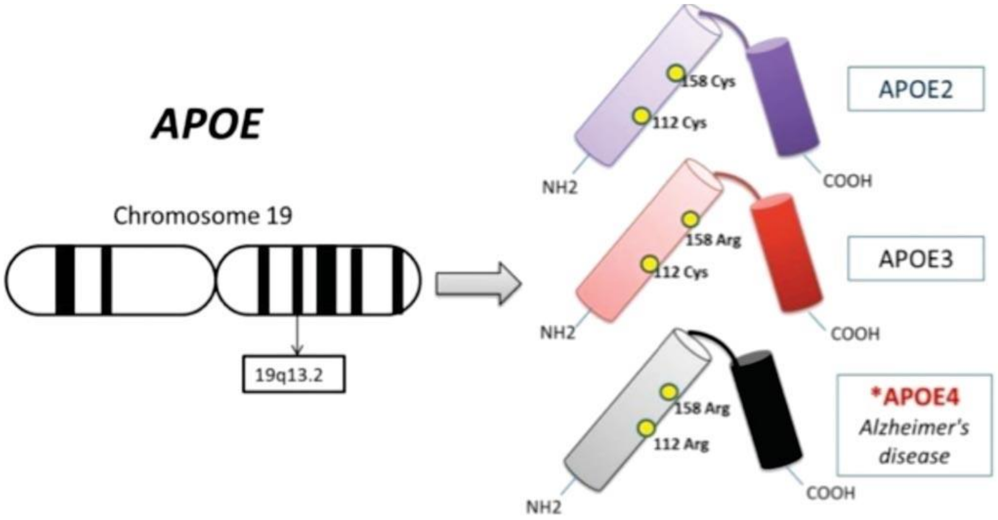


Figure 14 APOE can be found in three different allele variants, and only one is of danger of causing AD [52].

APOE has three predominant alleles in humans: ε2 (APOE2), ε3 (APOE3), and ε4 (APOE4) which is the main genetic risk factor that increases risk up to 15-fold in homozygotes, whereas APOE2 reduces risk and APOE3 is believed to be neutral. APOE's main function is the mediation of lipid transportation into the brain and periphery, but in pathological cases, the ε4 allele of APOE, ApoE4, causes seeding of Aβ plaques in the brain causing more amyloid accumulation. It also shows neuroinflammation and tau phosphorylation in the presence or absence of Aβ plaques, by inducing tau phosphorylation and cell death.

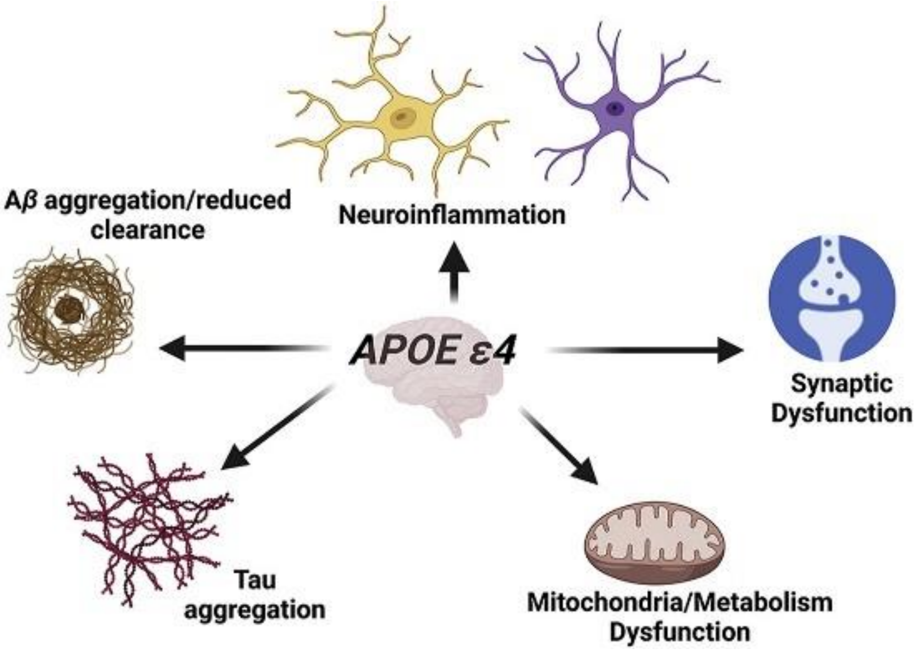

Aβ aggregation/reduced clearance
Neuroinflammation
APOE ε4
Synaptic Dysfunction
Tau aggregation
Mitochondria/Metabolism Dysfunction

Figure 15 All the effects of the APOE ε4 allele has on the aspects of AD [53].

The aggregation mechanism is suspected to be caused by the formation of tau/apoE complexes. This provides evidence of the role of understanding the role of apoE in tauopathy in identifying new approaches of therapy.

Therefore, this, and more studies like monozygotic twin studies (which play a major role in difficult diseases, to understand why one identical twin could present with a disease while the other remains with no disease), suggest that more factors are responsible that do not involve genetic factors. Furthermore, dietary and environmental factors of poor condition seem to be present in almost one-third of AD patients in studies [54] [55], such conditions involve low educational status, diabetes mellitus, depression, hypertension, . . . etc. which are all considered major contributing risk factors for AD and other neurodegenerative diseases as well. The main process by which this occurs is still not completely clear, but most studies suggest that early exposure (prenatal and in childhood) to metals, electromagnetic fields, and poor diet can have an impact on development and diseases later in life by perturbing how methylated CpGs interact with binding proteins [54]. This, therefore, leads to epigenetics, since epigenetics is the method by which environmental stimuli affect gene expression [56].

In AD, and other age-related diseases such as Parkinson's disease, senescent cells (aging cells) of the brain are accumulated, blocking tissue regeneration and causing inflammatory responses. Senescence, or cell aging, is a state of cells that is both non-proliferative and pro-inflammatory and is an active state in the human body. But, in age-related disorders, senescent cells start competing with normal cells, causing epigenetic changes that corrupt the epigenetic mechanism [57]. Therefore, DNA hypermethylation, histone deacetylation, and a repressed chromatin state all become characteristic of the disease, thereby changing gene expression by gene silencing [58].

Epigenetics play a major role in the pathophysiology of AD, for example, transcriptional activator sites *AP2*, *SP1*, and *GCF* have been proven to be of key roles in the epigenetics of tau hyperphosphorylation which causes the collapse of neurons and accumulation of neurofibrillary tangles, which is a hallmark of AD pathogenesis.

As for the epigenetics of APOE, it has recently been discovered that DNA methylation is differentiated in APOE CG islands in AD [58] [59].

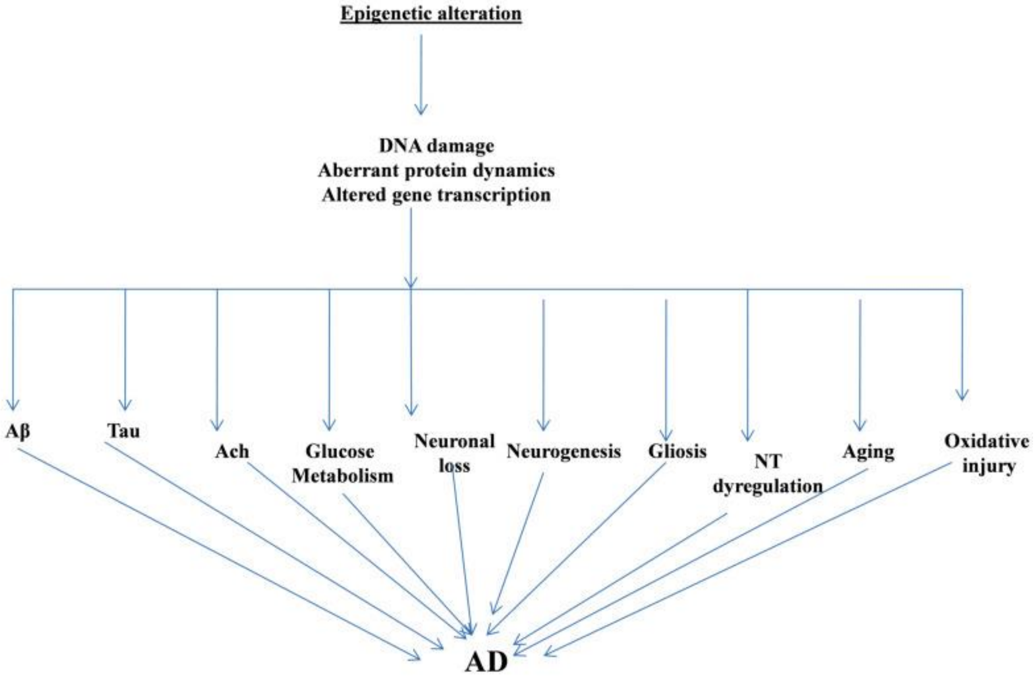


Figure 16 The Epigenetic alterations influencing Alzheimer's Disease from various sides. Aβ: Amyloid Beta; Ach: Acetylcholine; NT: Neurotransmitter; AD: Alzheimer's disease [58].

Impact of AD

According to the Cleveland Clinic, an estimated 24 million people worldwide suffer from AD [60], and the Alzheimer's Association declares that "An estimated 6.5 million Americans age 65 and older are living with Alzheimer's dementia today" [33]. It starts with short-term memory loss and progresses to a stage of declined coordination, mood swings, and unpredicted behavior which all require constant need for care from others.

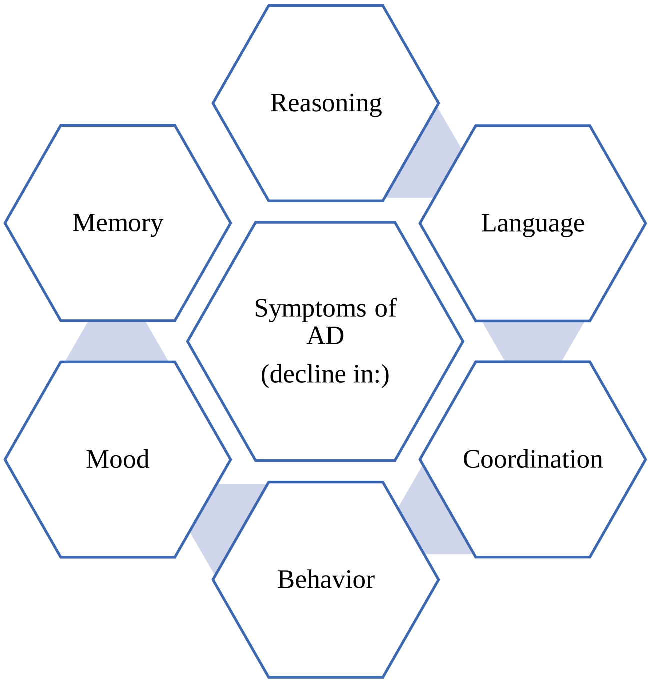


Figure 17 Symptoms of AD involve a gradual decline in some, most, or all of the above [60].

Measuring Alzheimer's Disease

To measure the severity and progression of the disease, many measurements are used that describe and categorize the symptoms either histologically, from imaging, or from direct observation by the clinician. The main traits used for measuring AD are:

1. Beta-amyloid load which quantifies the global Aβ burden from PET imaging *in vivo*.

2. Neurofibrillary tangle density (NFTs) which are measured in the cortex using immunohistochemistry with the use of phosphorylated-tau-specific antibodies [61].
3. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD) is semiquantitative because it relies on the clinician's assessment along with the amount and distribution of silver-stain-identified neuritic and diffuse plaque [62].
4. Global AD pathology burden which quantitavely summarizes the AD pathology measured from silver staining from three pathologies of AD: neuritic plaques, diffuse plaques, and neurofibrillary tangles, taken from 5 regions from the brain [63].
5. Cognitive trajectory/decline based on performance of patients in various examinations.
6. Braak staging: a semiquantitative measure to estimate the severity of neurofibrillary tangle pathology. Silver staining is used for the visualization of cortex areas frontal, temporal, parietal, entorhinal cortex, and the hippocampus [64]. The diagnosis depends on the opinion of the neuropathologist along with an algorithm that quantifies the amount of silver staining of neurofibrillary pathologies. It consists of

six stages from one to six, and it is coded as binary (controls: stages 1, 2, 3), (affected: 4, 5, 6)

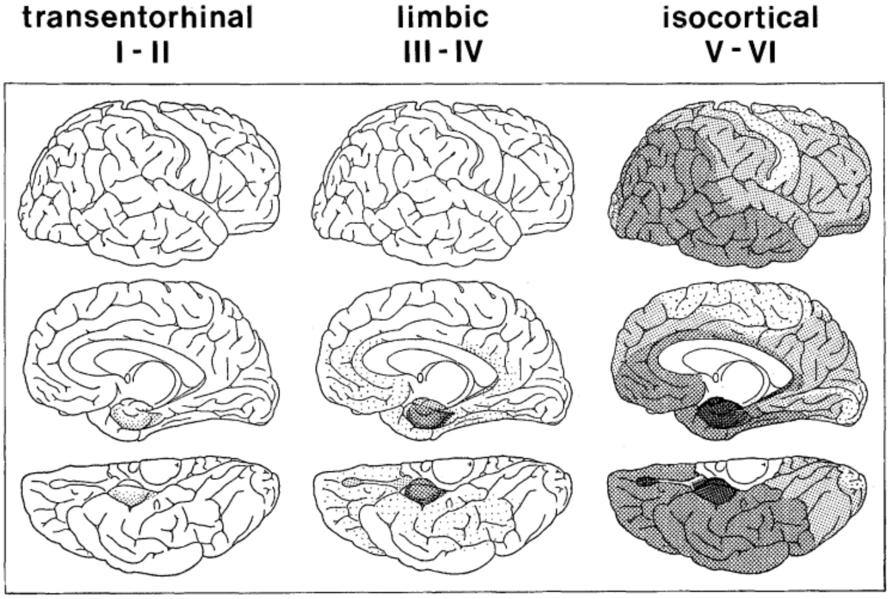


Figure 18 Changes in neurofibrillary distribution among different Braak stages, the density of the shading indicates the increasing severity of NF changes [61].

Therapy:

Unfortunately, to date, no treatment exists to erase Alzheimer’s disease or cure it. Some medication can be administered for slowing down symptom progression.

So far, only five drugs have been approved by the FDA which are tacrine, donepezil, rivastigmine, galantamine, and memantine, with little success. They are only targeted towards disorders of the memory without an actual cure or delay of the disease [65].

Impaired memory and concentration loss is said to be associated with a loss of cholinergic neurons. Therefore, most approaches nowadays target the promotion of cholinergic synapses to lessen neuron toxicity or stop plaque formation.

