UAB

Universitat Autònoma
de Barcelona

UNIVERSITAT AUTONOMA DE BARCELONA

DEPARTMENT OF GENETICS

Bayesian neural networks to predict aging and disease risk

Author:
Gerardo Alfonso Pérez

Supervisor:
Dr. Juan Ramón Gonzalez Ruiz
Tutor:
Dr. Mauro Santos Marono

A doctoral dissertation submitted for the degree of

PhD Genetics

September, 2019

Abstract

There has been an increased focus on personalized medicine in recent times. Significant technological improvements in the last few decades generating an explosion in the data available has been one the drivers of the expansion of this field. For instance, the amount of DNA methylation data as well as SNPs data available has increased very substantially. This dissertation focuses on the developments of techniques for analyzing that data applied to the field of aging as well of illness detection, more specifically for cancer and diabetes identification. It will be shown that using a two-step approach consisting of a first stage in which the dimensionality of the data is reduced using algorithms such as Elastic Net, followed with a robust forecasting techniques such as Bayesian Neural Networks is a viable option generating accurate forecast. Other algorithm were also used for illness detection such as Support Vector Machines as well as K-Nearest Neighbors.

This dissertation can be divided into three main sections with the first section covering the topic of biological clocks using DNA methylation data and the previously mentioned reduction of dimensionality combined with Bayesian Neural Networks. The biological clock presented in this dissertation generates age forecasts that are more accurate than some well-known existing clocks. This improvement is accomplished by using a non-linear algorithm. The second section covers the issue of cancer identification using, as in the previous case, DNA methylation data and Support Vector Machines as well as K-nearest Neighbor algorithm. It will be shown that for a large amount of different types of cancer, such as lung, colon, cervical or bladder the usage of DNA methylation data in conjunction with SVM generate accurate forecasts. Other algorithms, such as for instance K-Nearest Neighbors, were also used for cancer detection purposes. The last section cover the study of diabetes using in this case SNPs data and Bayesian Neural Networks that also generates accurate diabetes detection.

Given the ever increasing amount of DNA methylation data as well as SNPs data available as well as advances in data storage there is an increasing need to have more suitable and sophisticated methods for analyzing such data. One of the base assumptions in this dissertation is that the relationship between DNA methylation and aging and cancer as well as between SNPs and diabetes do not necessarily need to follow a linear model and hence non-linear models, such as Bayesian Neural Networks, can generate more accurate results. It will be shown that this is the case with models generating fairly accurate outcomes.

Acknowledgements

I would like to thank my thesis supervisor, Dr. Juan Ramón Gonzalez for all the support during this journey.

Bayesian neural networks to predict aging and disease risk

Gerardo Alfonso Pérez

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Personalized medicine

The field of personalized medicine has experienced a significant expansion in recent times driven, among other factors by an explosion in the amount of genomic data available and the ever increasing need to have more effective and targeted treatments and diagnosis for patients. The National Health Service in the United Kingdom for instance is starting to include genome sequencing in routine care for patients in a new initiative. Some of the early results of that initiative have been on the treatment of diabetes with medication in form of tablets, rather than injections, being offered to some of the patients depending on the results from the genomic results [Pinello, 2018]. According to the NHS this diabetes initiative is not only