Epigenetics play a more important role in finding new therapy than genetics do, since the main goal of epigenetic therapies is the reversal of the epigenetic mechanism itself, which therefore leads to the need for better insight and understanding of AD. For example, vitamin B deficiency leads to hypomethylation of the (*GSK3β*) gene at the promoter region, which leads to the expression of its protein kinase (glycogen synthase kinase 3β), causing hyperphosphorylation of tau, leading to cell death [66].

The same hypomethylation mechanism causes gene expression to genes responsible for cell death and neuroinflammation such as *(BIN1)*, *(CR1)*, *(CD33)*, *(TNF-α)*. Since epigenetic mechanisms may appear in early asymptomatic stages of the disease, and since many can be reversible, it is believed that epigenetics could be the key to finding new treatment by using **epidrugs** (epigenetic-based drugs), that target DNA methylation or other mechanisms of epigenetics [65].

Epidrugs could be activators or inhibitors of DNMTs (DNA methyltransferases), histone deacetylase inhibitors, sirtuin activators, modulators of histone acetylation and histone methylation, as well as RNA interference analogs [65].

DNA methylation activators, for example, attempt to restore DNA methylation conditions, which could put the metabolic pathways back on track. Diets rich in Vitamin B complex can be of benefit to target this method of treatment by increasing the SAM/SAH ratio, reducing inflammation, and toxicity caused by Aβ aggregation. Several clinical trials are currently working on Vitamin B6 and folate in the hopes of reducing cognitive impairment by reducing homocysteine levels [65].

As for DNA methylation inhibitors, they are small molecules or natural products that attempt to stop the hypermethylation of pathogenic genes, because even hypermethylation can cause neurodegeneration. Clinical trials are underway for the epigallocatechin-3-gallate (EGCG) which is the main polyphenol of the green tea (*Camilla sinensis*). It is a DNMT inhibitor which is currently under trial to test its efficacy in preventing Aβ aggregation to toxic polymers by binding directly to the unfolded peptide. Other products include curcumin derivatives, bioflavonoids, catechins, and psammaplins [65].

It is worth noting that in the pharmacoepigenomics field, recent evidence suggests that epigenetic changes can determine the pathogenesis of medical conditions along with drug response and resistance. This depends on gene variants responsible for drug response, along with the epigenetic modifications of the genes which can change their expression.

For example, an epigenetic modification in genes that play a role in drug response in AD is hypomethylated upregulated mRNA of genes ABCA1, ABCB1, ABCA7, and SLC24A4 [65].

Amyloid clinical trials, however, which include active or passive immunotherapy or anti-amyloid antibodies have shown no efficacy in the past 20 years. Since active immunotherapy has shown clinical complications, passive immunotherapy has shown repeated failures [67], and anti-amyloid antibodies aducanumab and BAN2401 still need much research [16], the tau line has been of more interest. Therefore, Braak staging, which is a measure of the pathology of neurofibrillary tangles which are made of tau protein, is the new direction for new therapy attempts – laboratories can design tau-based therapies, which are currently in progress [68]. It should be noted that several antibodies to the tau protein are currently in the clinical trial phase, and even some vaccines as well.

As is known nowadays, a strong correlation is shown between neurofibrillary tangle accumulation and cognitive function deterioration (disease worsens as the Braak stage progresses). Therefore, strong evidence suggests that the etiological role of tau protein cannot be ignored and must be further looked into. And, even though tauopathy is still not completely understood due to the absence of extensive research, clinical trials, and even epigenetic studies, it is an essential key to unlock and solve the mysteries of AD.

This thesis utilizes a deep learning model that may help in identifying new methylated CpG sites in the genome, which offer better understanding of AD pathogenesis, and especially how the disease evolves throughout the Braak stages and how tau and NFTs distribute.

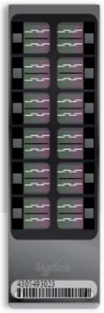
Analysis of the Epigenetics of AD

AD still remains partly in the dark regarding what exactly occurs in epigenetics of the brain and the body, which suggests the growing need for proper analysis methods that offer a more comprehensive analysis of the disease

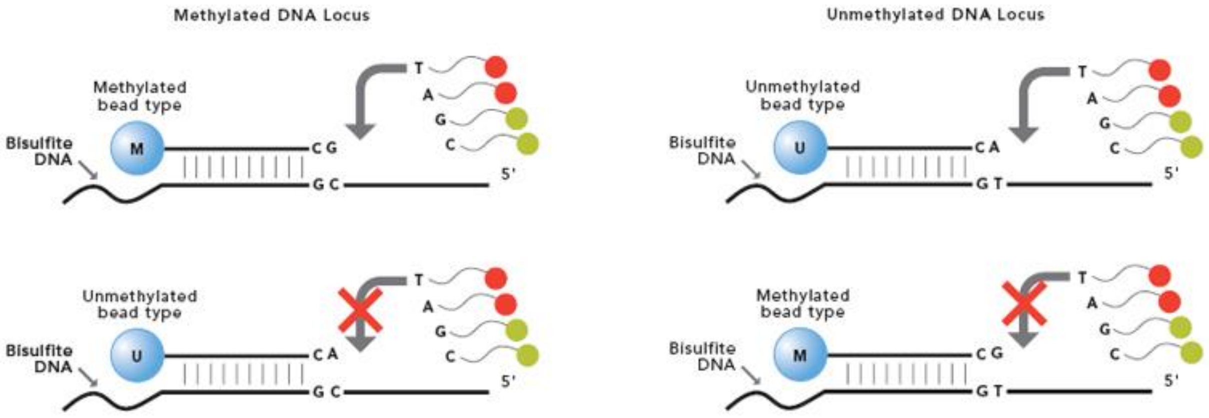
Classical/Present methods of analysis of AD epigenetics

The current approach of analysis of DNA methylation in AD is with technologies such as PCR-based bisulfite sequencing, MSP, pyrosequencing, RRBS, MRE-Seq, MeDIP, ELISA-based assay, single-cell bisulfite sequencing, SMRT sequencing, nanopore sequencing, OxBS-seq, HumanMethylation450, and WGBS [49]. All of which has strengths and weaknesses, but the main focus will be on HumanMethylation450 and WGBS since they are the technologies referenced in this study.

Profiling of DNA methylation: Infinium® HumanMethylation450 BeadChip



To obtain DNA methylation (DNAm) data, in epigenetic studies studying DNAm, on a genome-wide scale, it is usually preferred to use the Illumina Infinium® HumanMethylation450 BeadChip array, which covers approximately 480000 CpG sites and all at the single-CpG-site-level. Its method works on the basis of detecting methylated cytosine in CpG islands based on “highly multiplexed genotyping of bisulfite-converted genomic DNA” [69]. It requires no PCR, is relatively affordable and less laborious, and takes up to 12 samples and an amount as small as 500 nanograms of genomic DNA. It contains two probe types: Infinium I (n=135501) and Infinium II (350076). In Infinium I, every CpG site becomes the target of two 50bp probes (one for methylated density M, the other for unmethylated density U). In Infinium II, however, only one probe is utilized for the distinction of M and U through dye colors (green and red). When treated with bisulfite, the methylated cytosine base remains unchanged, while the unmethylated cytosine converts to uracil. The assay therefore identifies these two different bases that become chemically differentiated loci by utilizing two site-specific probes. (M bead type for the methylated locus, and U for the unmethylated locus). The probes incorporate a labeled ddNTP stained with a fluorescent reagent, and the level of methylation of a single locus is determined by calculating the ratio of fluorescent signals between the methylated and unmethylated sites.



The methylation value of a single CpG site is referred to as the β-value and is calculated as [70]:

$\beta = M/(M+U+\alpha^{*})$

M = methylated signal intensity (M>0)

U = unmethylated signal intensity (U<0)

$^{*}\alpha$ = 100 generally, as recommended by Illumina. Alpha is usually added to M and U for the need of stabilization of the beta value when M and U are small.

Under perfect conditions, when beta is zero, it means that all copies of the certain CpG sites are unmethylated, whereas a value of one means that every copy was methylated.

Another unit used for analysis is M-value [71], which is the logit-transformed β-value:

$M = log_2(\beta/(1-\beta))$

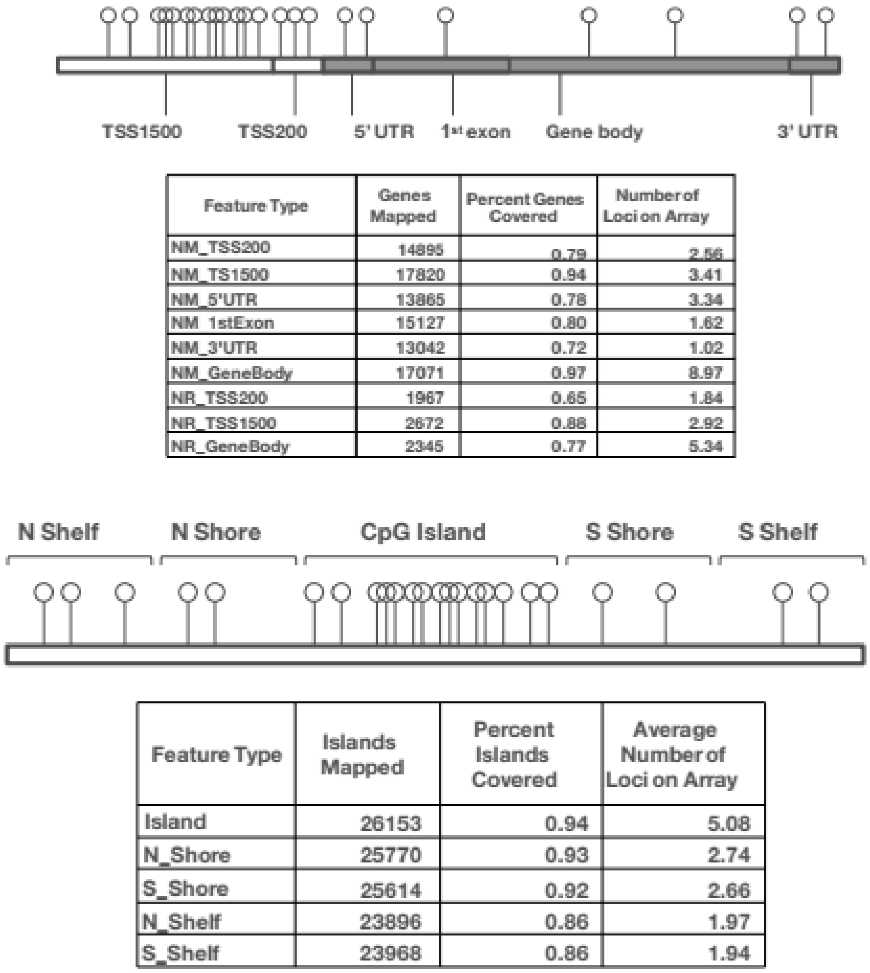
or

$M = log_2((M + \alpha)/(U + \alpha))^{*}$

*This is an alternative index that is not bounded by 0 or 1 [72].

In general, β-value is more preferred than M-value do its having more biological meaning and is more intuitive which helps reach more informative results. Even though it shows some heteroscedaticity when caught in the range outside of the middle methylation range (0.2-0.8) [73], it still remains better than M-value for gleaning biological insight.

The outcome offers a range of more than 480 thousand CpG sites and also covers 99% of RefSeq genes along with regions of low CpG island density that might be missed entirely using other methods. The target of coverage is promoter region gene sites, 5’UTR, first exon, gene body, and 3’UTR, and this enables coverage of 98% of the islands, island regions, island shores and island shelves [74].



Feature Type	Genes Mapped	Percent Genes Covered	Number of Loci on Array
NM_TSS200	14895	0.79	2.56
NM_TS1500	17820	0.94	3.41
NM_5'UTR	13865	0.78	3.34
NM 1stExon	15127	0.80	1.62
NM_3'UTR	13042	0.72	1.02
NM_GeneBody	17071	0.97	8.97
NR_TSS200	1967	0.65	1.84
NR_TSS1500	2672	0.88	2.92
NR_GeneBody	2345	0.77	5.34

Feature Type	Islands Mapped	Percent Islands Covered	Average Number of Loci on Array
Island	26153	0.94	5.08
N_Shore	25770	0.93	2.74
S_Shore	25614	0.92	2.66
N_Shelf	23896	0.86	1.97
S_Shelf	23968	0.86	1.94

Figure 19 Regions covered by HumanMethylation450 BeadChip [74].

Whole Genome Bisulfite Sequencing (WGBS)

In comparison with the technology of HumanMethylation450, WGBS is a protocol that detects the methylation of DNA at the fifth position in cytosine (5mC) [75] in *genomic DNA*, which in whole amounts to 28,084,558 CpG sites in the human reference genome [76], as opposed to the former which gives an output of approximately 480,000 sites. Genomic DNA is first treated with sodium bisulfite, sequenced, and finally it gives out single-base results of methylated cytosines of the genome. When treated with bisulfite, the unmethylated cytosine is deaminated to uracil, and when sequenced, is converted to thymidine.

As for methylated cytosine, it resists deamination so it is read as cytosine without change [77].

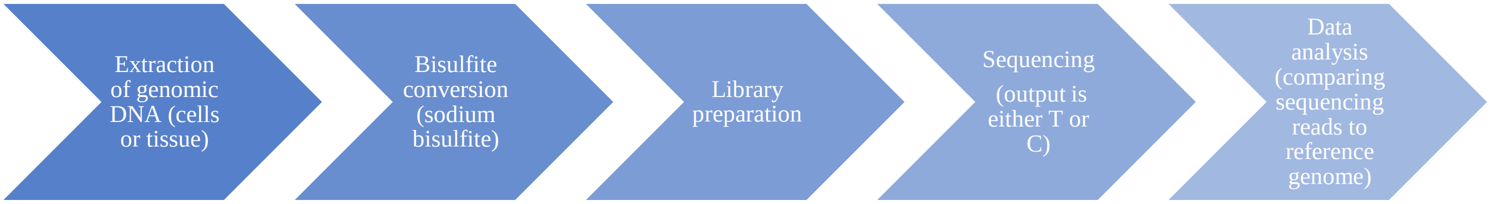

Extraction of genomic DNA (cells or tissue)
Bisulfite conversion (sodium bisulfite)
Library preparation
Sequencing
(output is either T or C)
Data analysis (comparing sequencing reads to reference genome)

Figure 20 The process in which WGBS works, from the extraction of DNA, to bisulfite treatment and 5' and 3' tagging, to the introduction of Illumina adapters using PCR amplification [78].

Epigenome-Wide Association Studies (EWAS)

Are studies that concentrate on the relationship between a phenotype (i.e. disease presence) and the underlying epigenetic variants. They are more and more accessible every day due to the decreasing cost which enables wider usage along with the emergence of new bioinformatics pipelines that help study the epigenetic mechanisms much more efficiently on a genome-wide scale. It can study any epigenetic mechanism. DNA methylation, for example, is studied through bisulfite conversion (mentioned earlier).

EWASs are mostly conducted using unrelated case-control and longitudinal designs, but there are also family studies and disease-discordant monozygotic twins.

Longitudinal studies can discriminate the relationship between epigenetic changes and phenotype in the gradual stages of the disease (right now, they only measure two time points), they are beneficial due to recruitement before disease and their avoidance of confounding and bias. A combination of longitudinal studies and monozygotic twin studies could be of double importance due to the benefit of no genetic difference or influence

Case-control studies, however, can dive into the understanding of dichotomous traits and methylation – they are considered more practical and affordable in comparison with longitudinal studies, due to the availability of actual cohorts that one can compare with, but since they are retrospective, environmental control and genetic confounders are potential disadvantages [79] [80].

Family studies, for instance are beneficial for the analysis of potential epigenetic inheritance, but family cohorts are not as plenty as other available cohorts [79].

Monozygotic twin studies are very beneficial for epigenetic studies, but the same issue with family studies arises, which is the scarcity of such cases, along with the need for longitudinal recruitement to better understand causes of disease [79].

Identification of CpGs

CpGs do not have a database accurately naming every cytosine guanine dinucleotide the way SNPs or genes do, so Illumina developed a new method for the consistent designation of CpG loci based on actual or contextual CpG locus sequence [81], by using the flanking sequences to give a unique CpG locus cluster ID. It only relies on sequence information. It also concerns DNA strands, because of the symmetry of CpGs on forward and reverse DNA strands, and it is important that we refer to the cytosine -without ambiguity - whether it is on the forward strand, reverse strand, or on both. Therefore, a naming is applied for the top and bottom strands as well (TOP/BOT). For example:

Table 2 Example of CpG naming and coordinates

Cluster CG#	Chromosome	Coordinate	Genome build	Sequence	TOP/BOT
cg00009407	14	88,360,674	36	...GGCG[CG]CTGC...	BOT
cg00003994	7	15,692,387	36	...TCTT[CG]TTGG...	TOP
cg00005847	2	176,737,319	36	...ATGG[CG]CTTT...	BOT

We use the sequences which flank the CpG to generate CpG cluster IDs (cg# column). A 122-base sequence is used which is made up of 60 bases that flank either side of the CpG locus (Sequence column). One CpG cluster may have several other members that can map onto different loci in the genome (**only if** they have *identical* sequences).

For CpG identification, we need three basic pieces of information to know every individual member of a cluster (above).

Chromosome number

- Chromosome numbers range from 1 to 23, in this study only 1 to 22, to evaluate without the gender chromosome.

Genomic coordinate

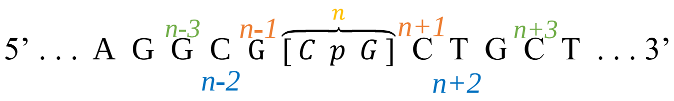
- One for G, one for C, the lesser of C or G is used as the CpG locus coordinate.

Genome build

- GRCh38 or hg19, for example, in which both are human genomes, but GRCh Build 38 is more accurate and provides alternative alt_sequences and it is the primary build referenced in studies.

As for the strand (TOP/BOT), a sequence walking method is used, and the CpG dinucleotide is considered as position ‘n’. The base before it would be ‘n-1’ and before is ‘n+1’, then ‘n+2’ and so on until an unambiguous pairing is found (unambiguous pair: two bases that are in equal distance from the CpG in which only one of the two is an A or T like A/G or A/C or T/C or T/G).

If the A or T of the unambiguous pair is from the 5’ side of the CpG then it is TOP, if on the 3’ side then it is a BOT. The CpG sequence below is on the TOP strand because the ‘n-2’, ‘n+2’ pair is the first unambiguous pair C/T, and since A or T are responsible for determining TOP or BOT, we have T which is on the 3’ side, we determine that the strand is BOT.



Chapter 2: Relevant Studies & Reference Study

Modern / Advanced methods of epigenetic analysis of AD

We are now in the era of artificial intelligence, including its most beneficial subsets in the field of biology/bioinformatics – machine learning and deep learning (AI/ML/DL).

Simply put, artificial intelligence is a computational field that enables the machine/model to extract information from large amounts of data which is not humanly possible. Its simpler form is machine learning, that relies on the basis of training a model on information and then testing its validity, followed by the prediction it makes on new unexplored data.

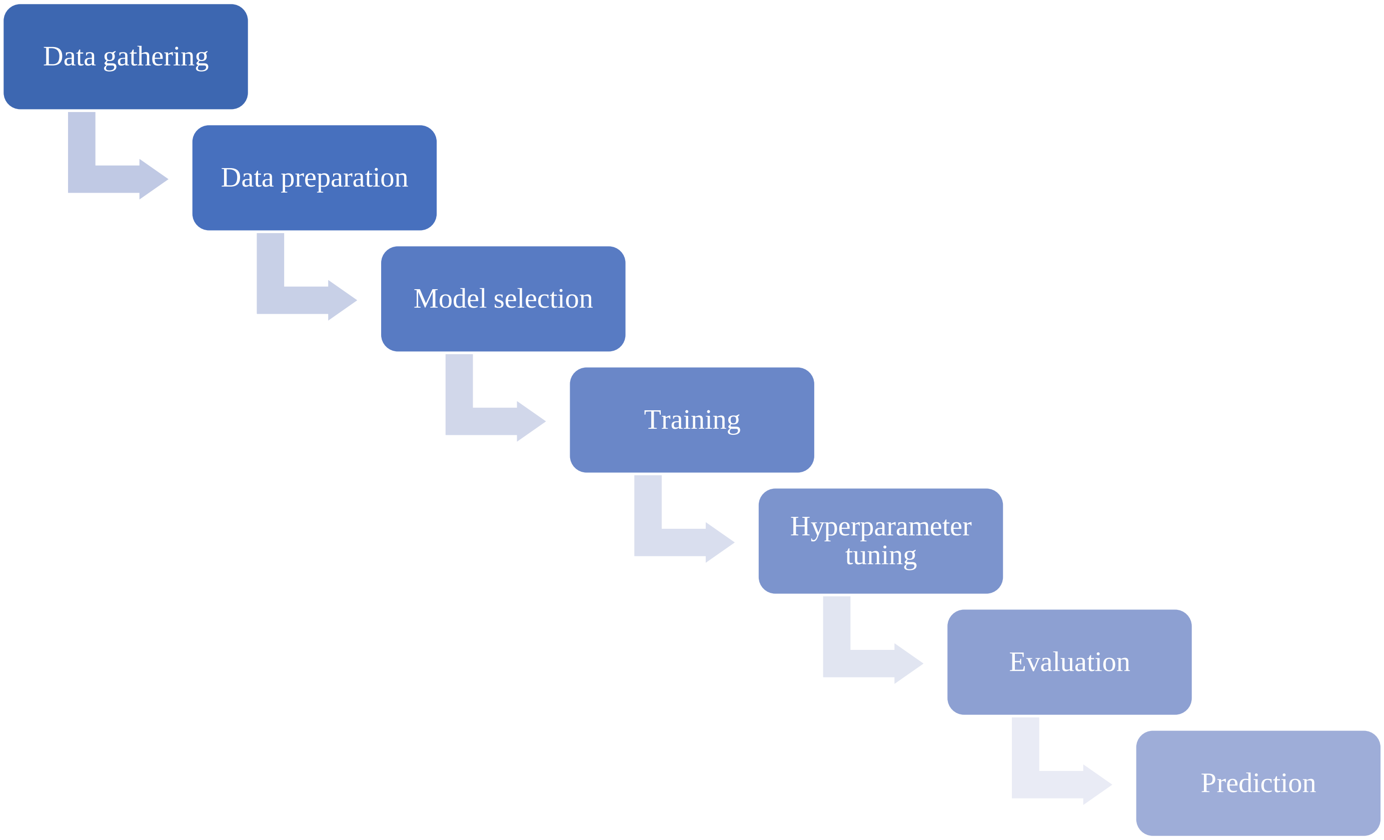


Figure 21 Steps of machine learning [82].

1. **Data gathering:** Important and time-consuming, the data chosen for a model is essential to its performance. The data used in this study for example, was obtained/gathered from EWAS data and ENCODE WGBS data for proper training, testing, and prediction.
2. **Data preparation:** Includes choosing the proper file format, columns, lines, and preprocessing (units, nominal or numerical, . . . etc). txt files, bed files, csv files, all can be of benefit in such studies. Here, we used bed files and csv files that are descriptive of the fields needed for the model.
3. **Model selection:** The choice of the model/algorithm used in learning is also essential, since some algorithms are supervised (labels are used as input in the process), or unsupervised, and others are semi-supervised. The algorithm used for some data may not be of benefit for another type, depending on what is hoped to be achieved from the learning process. Some algorithms include SVM, linear regression, random forest, . . . etc. The model used in our reference study, for example, applied random forest, xgboost, logistic regression with L2-regularization and SVC with linear kernel, they can be used as an ensemble, by calculating the performance of the solo algorithms or a combination of them.
4. **Training:** In case no tuning is needed, the model can start to learn the weights from the data at hand [83]. This is the “learning” step in the process of machine learning, and the outcome of the predicition heavily relies on this step [84].
5. **Hyperparameter tuning:** For best behavior/results of model, this process can help modify and tweak the hyperparameters of the model like modifying the learning rate of the model [85]. For example, when using multiple models (to choose an ensemble

model), one must select the optimal hyperparameters for each base learner (model) and the best combination as well [86].

6. **Evaluation:** Several evaluation metrics can be used to assess the performance of the model, including, accuracy, precision and confusion matrices [87].
7. **Prediction:** Ready for performance on real unknown data.

Recent papers using AI for AD and/or epigenetics

Since epigenetics of AD have only recently taken interest, only a handful of studies address them from an AI perspective. For example, many studies have utilized AI and its subfields for the diagnosis of AD or predicting its stages, using either clinical cognitive information, or brain imaging techniques as data for their models. The table below sorts some of the prominent applications:

Table 3 Artificial intelligence studies on Alzheimer's disease

Method	Study title	Use	Reference
DL	*Multimodal deep learning models for early detection of Alzheimer's disease stage*	Use of DL (3D CNNs) for the analysis of MRI, SNPs, and clinical test data for the classification of patients into AD, mild cognitive impairment, or controls	[88]
DL	*Deep Learning Approach for Early Detection of Alzheimer's Disease*	Utilizing an end-to-end framework capable of early diagnosis of AD and classification of AD stages from images using CNNs.	[89]

DL	*Development and validation of an interpretable deep learning framework for Alzheimer's disease classification*	Identification of unique AD signatures from multimodal inputs of MRI, age, gender, and Mini-Mental State Examination score using a fully convolutional network.	[90]
ML	*Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models*	Employment of Decision Tree, Random Forest, Support Vector Machine, Gradient Boosting, and Voting Classifiers, for the identification of bets parameters for AD prediction, all using imaging data.	[91]
ML	*A Novel Approach Utilizing Machine Learning for the Early Diagnosis of Alzheimer's Disease*	Design of a ML model comprising of GaussianNB, Decision Tree, Random Forest, XGBoost, Voting Classifier, and GradientBoost using imaging data for early diagnosis of AD	[92]

Reference study

A machine learning approach to brain epigenetic analysis reveals kinases associated with Alzheimer's disease [62]

This study conducted by Huang et al. developed a machine learning model depending on brain epigenetic data from Alzheimer's disease patients in the aim of identifying novel CpG loci across the whole genome, and was able to identify hundreds of new CpGs that cannot be detected using classical methods like HumanMethylation450 used in EWASs (ergo, the name: EWAS*plus*). It used supercomputers, extensive funding, and enormous data and resources, along with using AWS, Jupyter, R language, and Python. EWASplus is a supervised machine learning binary classifier, that is trained from data derived from array EWASs and several features were used from WGBS, RNAseq, ATACseq, and more, all include genomic and epigenomic profiling data. Several traits of AD were used to further supply information, which are beta-amyloid load, Braak staging, cognitive decline, global pathology, neurofibrillary tangles, and CERAD score. The model was then used for predicting new loci after its training, using an ensemble learning method that included regularized logistic regression (RLR), support vector machine (SVM), random forest (RF), gradient boosting decision trees (GBDT). It identified several new loci, and its performance was further checked and evaluated by conducting targeted bisulfite sequencing that validated the *in-silico* predictions reached. Its results reached that *"predicted CpGs are 2.2 times more like to be associated with AD ($p < 1.00 \times 10^{-9}$) than negative control CpGs"*

Chapter 3: Machine Learning/Deep Learning Algorithms Used in Study

The models used in this study

For this thesis, Google Colab was employed to run the Python code of this study.

`Random Forest Regressor`: a machine learning model used by utilizing several decision trees to identify the best decision and give an output of the best overall performance. It also ranks the importance of features used in the process [93].

`Keras Regressor` is a deep learning model, with the ability to predict continuous labels, also used in this study for the final prediction stage [94].

Libraries used in study

Table 4 Most prominent libraries used in the study

Pandas	Beneficial for dataframe handling
Numpy	Useful for computation and arrays
Keras	For deep learning
Random Forest	For Random Forest Regressor

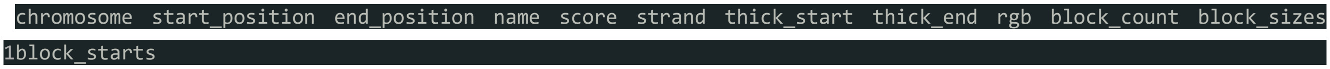
Brief explanation of .bed files used in study:

Bed (browser extensible data) files are usually used in gene annotation for their compatibility in representing genetic information. For example, the bed files used in this study contains output files from WGBS processing in which the inputs are fastq files (containing raw reads from DNA sequencing) and tar files (contain genome index).

Bed files here represent the methylation state (percentage) at CpG sites. The bed file stores the information as coordinates and annotations. This format was first created for the Human Genome Project in 2003, and was later adopted by the Encode Project (The Encyclopedia of DNA Elements) [95].

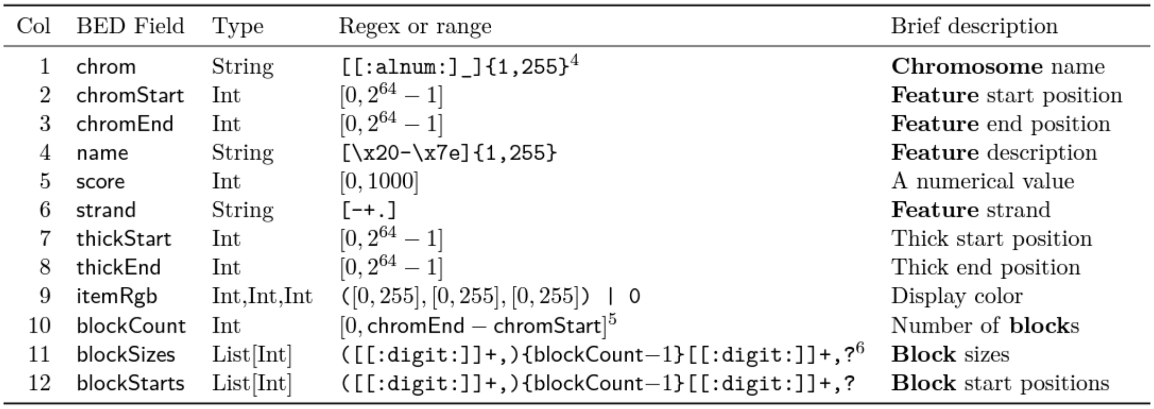
Started in 2003, ENCODE's goal is to comprehensively collect data for all functional parts of the human genome mainly by discovery and annotation of gene elements, making use of technologies such as assays of DNA methylation, DNA hypersensitivity, and immunoprecipitation of proteins interacting with both DNA and RNA. Its most beneficial aspect is its freely-accessible database. It first focused on 1% of the whole genome and has now conducted whole-genome analyses of human and mouse genomes [95].

Obligatory columns of bed files are the chromosome names, start, and end sequences of the CpGs concerned, and then optional or additional columns might contain read number, percentage of methylation-showing reads, color value (RGB) and more



```
 chromosome  start_position  end_position  name  score  strand  thick_start  thick_end  rgb  block_count  block_sizes
1block_starts
```

Figure 22 Column labels of most .bed files [96]

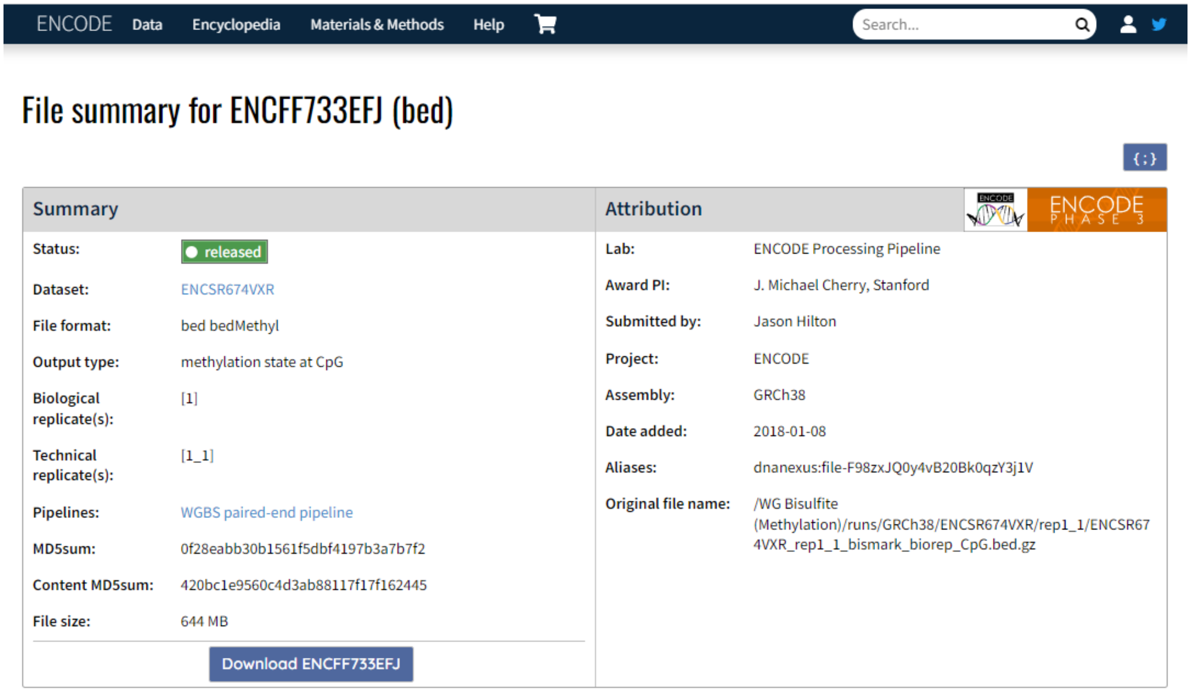


Col	BED Field	Type	Regex or range	Brief description
1	chrom	String	`[[:alnum:]_]{1,255}`[4]	**Chromosome** name
2	chromStart	Int	$[0, 2^{64} - 1]$	**Feature** start position
3	chromEnd	Int	$[0, 2^{64} - 1]$	**Feature** end position
4	name	String	`[\x20-\x7e]{1,255}`	**Feature** description
5	score	Int	$[0, 1000]$	A numerical value
6	strand	String	`[-+.]`	**Feature** strand
7	thickStart	Int	$[0, 2^{64} - 1]$	Thick start position
8	thickEnd	Int	$[0, 2^{64} - 1]$	Thick end position
9	itemRgb	Int,Int,Int	([0, 255], [0, 255], [0, 255]) | 0	Display color
10	blockCount	Int	[0, chromEnd − chromStart][5]	Number of **blocks**
11	blockSizes	List[Int]	`([[:digit:]]+,){blockCount−1}[[:digit:]]+,?`[6]	**Block** sizes
12	blockStarts	List[Int]	`([[:digit:]]+,){blockCount−1}[[:digit:]]+,?`	**Block** start positions

Figure 23 The official documentation of bed files, indicating the first 3 obligatory columns and additional files, along with purpose of each column [97]

WGBS bed files used for this paper describe the methylation state of CpGs, and follow a WGBS paired-end pipeline (meaning it takes in paired-end fastqs as inputs).

Another sometimes simpler option, which is the option used in this study is the use of Python libraries like pybedtools and bed-reader. Both work well with numpy and pandas for proper reading.



File summary for ENCFF733EFJ (bed)

Summary	
Status:	● released
Dataset:	ENCSR674VXR
File format:	bed bedMethyl
Output type:	methylation state at CpG
Biological replicate(s):	[1]
Technical replicate(s):	[1_1]
Pipelines:	WGBS paired-end pipeline
MD5sum:	0f28eabb30b1561f5dbf4197b3a7b7f2
Content MD5sum:	420bc1e9560c4d3ab88117f17f162445
File size:	644 MB

Download ENCFF733EFJ

Attribution	
Lab:	ENCODE Processing Pipeline
Award PI:	J. Michael Cherry, Stanford
Submitted by:	Jason Hilton
Project:	ENCODE
Assembly:	GRCh38
Date added:	2018-01-08
Aliases:	dnanexus:file-F98zxJQ0y4vB20Bk0qzY3j1V
Original file name:	/WG Bisulfite (Methylation)/runs/GRCh38/ENCSR674VXR/rep1_1/ENCSR674VXR_rep1_1_bismark_biorep_CpG.bed.gz

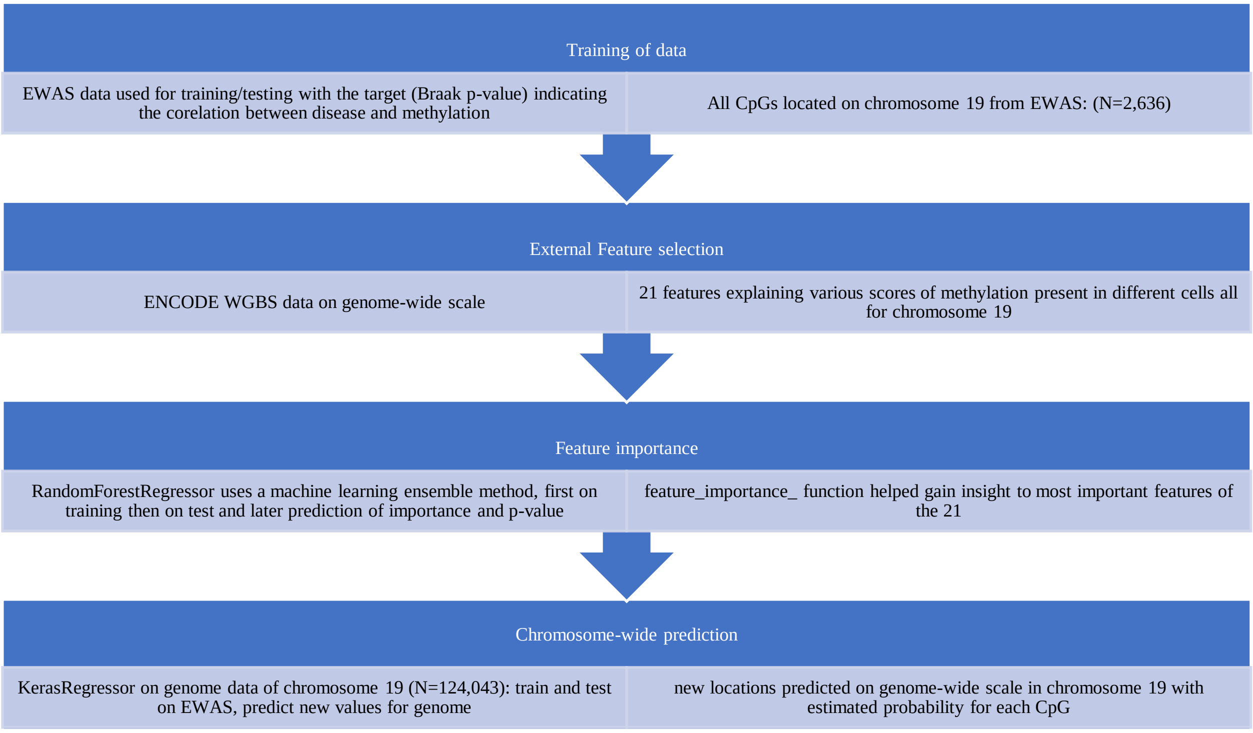
Figure 22 Example of WGBS .bed file page on ENCODE website.

`bed-reader` works with numpy and pandas (in pandas, it reads the bed file as a csv) and was used as the bed library in this study and can be used like the following example [98].

```
 pip install bed-reader
import numpy as np
from bed_reader import open_bed, sample_file
file_name = sample_file("small.bed")
bed = open_bed(file_name)
val = bed.read()
print(val)
[[ 1.  0. nan  0.]
 [ 2.  0. nan  2.]
 [ 0.  1.  2.  0.]]
```

Work conducted in this thesis

This study/project was conducted in the aim of shedding light on epigenetics in Alzheimer’s disease, and the role of machine learning and deep learning in discovering new epigenetic marks (DNA methylation specifically). An attempt was made for employing a deep learning model to predict new locations in the epigenetic field related to Alzheimer’s disease.



Training of data

EWAS data used for training/testing with the target (Braak p-value) indicating the corelation between disease and methylation

All CpGs located on chromosome 19 from EWAS: (N=2,636)

External Feature selection

ENCODE WGBS data on genome-wide scale

21 features explaining various scores of methylation present in different cells all for chromosome 19

Feature importance

RandomForestRegressor uses a machine learning ensemble method, first on training then on test and later prediction of importance and p-value

feature_importance_ function helped gain insight to most important features of the 21

Chromosome-wide prediction

KerasRegressor on genome data of chromosome 19 (N=124,043): train and test on EWAS, predict new values for genome

new locations predicted on genome-wide scale in chromosome 19 with estimated probability for each CpG

Figure 23Workflow of the processes taken in this study 1) training of model on EWAS data of chromosme 19 (using both RandomForestRegressor and KerasRegressor) 2) Feature selection from external sources (WGBS) from which 21 features describing the methylation status of the Cpgs 3) applying the machine learning model on our data for feature importance ranking 4) Prediction of methylation of CpGs using a machine learning model and a deep learning model.

Pipeline: Employing AI/ML/DL for better insight into DNA methylation of AD

In this thesis, both a machine learning model and deep learning were trained to find CpGs related to Alzheimer's disease from EWAS data that contains more than 2,636 CpG sites along with their Braak stage p-values.

Data Collection and Exploration

For collection of data, several data files were needed for the proper training and prediction of the model.

It is worth noting that this step of the study was the most time-consuming due to regional difficulties and computational setbacks in terms of internet speed and processing quality. Files used in the pipeline of this study are 21 WGBS bed files (feature files) from the ENCODE project, containing description of methylation states at CpG sites, along with an EWAS file from an Arizona cohort containing β-values indicating methylation values of 450 thousand CpG sites from (N=302) participants and it is the main cohort used in the study. Our Arizona EWAS data contains 411,714 CpGs with their estimates, standard errors, t value, Braak p-value and mean β-values.

After obtaining the data files (1 EWAS file and 22 WGBS files), a step was taken to make use of uniform numbers/indexing of CpG sites, since WGBS files have coordinates of CpGs, while EWAS files has CpG labels without coordinates.

The step generated an EWAS file containing start sites (coordinates) of the CpG site along with its Braak p-values, and 21 WGBS scores

Features in this study (the 21 WGBS features) are:

Table 5 WGBS feature files used in study

Accession no.	ENCFF157POM	ENCFF003JVR	ENCFF428TVT	ENCFF588ETU	ENCFF110AZO

Cell source	sigmoid colon	A549	mesenchymal stem cell	muscle of leg	right cardiac atrium

Accession no.	ENCFF763RUE	ENCFF489CEV	ENCFF774GXJ	ENCFF699RBP	ENCFF366UWF
Cell source	pancreas	stomach	skeletal muscle myoblast	body of pancreas	hepatocyte

Accession no.	ENCFF116DGM	ENCFF103DNU	ENCFF092FNE	ENCFF064GJQ	ENCFF043NUK
Cell source	GM23248	adipose tissue	H1	HepG2	endodermal cell

Accession no.	ENCFF536RSX	ENCFF801OHX	ENCFF601NBW	ENCFF435ETE	ENCFF451WIY
Cell source	heart left ventricle	mesenchymal stem cell	H1	natural killer cell	CD14-positive monocyte

Accession no.	ENCFF355UVU
Cell source	T-cell

Data Cleaning

The main step taken for data cleaning was the focus on chromosome 19 in particular, since it is the chromosome studied in this thesis, and the reason for choosing a single chromosome was the inability to apply the computations to the entire genome. Therefore, using Python, all other chromosomes in WGBS and EWAS files were omitted, leaving only details of chromosome 19, such as coordinates, start and end, and all other 9 columns.

Next, 21 features (the 21 WGBS.bed files) are assigned to CpG sites. This measure is taken on the EWAS file, to give columns of the CpG site, Braak p-value, and score of all 21 features.

The same process is performed on the genome file to give an outcome of 124,043 sites, with their CpG sites and scores.

Feature importance

As proven previously in the reference study, this type of data requires a feature ranking/importance step to identify the most important features. `RandomForestRegressor` was chosen here. It depends on a number of decision trees and was called using `sklearn.ensemble`:

```
from sklearn.ensemble import RandomForestRegressor
```

Then the feature importances are taken after fitting the model (Model details in next step)

```
regressor.feature_importances_
```

The feature importances were of low quality due to the relatively small number of features, with the highest feature importance being for mesenchymal stem cell score: 0.06255308

Modeling

Since the reference study only used machine learning models (ensemble learning model of xgboost, RLR, SVC, . . . etc.), an attempt was made to employ a deep learning model to predict CpGs just as the reference study accomplished, but with admittedly less data due to connection and computational limitations in most third-world countries.

This study employed two artificial intelligence models: `RandomForestRegressor` and `KerasRegressor`.

The machine learning model, `RandomForestRegressor` was deployed first after importing `pandas`, `numpy`, and all necessary for the model from `sklearn.model_selection`, `sklearn.preprocessing`, `sklearn.ensemble`

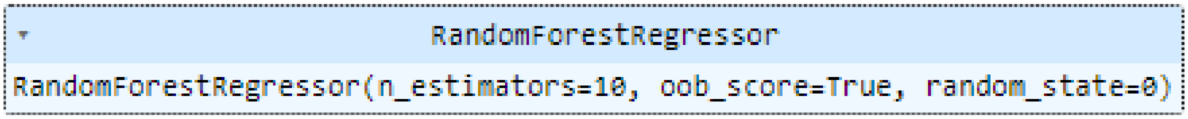
Next was the splitting of the EWAS dataset into train set and test set.

```
dataset = dataframe.values
X = dataset[:,1:22]
Y = dataset[:,22]
seed = 7
test_size = 0.2
X_train, X_test, y_train, y_test = train_test_split(X, Y,
test_size=test_size, random_state=seed)
```

Next step, fitting of the model to the dataset:

```
regressor = RandomForestRegressor(n_estimators=10,
random_state=0, oob_score=True)
regressor.fit(X_train, y_train)
```

This step will give the following output:



```
RandomForestRegressor(n_estimators=10, oob_score=True, random_state=0)
```

Our regressor score is

```
regressor.score(X_test,y_test)
```

-0.1114018192069881

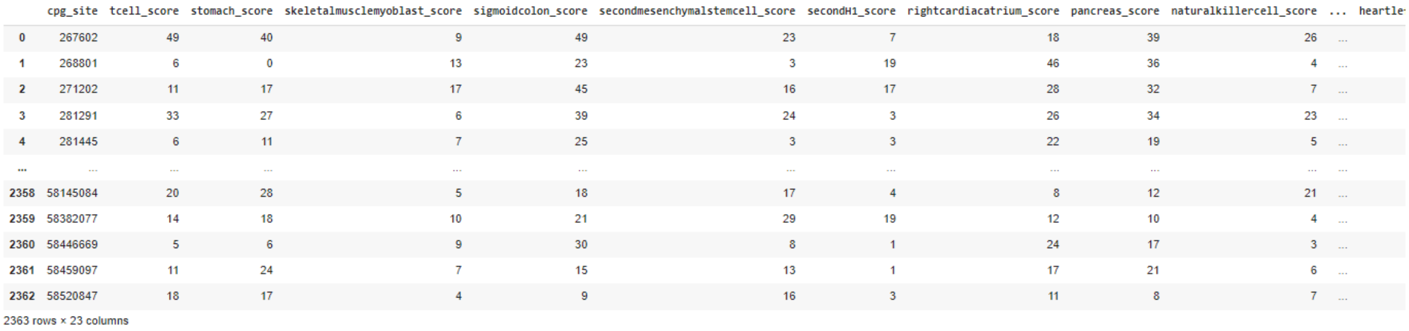
Next, we start prediction of p-value on test samples from genome data:

```
genome_data['predicted_pvalue']=regressor.predict(test_samples)
```

This gives an output of predicted Braak p-values of 123042 CpG sites.

Our next model deployed here was the deep learning model `KerasRegressor` after importing: `pandas`, `numpy`, and necessary `Keras` components from `scikeras.wrappers`, `sklearn.model_selection`, `tensorflow.keras.layers`, and `tensorflow.keras.models`.

we start loading the dataset



	cpg_site	tcell_score	stomach_score	skeletalmusclemyoblast_score	sigmoidcolon_score	secondmesenchymalstemcell_score	secondH1_score	rightcardiacatrium_score	pancreas_score	naturalkillercell_score	...	heartle
0	267602	49	40	9	49	23	7	18	39	26	...	
1	268801	6	0	13	23	3	19	46	36	4	...	
2	271202	11	17	17	45	16	17	28	32	7	...	
3	281291	33	27	6	39	24	3	26	34	23	...	
4	281445	6	11	7	25	3	3	22	19	5	...	
...	...	...	...	...	...	...	...	...	...	...	...	
2358	58145084	20	28	5	18	17	4	8	12	21	...	
2359	58382077	14	18	10	21	29	19	12	10	4	...	
2360	58446669	5	6	9	30	8	1	24	17	3	...	
2361	58459097	11	24	7	15	13	1	17	21	6	...	
2362	58520847	18	17	4	9	16	3	11	8	7	...	

2363 rows × 23 columns

And splitting the dataset

```
dataset = dataframe.values
X = dataset[:,1:22]
Y = dataset[:,22]
seed = 7
test_size = 0.2
X_train, X_test, y_train, y_test = train_test_split(X, Y,
test_size=test_size, random_state=seed)
```

Next is the definition of the base model, creation, and compilation:

```
def baseline_model():
  model = Sequential()
  model.add(Dense(21, input_shape=(21,),
kernel_initializer='normal', activation='relu'))
```

```
  model.add(Dense(1, kernel_initializer='normal'))
  model.compile(loss='mean_squared_error', optimizer='adam')
  return model
```

After evaluation (more detail in next section), the estimator is fit:

```
estimator.fit(X_train,y_train)
```

Output:

☐ KerasRegressor

```
KerasRegressor(
      model=<function baseline_model at 0x0000020FF45184C0>
      build_fn=None
      warm_start=False
      random_state=None
      optimizer=rmsprop
      loss=None
      metrics=None
      batch_size=5
      validation_batch_size=None
      verbose=0
      callbacks=None
      validation_split=0.0
      shuffle=True
      run_eagerly=False
      epochs=100
)
```

Next, the model is applied on the genome dataset to predict new values

```
genome_data['predicted_values']=estimator.predict(test_samples)
```

This gave us p-values predicted by `KerasRegressor` for all CpG sites on

chromosome 19

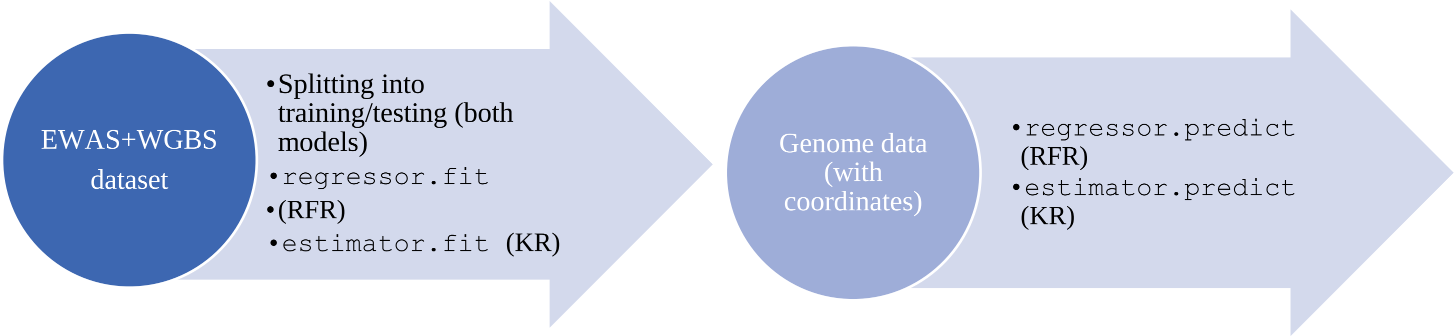


Figure 24 Simple explanation of input of each code in both RandomForestRegressor and KerasRegressor

Evaluation

`RandomForestRegressor` relies on `regressor.score` and its Root Mean Square Error (RMSE) for the evaluation of its performance, along with the p-value threshold of its predictions.

Its regressor score is -0.1114018192069881
RMSE is calculated:

```
rmse=np.sqrt(np.sum(error*error)/error.size)
```

0.3146324186691835

`KerasRegressor` relies on its MSE, RMSE, and an estimator score for its evaluation:

```
estimator = KerasRegressor(model=baseline_model, epochs=100,
batch_size=5, verbose=0)
kfold = KFold(n_splits=10)
results = cross_val_score(estimator, X_test, y_test, cv=kfold,
scoring='neg_mean_squared_error')
print("Baseline: %.2f (%.2f) MSE" % (results.mean(),
results.std()))
     Baseline: -0.12 (0.02) MSE
```

```
rmse=np.sqrt(np.sum(error*error)/error.size)
rmse
```

```
0.3039870671649546
```

```
estimator.score(X_test,y_test)
```

```
-0.03746721240964379
```

So far, no single metric has been unified and deployed for evaluation of regression analysis, since its evaluation is not as simple as that for classification models. Some studies use mean square error MSE, or its rooted variant RMSE. In general, a regressor’s score is the R^2 coefficient of determination. It ranges between 0.0 and 1.0 and can be either negative or positive (best possible score is 1.0) [99].

As for MSE and RMSE, they are the mean squared error and root of mean squared error: MSE identifies regression loss, and the closer to zero, the better the performance.

Chapter 4: Obtained Results and Analysis/Discussion

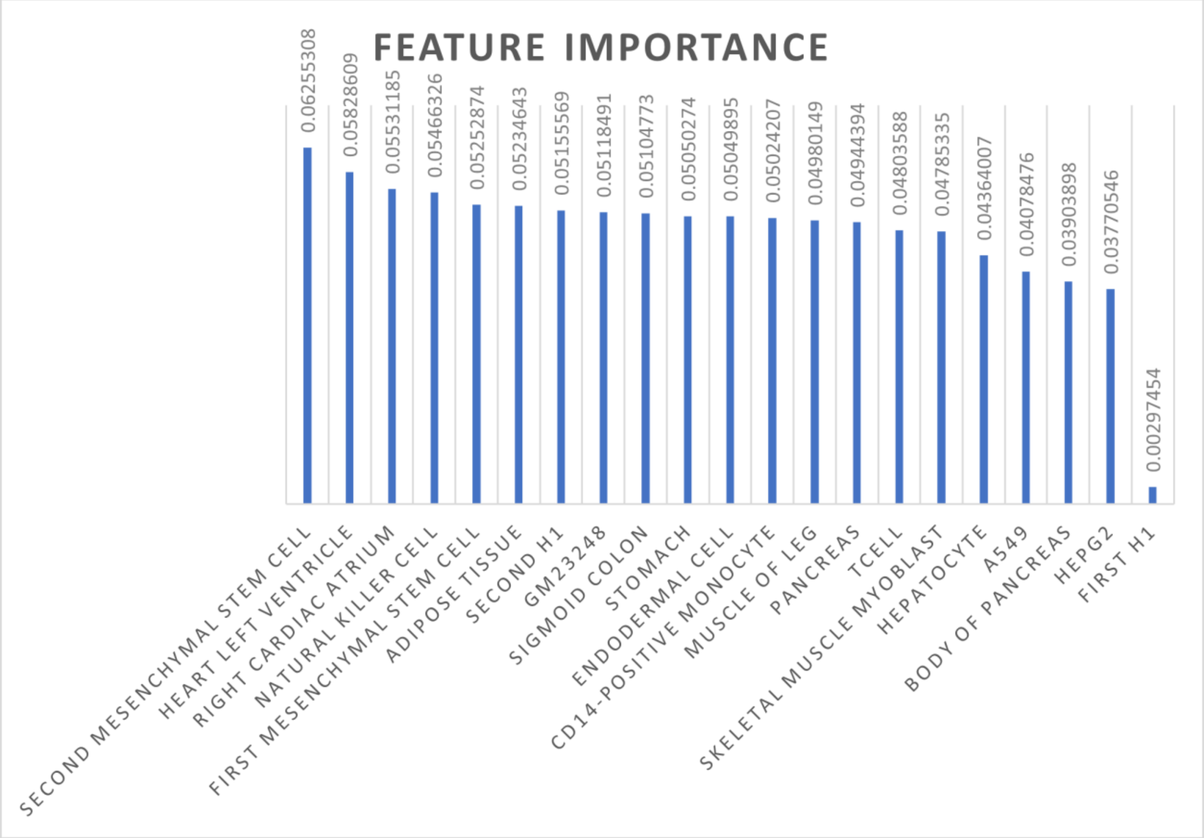
Results

The result of using the two models was two different predictions on the genome dataset. One using `RandomForestRegressor`: a machine learning model, while the other was `KerasRegressor`: a deep learning model.

Feature Importance

The application of `RandomForestRegressor` first achieved insight into the 21 features used in the models, and ranked the importances of the features as follows:

Table 6 Bar chart describing feature importance of all 21 features



The highest ranked feature in terms of importance was the second mesenchymal stem cell and its WGBS methylation score (ENCODE accession number ENCFF801OHX) with a feature importance score of 0.06255308, followed by the heart left ventricle score (ENCFF536RSX) with a score of 0.05828609 and then right cardiac atrium score (ENCFF110AZO): 0.05531185.

The next results obtained were those of the `RandomForestRegressor` performance evaluation and predictions:

The `RandomForestRegressor` performance was evaluated as `regressor.score` of -0.1114018192069881

Followed by evaluation using rmse = 0.3146324186691835

Prediction

The predictions of the `RandomForestRegressor` were predictions of the Braak p-value of all 124,043 sites on chromosome 19 in the aim of predicting methylated loci that were previously undetected.

The lowest p-values predicted (must be less than p-value threshold of 0.04) were of:

Table 7 Best predicted CpG sites believed to play role in methylation

CpG site	Predicted p-value
474445	0.005540819
18768688	0.009248729
1580576	0.033382792
939810	0.039087401

The score of the regressor is -0.03746721240964379, and its RMSE is 0.3039870671649546

KerasRegressor also predicted similar results to RandomForestRegressor, indicating that the full potential of deep learning cannot be unlocked at this relatively low data size.

Discussion

Alzheimer's disease is a very complex and mysterious disease – it is a multifactorial disease caused by genetic factors (genes such as APOE, ANK1, . . . etc.) and environmental factors that could potentially move the patients towards or away from the disease. These environmental factors such as diet and stress affect the genes through epigenetics and the many epigenetic mechanisms of DNA methylation, histone modifications, and noncoding RNAs. DNA methylation has been studied extensively and it is implied that it may play a major role in the understanding of AD.

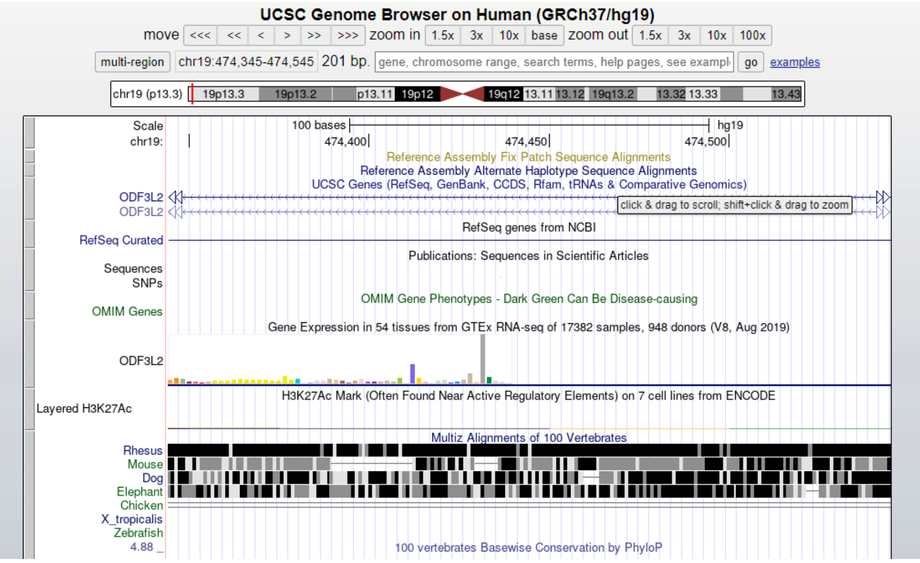
Current studies of epigenetics such as epigenome-wide association studies (EWAS) only detect a small percentage of methylation in the genome (approximately 450,000 sites from the entire genome), and the need for better detection is necessary. This thesis took inspiration from the EWASplus study and referred to it as a reference study in the aim of using a small fraction of its methods (due to technical and regional limitations) in the attempt of replicating the methodology for applying a deep learning model for detection of new methylated CpG sites in the genome.

In this thesis, deep learning was applied to attempt prediction on CpGs on chromosome 19. It has recently been suggested that deep learning can be the key for clearer understanding of epigenetics and its many mechanisms to further pave the way towards epigenetics-targeted drugs. An EWAS dataset was utilized as a training/testing set for the artificial intelligence models utilized, and a dataset containing all CpG sites of the genome and their coordinates was used for the prediction stage. Both datasets were downscaled to the range of chromosome 19 only, due to computational and connection limitations present in most third-world countries, along with the fact that chromosome 19 contains the APOE gene, a major risk gene in Alzheimer’s disease, and epigenetic marks usually occur near key genes of the disease. We joined the EWAS dataset with 21 feature datasets from the ENCODE project of WGBS from different cells indicating scores of methylation at CpG sites.

The models of this thesis were `RandomForestRegressor` and `KerasRegressor`. Regressors are used here as opposed to classifiers, due to the need for predicting continuous rather than binary data. `RandomForestRegressor`, a machine learning model, was chosen due to its flexibility with large datasets and its benefit of utilizing the `feature_importance_` function, which is very similar to a step taken in the reference study which used a more complex feature selection stage due to its high number of features (2256) as opposed to our 21 features (the features used in this study were previously used in the reference study and were chosen due to their high importance ranking). `RandomForestRegressor` was later used for prediction of new CpG sites on chromosome 19. Then came the `KerasRegressor`, which is a deep learning model. No deep learning was used in the reference study, and our main concept was applying it to smaller data and samples to evaluate its efficacy on a smaller scale.

`RandomForestRegressor` ranked the importance of features and gave a result that the top features were of the mesenchymal stem cell, the heart left ventricle, and the right cardiac atrium. Even though the scores (in the results section) were not very high, and could either indicate authentic results of a role in the detection, or not, one can hypothesize that stem cells present potential in disease and are more prone to environmental factors. As for the heart left ventricle and right cardiac atrium, since they are both cells of the heart, this can be backed up by Figure 1 in this document, which lists strokes, cardiovascular disease, and congestive heart failure as risk factors of Alzheimer's disease.

Performance of `RandomForestRegressor` was evaluated using `regressor.score` and `RMSE`. `KerasRegressor` was evaluated using `estimator.score` and `RMSE`. Both performances were not high due to the low number of features/columns and samples/rows used in the training/testing EWAS dataset. The case at hand (the problem of both models) was not a malfunction or problem caused by the actual models, nor by the data present in the datasets, but by the quantity, causing what is known as '*overfitting*'. Overfitting is when the model does not behave in a desirable manner in which it gives accurate predictions for its training data (it learned the training data too well). As mentioned earlier, it is computationally impossible in this region and this short time to benefit of publicly available data to achieve higher and more complex genomic results and conclusions, which is why only 21 features were chosen and only samples of chromosome 19 were used. However, `RandomForestRegressor` and `KerasRegressor` could be used in the future with similar data sizes to the reference study with the utilization of super computers, of course, along with using higher feature sizes (more than 2000) and using the original window IDs of CpGs of the genome (which was also technically impossible in this thesis). Hypothesizing that our predictions could reach some insight into DNA methylation of the genome in Alzheimer's disease, we could reach the results that loci 474445, 18768688, and 1580576 show high correlation with methylation in Alzheimer's disease, but are not very close to APOE on chromosome 19.

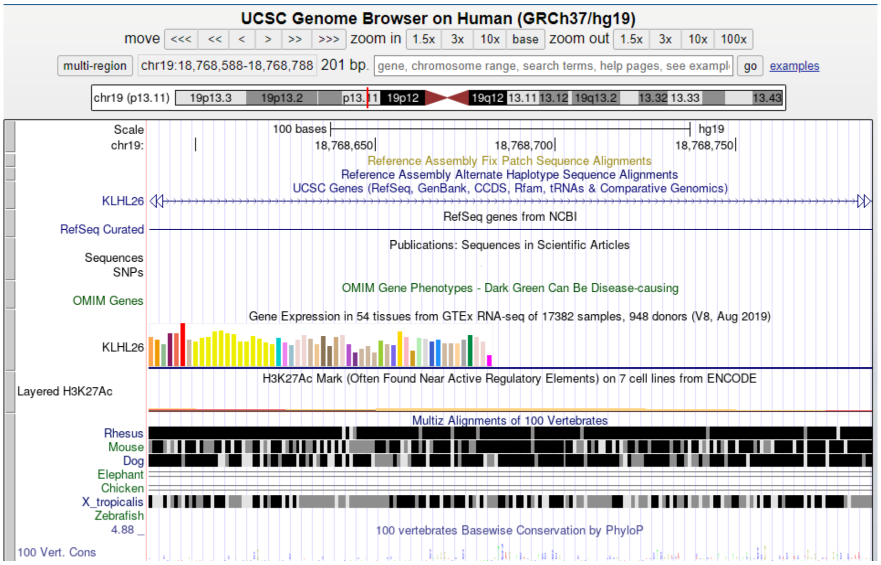


UCSC Genome Browser on Human (GRCh37/hg19)

chr19:474,345-474,545 201 bp.

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
multi-region
gene, chromosome range, search terms, help pages, see exampl
go
examples
chr19 (p13.3)
19p13.3 19p13.2 p13.11 19p12 19q12 13.11 13.12 19q13.2 13.32 13.33 13.43
Scale
chr19:
100 bases
hg19
474,400
474,450
474,500
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
ODF3L2
ODF3L2
click & drag to scroll; shift+click & drag to zoom
RefSeq genes from NCBI
RefSeq Curated
Publications: Sequences in Scientific Articles
Sequences
SNPs
OMIM Gene Phenotypes - Dark Green Can Be Disease-causing
OMIM Genes
Gene Expression in 54 tissues from GTEx RNA-seq of 17382 samples, 948 donors (V8, Aug 2019)
ODF3L2
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
4.88 _
100 vertebrates Basewise Conservation by PhyloP

Figure 25 Location of CpG coordinate 474445 on chromosome 19 via UCSC Genome Browser



UCSC Genome Browser on Human (GRCh37/hg19)

chr19:18,768,588-18,768,788 201 bp.

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
multi-region
gene, chromosome range, search terms, help pages, see exampl
go
examples
chr19 (p13.11)
19p13.3 19p13.2 p13.11 19p12 19q12 13.11 13.12 19q13.2 13.32 13.33 13.43
Scale
chr19:
100 bases
hg19
18,768,650
18,768,700
18,768,750
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
KLHL26
RefSeq genes from NCBI
RefSeq Curated
Publications: Sequences in Scientific Articles
Sequences
SNPs
OMIM Gene Phenotypes - Dark Green Can Be Disease-causing
OMIM Genes
Gene Expression in 54 tissues from GTEx RNA-seq of 17382 samples, 948 donors (V8, Aug 2019)
KLHL26
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
4.88 _
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons

Figure 26 Location of CpG coordinate 18768688 on chromosome 19 via UCSC Genome Browser

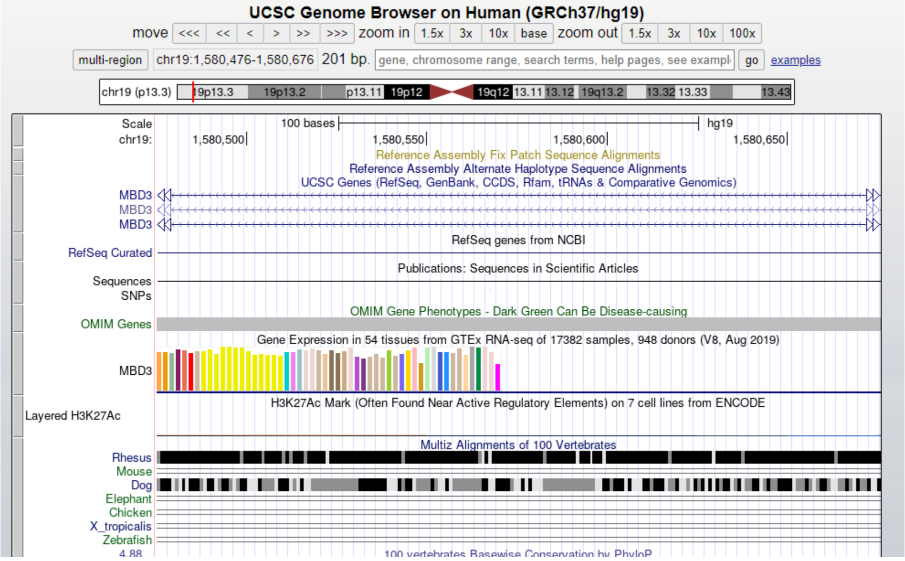

UCSC Genome Browser on Human (GRCh37/hg19)
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
multi-region chr19:1,580,476-1,580,676 201 bp. gene, chromosome range, search terms, help pages, see exampl go examples
chr19 (p13.3) 19p13.3 19p13.2 p13.11 19p12 19q12 13.11 13.12 19q13.2 13.32 13.33 13.43
Scale 100 bases hg19
chr19: 1,580,500 1,580,550 1,580,600 1,580,650
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
MBD3
MBD3
MBD3
RefSeq genes from NCBI
RefSeq Curated
Publications: Sequences in Scientific Articles
Sequences
SNPs
OMIM Gene Phenotypes - Dark Green Can Be Disease-causing
OMIM Genes
Gene Expression in 54 tissues from GTEx RNA-seq of 17382 samples, 948 donors (V8, Aug 2019)
MBD3
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
4.88
100 vertebrates Basewise Conservation by PhyloP

Figure 27 Location of CpG coordinate 1580576 on chromosome 19 via UCSC Genome Browser

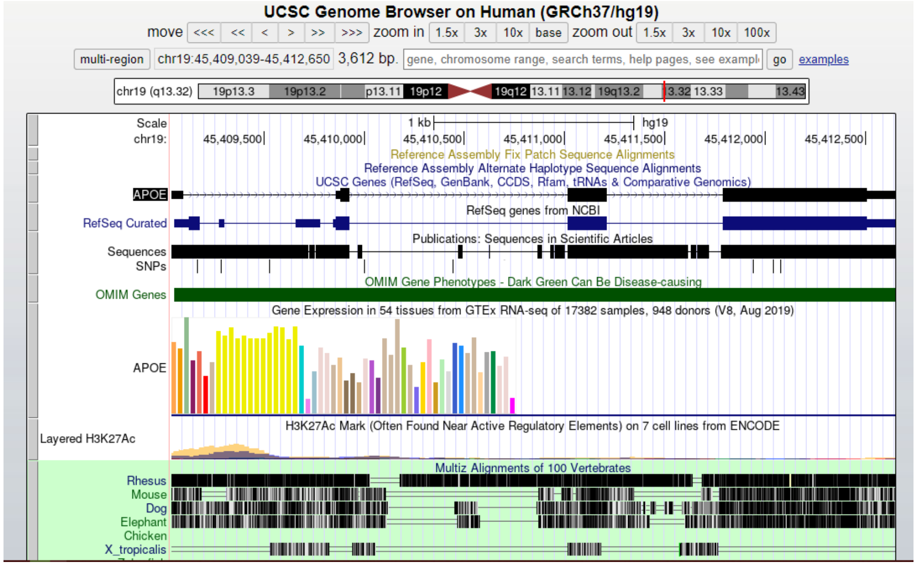
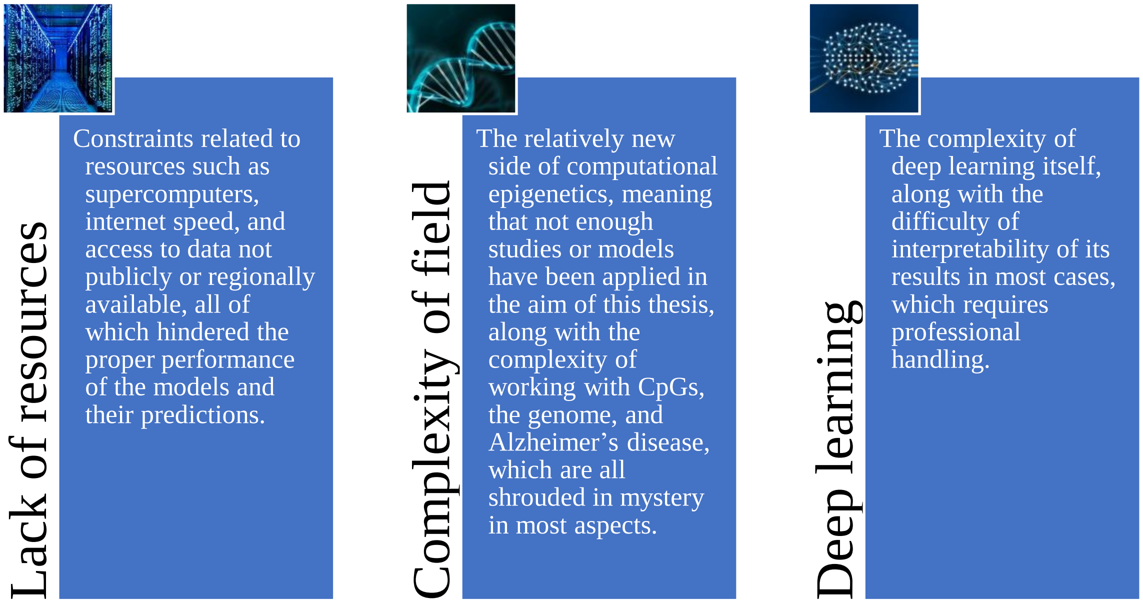

UCSC Genome Browser on Human (GRCh37/hg19)
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
multi-region chr19:45,409,039-45,412,650 3,612 bp. gene, chromosome range, search terms, help pages, see exampl go examples
chr19 (q13.32) 19p13.3 19p13.2 p13.11 19p12 19q12 13.11 13.12 19q13.2 13.32 13.33 13.43
Scale 1 kb hg19
chr19: 45,409,500 45,410,000 45,410,500 45,411,000 45,411,500 45,412,000 45,412,500
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
APOE
RefSeq genes from NCBI
RefSeq Curated
Publications: Sequences in Scientific Articles
Sequences
SNPs
OMIM Gene Phenotypes - Dark Green Can Be Disease-causing
OMIM Genes
Gene Expression in 54 tissues from GTEx RNA-seq of 17382 samples, 948 donors (V8, Aug 2019)
APOE
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis

Figure 28 Location of APOE gene on chromosome 19

This indicates that these loci, with further proof using higher computational capabilities (super computers) and clinical experiments, could be beneficial in identifiying methylation in this area on chromosome 19, along with predicting more accurate methylated CpG loci in the vicinity of the APOE gene. As for `KerasRegressor`, which was the aim of this thesis: deep learning in clinical epigenetics, it also achieved a low estimator score for the same reasons mentioned above: low computational and hardware capabilities. It still shows promise because deep learning has been recently believed to be the key for the identification of epigenetic mechanisms of the genome, and precisely in DNA methylation.

It is worth noting, however, that many limitations faced this study (*Figure 31*):



Lack of resources

Constraints related to resources such as supercomputers, internet speed, and access to data not publicly or regionally available, all of which hindered the proper performance of the models and their predictions.

Complexity of field

The relatively new side of computational epigenetics, meaning that not enough studies or models have been applied in the aim of this thesis, along with the complexity of working with CpGs, the genome, and Alzheimer's disease, which are all shrouded in mystery in most aspects.

Deep learning

The complexity of deep learning itself, along with the difficulty of interpretability of its results in most cases, which requires professional handling.

Figure 29 Constraints and limitations of study

Since the regressor scores in both models (machine learning and deep learning) were not suitable, one can either hypothesize that the results may be of some truth, thereby extracting some information from CpG sites on chromosome 19 that show higher methylation levels than other loci on the genome. This is of great benefit for the newly emerging field of computational epigenetics from the therapeutic approach, which is recently targeting the APOE gene: APOE-targeted epigenome therapy [100]. The utilization of both models can show promise and be of great benefit in the field when used with larger data and better computational resources, time, and funding. The work and model completed here can also be tried with different epigenetic mechanisms, such as histone modifications (acetylation, methylation, ubiquitinylation, and more), since the concept is quite similar in which the model targets and computationally quantifies the methylation (or any other chemical group) and then predicts such levels in novel loci. The concept can also work similarly with single nucleotide polymorphisms, since the models at hand deal with regression rather than classification. Such alterations can be useful in extending the benefit achieved from the models in this study and enable a deeper level of understanding of the effects the epigenome (DNA methylation, ncRNA, and histone modifications) and SNPs have on gene expression in AD. Future work can aim at using robust cloud computing resources, supercomputers, and better funding and extensive periods of time to achieve better results with deep learning models such as `KerasRegressor`, along with joining SNPs to the studied data and more epigenetic marks to ensure better studies of the epigenome.

Conclusion

In conclusion, this thesis's main aim was to utilize deep learning in clinical epigenetics and discuss therapeutic approaches. Both a machine learning model and a deep learning model were used in this thesis to identify methylated CpG loci on chromosome 19 in Alzheimer's disease. Due to computational limitations, short time, and low internet speed, only a portion of the data use in this thesis's reference study were used, which resulted in low scores and inaccurate predictions. Even though the resulting data may be inaccurate, this still offers potential for these models to be used with larger data and better resources in the aim of extracting beneficial information regarding DNA methylation of CpG sites of the genome, paving the way for higher accuracy in terms of DNA-methylation-targeted therapy. Therefore, with higher computational power, better funding, and time, serious steps can be taken towards epigenetics-targeted drugs to help reverse the reversable sides of epigenetics in early stages of the disease before it is too late.

References

[1] Alzheimer, A., Stelzmann, R. A., Schnitzlein, H. N., & Murtagh, F. R. (1995). **An English translation of Alzheimer's 1907 paper, "Uber eine eigenartige Erkankung der Hirnrinde"**. Clinical anatomy (New York, N.Y.), 8(6), 429–431. https://doi.org/10.1002/ca.980080612

[2] Dong, H. K., Gim, J. A., Yeo, S. H., & Kim, H. S. (2017). **Integrated late onset Alzheimer's disease (LOAD) susceptibility genes: Cholesterol metabolism and trafficking perspectives**. Gene, 597, 10–16. https://doi.org/10.1016/j.gene.2016.10.022

[3] Mendez M. F. (2017). **Early-Onset Alzheimer Disease**. Neurologic clinics, 35(2), 263–281. https://doi.org/10.1016/j.ncl.2017.01.005

[4] World Health Organization. (2023, December 23). **Dementia**. https://www.who.int/news-room/fact-sheets/detail/dementia

[5] Li, X., Feng, X., Sun, X., Hou, N., Han, F., & Liu, Y. (2022). **Global, regional, and national burden of Alzheimer's disease and other dementias, 1990-2019**. Frontiers in aging neuroscience, 14, 937486. https://doi.org/10.3389/fnagi.2022.937486

[6] A Armstrong R. (2019). **Risk factors for Alzheimer's disease**. Folia neuropathologica, 57(2), 87–105. https://doi.org/10.5114/fn.2019.85929

[7] Afzal, S., Maqsood, M., Khan, U., Mehmood, I., Nawaz, H., Aadil, F., Song, O.-Y., & Nam, Y. (2021). **Alzheimer Disease Detection Techniques and Methods: A Review.** International Journal of Interactive Multimedia and Artificial Intelligence, 6(7), 26. https://doi.org/10.9781/ijimai.2021.04.005

[8] Hampel, H., Hardy, J., Blennow, K., Chen, C., Perry, G., Kim, S. H., Villemagne, V. L., Aisen, P., Vendruscolo, M., Iwatsubo, T., Masters, C. L., Cho, M., Lannfelt, L., Cummings, J. L., & Vergallo, A. (2021). **The Amyloid-β Pathway in Alzheimer's Disease**. Molecular Psychiatry, 26(10). https://doi.org/10.1038/s41380-021-01249-0

[9] Medeiros, R., Baglietto-Vargas, D., & LaFerla, F. M. (2011). **The role of tau in Alzheimer's disease and related disorders**. CNS neuroscience & therapeutics, 17(5), 514–524. https://doi.org/10.1111/j.1755-5949.2010.00177.x

[10] Bright Focus. (2023 December 23). **Amyloid Plaques and Neurofibrillary Tangles**. https://www.brightfocus.org/news/amyloid-plaques-and-neurofibrillary-tangles

[11] Weinhold B. (2006). **Epigenetics: the science of change**. Environmental health perspectives, 114(3), A160–A167. https://doi.org/10.1289/ehp.114-a160

[12] Kato, K. (2022, January 1). **Chapter 5 - What is epigenetics?** (H. Tomita, Ed.). ScienceDirect; Academic Press. https://www.sciencedirect.com/science/article/abs/pii/B9780323885263000051

[13] Jin, B., Li, Y., & Robertson, K. D. (2011). **DNA methylation: superior or subordinate in the epigenetic hierarchy?**. Genes & cancer, 2(6), 607–617. https://doi.org/10.1177/1947601910393957

[14] Pi, T., Liu, B., & Shi, J. (2020, September 8). **Abnormal Homocysteine Metabolism: An Insight of Alzheimer’s Disease from DNA Methylation**. Behavioural Neurology. https://www.hindawi.com/journals/bn/2020/8438602/

[15] Campagna, M. P., Xavier, A., Lechner-Scott, J., Maltby, V., Scott, R. J., Butzkueven, H., Jokubaitis, V. G., & Lea, R. A. (2021). **Epigenome-wide association studies: current knowledge, strategies and recommendations**. Clinical Epigenetics, 13(1). https://doi.org/10.1186/s13148-021-01200-8

[16] Cummings, J. L., Leisgang, A. M., Cammann, D., Powell, J., & Chen, J. (2023). **Anti-Amyloid Monoclonal Antibodies for the Treatment of Alzheimer’s Disease**. BioDrugs. https://doi.org/10.1007/s40259-023-00633-2

[17] Alzheimer’s Organization. (2023 December 23). **Drug Treatments for Alzheimer’s Disease**. https://www.alzheimers.org.uk/sites/default/files/pdf/factsheet_drug_treatments_for_alzheimers_disease.pdf

[18] Weill Institute for Neurosciences. (2023 December 23) **Familial Alzheimer’s Disease**. https://memory.ucsf.edu/genetics/familial-alzheimer-disease

[19] Iqbal, K., & Grundke-Iqbal, I. (2010). **Alzheimer's disease, a multifactorial disorder seeking multitherapies**. Alzheimer's & dementia : the journal of the Alzheimer's Association, 6(5), 420–424. https://doi.org/10.1016/j.jalz.2010.04.006

[20] Liu, P.-P., Xie, Y., Meng, X.-Y., & Kang, J.-S. (2019). **History and progress of hypotheses and clinical trials for Alzheimer’s disease**. Signal Transduction and Targeted Therapy, 4(1). https://doi.org/10.1038/s41392-019-0063-8

[21] Huang, L.-K., Chao, S.-P., & Hu, C.-J. (2020). **Clinical trials of new drugs for Alzheimer disease**. Journal of Biomedical Science, 27(1). https://doi.org/10.1186/s12929-019-0609-7

[22] Liu, P.-P., Xie, Y., Meng, X.-Y., & Kang, J.-S. (2019). **History and progress of hypotheses and clinical trials for Alzheimer's disease**. Signal Transduction and Targeted Therapy, 4(1). https://doi.org/10.1038/s41392-019-0063-8

[23] Hardy, J., & Allsop, D. (1991). **Amyloid deposition as the central event in the aetiology of Alzheimer's disease**. Trends in pharmacological sciences, 12(10), 383–388. https://doi.org/10.1016/0165-6147(91)90609-v

[24] O'Brien, R. J., & Wong, P. C. (2011). **Amyloid Precursor Protein Processing and Alzheimer's Disease**. Annual Review of Neuroscience, 34(1), 185–204. https://doi.org/10.1146/annurev-neuro-061010-113613

[25] Hur J. Y. (2022). **γ-Secretase in Alzheimer's disease**. Experimental & molecular medicine, 54(4), 433–446. https://doi.org/10.1038/s12276-022-00754-8

[26] Shinobu Kitazume, Tachida, Y., Oka, R., Keiro Shirotani, Saido, T. C., & Hashimoto, Y. (2001). **Alzheimer's β-secretase, β-site amyloid precursor protein-cleaving enzyme, is responsible for cleavage secretion of a Golgi-resident sialyltransferase**. Proceedings of the National Academy of Sciences of the United States of America, 98(24), 13554–13559. https://doi.org/10.1073/pnas.241509198

[27] Levkovich, S. A., Gazit, E., & Laor Bar-Yosef, D. (2021). **Two Decades of Studying Functional Amyloids in Microorganisms**. Trends in Microbiology, 29(3), 251–265. https://doi.org/10.1016/j.tim.2020.09.005

[28] Ristori, E., Donnini, S., & Ziche, M. (2020). **New Insights Into Blood-Brain Barrier Maintenance: The Homeostatic Role of β-Amyloid Precursor Protein in Cerebral Vasculature**. Frontiers in Physiology, 11. https://doi.org/10.3389/fphys.2020.01056

[29] Mucke, L., & Selkoe, D. J. (2012). **Neurotoxicity of amyloid β-protein: synaptic and network dysfunction**. Cold Spring Harbor perspectives in medicine, 2(7), a006338. https://doi.org/10.1101/cshperspect.a006338

[30] PDB101. (2023 December 23). **Molecule of the Month: Amyloids**. https://pdb101.rcsb.org/motm/189

[31] PDB101. (2023 December 23). **Molecule of the Month: Amyloid-beta Precursor Protein**. https://pdb101.rcsb.org/motm/79

[32] Wake, H., Moorhouse, A. J., & Nabekura, J. (2011). **Functions of microglia in the central nervous system--beyond the immune response**. Neuron glia biology, 7(1), 47–53. https://doi.org/10.1017/S1740925X12000063

[33] Alzheimer's Association. (2023). **2023 Alzheimer's disease facts and figures**. Alzheimer's & Dementia, 19(4). https://doi.org/10.1002/alz.13016

[34] Biology Dictionary (2023 December 23). **Simple Neuron**. https://biologydictionary.net/wp-content/uploads/2021/04/Simple-neuron.jpg

[35] Butler, C. A., Popescu, A., Kitchener, E., Allendorf, D. H., Puigdellívol, M., & Brown, G. C. (2021). **Microglial phagocytosis of neurons in neurodegeneration, and its regulation**. Journal of Neurochemistry, 158(3). https://doi.org/10.1111/jnc.15327

[36] Fowler, D. M., Koulov, A. V., Alory-Jost, C., Marks, M. S., Balch, W. E., & Kelly, J. W. (2005). **Functional Amyloid Formation within Mammalian Tissue**. PLoS Biology, 4(1), e6. https://doi.org/10.1371/journal.pbio.0040006

[37] Kosik, K. S., & Shimura, H. (2005). **Phosphorylated tau and the neurodegenerative foldopathies**. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 1739(2-3), 298–310. https://doi.org/10.1016/j.bbadis.2004.10.011

[38] StressMarq Biosciences inc. (2023 December 23). **Tangles and Tau**. https://www.stressmarq.com/research/neuroscience/neurodegeneration/alzheimers-disease/tangles-and-tau

[39] Waddington, C. H. (1942). **The Epigenotype**. International Journal of Epidemiology, 41(1), 10–13. https://doi.org/10.1093/ije/dyr184

[40] Tronick, E., & Hunter, R. G. (2016). **Waddington, Dynamic Systems, and Epigenetics**. Frontiers in behavioral neuroscience, 10, 107. https://doi.org/10.3389/fnbeh.2016.00107

[41] Zakhari, Samir. (2013). **Alcohol Metabolism and Epigenetics Changes**. Alcohol research : current reviews. 35. 6-16.

[42] Jin, B., & Robertson, K. D. (2013). **DNA methyltransferases, DNA damage repair, and cancer**. Advances in experimental medicine and biology, 754, 3–29. https://doi.org/10.1007/978-1-4419-9967-2_1

[43] Christopher, M. A., Kyle, S. M., & Katz, D. J. (2017). **Neuroepigenetic mechanisms in disease**. Epigenetics & chromatin, 10(1), 47. https://doi.org/10.1186/s13072-017-0150-4

[44] National Human Genome Research Institute (2023 December 23). **Histone**. https://www.genome.gov/genetics-glossary/histone

[45] Bannister, A. J., & Kouzarides, T. (2011). **Regulation of Chromatin by Histone Modifications**. Cell Research, 21(3), 381–395. https://doi.org/10.1038/cr.2011.22

[46] Turunen, T. A., Väänänen, M.-A. ., & Ylä-Herttuala, S. (2018, January 1). **Epigenomics** (R. S. Vasan & D. B. Sawyer, Eds.). ScienceDirect; Elsevier. https://www.sciencedirect.com/science/article/abs/pii/B9780128096574995759?via%3Dihub

[47] Al, N. M., Simpson, B., & Ishwarlal Jialal. (2019, April 5). **Genetics, Epigenetic Mechanism**. Nih.gov; StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK532999/

[48] Bhat, A. A., Younes, S. N., Raza, S. S., Zarif, L., Nisar, S., Ahmed, I., Mir, R., Kumar, S., Sharawat, S. K., Hashem, S., Elfaki, I., Kulinski, M., Kuttikrishnan, S., Prabhu, K. S., Khan, A. Q., Yadav, S. K., El-Rifai, W., Zargar, M. A., Zayed, H., & Haris, M. (2020). **Role of non-coding RNA networks in leukemia progression, metastasis and drug resistance**. Molecular Cancer, 19(1). https://doi.org/10.1186/s12943-020-01175-9

[49] Li Y. (2021). **Modern epigenetics methods in biological research**. Methods (San Diego, Calif.), 187, 104–113. https://doi.org/10.1016/j.ymeth.2020.06.022

[50] Bekris, L. M., Yu, C. E., Bird, T. D., & Tsuang, D. W. (2010). **Genetics of Alzheimer disease**. Journal of geriatric psychiatry and neurology, 23(4), 213–227. https://doi.org/10.1177/0891988710383571

[51] Raulin, A.-C., Doss, S. V., Trottier, Z. A., Ikezu, T. C., Bu, G., & Liu, C.-C. (2022). **ApoE in Alzheimer's disease: pathophysiology and therapeutic strategies**. Molecular Neurodegeneration, 17(1). https://doi.org/10.1186/s13024-022-00574-4

[52] **Alzheimer Disease Stratified by APOE Genotype**. (2019, June 27). Labmedica.com. https://www.labmedica.com/molecular-diagnostics/articles/294778461/alzheimer-disease-stratified-by-apoe-genotype.html

[53] Troutwine, B. R., Hamid, L., Lysaker, C. R., Strope, T. A., & Wilkins, H. M. (2022). **Apolipoprotein E and Alzheimer's disease**. Acta Pharmaceutica Sinica. B, 12(2), 496–510. https://doi.org/10.1016/j.apsb.2021.10.002

[54] Migliore, L., & Coppedè, F. (2009). **Genetics, environmental factors and the emerging role of epigenetics in neurodegenerative diseases**. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 667(1-2), 82–97. https://doi.org/10.1016/j.mrfmmm.2008.10.011

[55] Killin, L. O. J., Starr, J. M., Shiue, I. J., & Russ, T. C. (2016). **Environmental risk factors for dementia: a systematic review**. BMC Geriatrics, 16(1). https://doi.org/10.1186/s12877-016-0342-y

[56] Nativio, R., Lan, Y., Donahue, G., Sidoli, S., Berson, A., Srinivasan, A. R., Shcherbakova, O., Amlie-Wolf, A., Nie, J., Cui, X., He, C., Wang, L.-S., Garcia, B. A., Trojanowski, J. Q., Bonini, N. M., & Berger, S. L. (2020). **An integrated multi-omics approach identifies epigenetic alterations associated with Alzheimer's disease**. Nature Genetics, 52(10), 1024–1035. https://doi.org/10.1038/s41588-020-0696-0

[57] Tecalco-Cruz, A. C., Ramírez-Jarquín, J. O., Alvarez-Sánchez, M. E., & Zepeda-Cervantes, J. (2020). **Epigenetic basis of Alzheimer disease**. World journal of biological chemistry, 11(2), 62–75. https://doi.org/10.4331/wjbc.v11.i2.62

[58] Sharma, V. K., Mehta, V., & Singh, T. G. (2020). **Alzheimer's Disorder: Epigenetic connection and associated risk factors**. Current Neuropharmacology, 18(8). https://doi.org/10.2174/1570159X18666200128125641

[59] Foraker, J., Millard, S. P., Leong, L., Thomson, Z., Chen, S., Keene, C. D., Bekris, L. M., & Yu, C.-E. (2015). **The APOE Gene is Differentially Methylated in Alzheimer's Disease. Journal of Alzheimer's Disease**, 48(3), 745–755. https://doi.org/10.3233/jad-143060

[60] Cleveland Clinic. (2023 December 23). **Alzheimer's Disease**. https://my.clevelandclinic.org/health/diseases/9164-alzheimers-disease

[61] Braak, H., & Braak, E. (1991). **Neuropathological stageing of Alzheimer-related changes**. Acta Neuropathologica, 82(4), 239–259. https://doi.org/10.1007/bf00308809

[62] Huang, Y., Sun, X., Jiang, H., Yu, S., Robins, C., Armstrong, M. J., Li, R., Mei, Z., Shi, X., Gerasimov, E. S., De Jager, P. L., Bennett, D. A., Wingo, A. P., Jin, P., Wingo, T. S., & Qin, Z. S. (2021). **A machine learning approach to brain epigenetic analysis reveals kinases associated with Alzheimer's disease**. Nature Communications, 12(1). https://doi.org/10.1038/s41467-021-24710-8

[63] Rush Alzheimer's Disease Center. (2023 December 23). **Global AD pathology burden**. https://www.radc.rush.edu/docs/var/detail.htm?category=Pathology&subcategory=Alzheimer%27s+disease&variable=gpath

[64] Rush Alzheimer's Disease Center. (2023 December 23). **Braak stage**. https://www.radc.rush.edu/docs/var/detail.htm;jsessionid=8A357F0B2AC2674C66E480CB56A8C424?category=Pathology&subcategory=Alzheimer%27s+disease&variable=braaksc

[65] Teijido, O., & Cacabelos, R. (2018). **Pharmacoepigenomic Interventions as Novel Potential Treatments for Alzheimer's and Parkinson's Diseases**. International journal of molecular sciences, 19(10), 3199. https://doi.org/10.3390/ijms19103199

[66] Nicolia, V., Fuso, A., Cavallaro, R. A., Di Luzio, A., & Scarpa, S. (2010). **B vitamin deficiency promotes tau phosphorylation through regulation of GSK3beta and PP2A**. Journal of Alzheimer’s Disease: JAD, 19(3), 895–907. https://doi.org/10.3233/JAD-2010-1284

[67] Plotkin, S. S., & Cashman, N. R. (2020). **Passive immunotherapies targeting Aβ and tau in Alzheimer’s disease**. Neurobiology of Disease, 144, 105010. https://doi.org/10.1016/j.nbd.2020.105010

[68] Pluta, R., & Ułamek-Kozioł, M. (2020). **Tau Protein-Targeted Therapies in Alzheimer’s Disease: Current State and Future Perspectives** (X. Huang, Ed.). PubMed; Exon Publications. https://www.ncbi.nlm.nih.gov/books/NBK566118/

[69] Illumina. (2023 December 23). **Infinium Methylation Assay**. https://www.illumina.com/science/technology/microarray/infinium-methylation-assay.html

[70] Weinhold, L., Wahl, S., Pechlivanis, S., Hoffmann, P., & Schmid, M. (2016). **A statistical model for the analysis of beta values in DNA methylation studies**. BMC Bioinformatics, 17(1). https://doi.org/10.1186/s12859-016-1347-4

[71] Wang, Z., Wu, X., & Wang, Y. (2018). **A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip**. BMC Bioinformatics, 19(S5). https://doi.org/10.1186/s12859-018-2096-3

[72] Pidsley, R., Y Wong, C. C., Volta, M., Lunnon, K., Mill, J., & Schalkwyk, L. C. (2013). **A data-driven approach to preprocessing Illumina 450K methylation array data**. BMC Genomics, 14(1), 293. https://doi.org/10.1186/1471-2164-14-293

[73] Life Epigenetics (2023 December 23). **Introduction to DNA Methylation Analysis**. https://life-epigenetics-methylprep.readthedocs-hosted.com/en/latest/docs/introduction/introduction.html

[74] Illumina. (2023 December 23). **HumanMethylation450 datasheet**. https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_humanmethylation450.pdf

[75] Olova, N., Krueger, F., Andrews, S., Oxley, D., Berrens, R. V., Branco, M. R., & Reik, W. (2018). **Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data**. Genome Biology, 19(1). https://doi.org/10.1186/s13059-018-1408-2

[76] Youk, J., An, Y., Park, S., Jake June-Koo Lee, & Young Seok Ju. (2020). **The genome-wide landscape of C:G > T:A polymorphism at the CpG contexts in the human population**. BMC Genomics, 21(1). https://doi.org/10.1186/s12864-020-6674-1

[77] Illumina. (2023 December 23). **BS-Seq/Bisulfite-seq/WGBS.** https://www.illumina.com/science/sequencing-method-explorer/kits-and-arrays/bs-seq-bisulfite-seq-wgbs.html

[78] CD Genomics. (2023 December 23). **Whole Genome Bisulfite Sequencing (WGBS)** https://www.cd-genomics.com/whole-genome-bisulfite-sequencing.html

[79] Rakyan, V. K., Down, T. A., Balding, D. J., & Beck, S. (2011). **Epigenome-wide association studies for common human diseases**. Nature Reviews Genetics, 12(8), 529–541. https://doi.org/10.1038/nrg3000

[80] Campagna, M. P., Xavier, A., Lechner-Scott, J., Maltby, V., Scott, R. J., Butzkueven, H., Jokubaitis, V. G., & Lea, R. A. (2021). **Epigenome-wide association studies: current knowledge, strategies and recommendations**. Clinical Epigenetics, 13(1). https://doi.org/10.1186/s13148-021-01200-8

[81] Illumina. (2023 December 23). **CpG loci identification tech note**. https://www.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/cpg-loci-identification-tech-note-m-gl-00921/cpg-loci-identification-tech-note-m-gl-00921.pdf

[82] My Great Learning. (2023 December 23). **Machine Learning Tutorial**. https://www.mygreatlearning.com/blog/machine-learning-tutorial/

[83] Google Developers. (2023 December 23). **Training and loss**. https://developers.google.com/machine-learning/crash-course/descending-into-ml/training-and-loss

[84] ODEN. (2023 December 23). **Model Training**. https://oden.io/glossary/model-training/

[85] Amazon Web Services. (2023 December 23). **Hyperparameter Tuning**. https://aws.amazon.com/what-is/hyperparameter-tuning/

[86] GitHub. (2023 December 23). **EWASplus**. https://github.com/xsun28/EWASplus/tree/remastered

[87] Domino AI. (2023 December 23). **Model Evaluation**. https://domino.ai/data-science-dictionary/model-evaluation

[88] Venugopalan, J., Tong, L., Hassanzadeh, H. R., & Wang, M. D. (2021). **Multimodal deep learning models for early detection of Alzheimer’s disease stage**. Scientific Reports, 11(1), 3254. https://doi.org/10.1038/s41598-020-74399-w

[89] Helaly, H. A., Badawy, M., & Haikal, A. Y. (2021). **Deep Learning Approach for Early Detection of Alzheimer’s Disease**. Cognitive Computation. https://doi.org/10.1007/s12559-021-09946-2

[90] Qiu, S., Joshi, P. S., Miller, M. I., Xue, C., Zhou, X., Karjadi, C., Chang, G. H., Joshi, A. S., Dwyer, B., Zhu, S., Kaku, M., Zhou, Y., Alderazi, Y. J., Swaminathan, A., Kedar, S., Saint-Hilaire, M.-H., Auerbach, S. H., Yuan, J., Sartor, E. A., & Au, R. (2020). **Development and validation of an interpretable deep learning framework for Alzheimer’s disease classification**. Brain, 143(6), 1920–1933. https://doi.org/10.1093/brain/awaa137

[91] Kavitha, C., Mani, V., Srividhya, S. R., Khalaf, O. I., & Tavera Romero, C. A. (2022). **Early-Stage Alzheimer’s Disease Prediction Using Machine Learning Models**. Frontiers in Public Health, 10. https://doi.org/10.3389/fpubh.2022.853294

[92] Mohi, M., Mir Jafikul Alam, Jannat-E-Anawar, Md. Ashraf Uddin, & Aryal, S. (2023). **A Novel Approach Utilizing Machine Learning for the Early Diagnosis of Alzheimer’s Disease**. Biomedical Materials & Devices. https://doi.org/10.1007/s44174-023-00078-9

[93] Segal, M. R. (2017). **Machine Learning Benchmarks and Random Forest Regression**. Escholarship.org. https://escholarship.org/uc/item/35x3v9t4

[94] Gulli, A., Kapoor, A., & Pal, S. (2019). **Deep Learning with TensorFlow 2 and Keras: Regression, ConvNets, GANs, RNNs, NLP, and more with TensorFlow 2 and the Keras API**, 2nd Edition. In Google Books. Packt Publishing Ltd. https://books.google.com/books?hl=en&lr=&id=BVnHDwAAQBAJ&oi=fnd&pg=PP1&dq=keras+regressor&ots=K-r69mVB_3&sig=0xe80WDd9O-zpK5ESnZHME0XpWQ

[95] **ENCODE project**. (2023 December 23). https://www.encodeproject.org/

[96] Intuitive Tutorials (2023 December 23). **BED File Format in Bioinformatics**. https://intuitivetutorial.com/2023/05/02/bed-file-format-in-bioinformatics/

[97] SAM tools. (2023 December 23). **BED**. https://samtools.github.io/hts-specs/BEDv1.pdf

[98] GitHub. (2023 December 23). **bed-reader**. https://github.com/fastlmm/bed-reader

[99] SciKit-Learn (2023 December 23). **r2 score**. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

[100] Neurology Live (2023 December 23). **Epigenetic Therapy Demonstrates Efficacy in APOE Reduction for Alzheimer Disease**. https://www.neurologylive.com/view/epigenetic-therapy-demonstrates-efficacy-apoe-reduction-alzheimer-disease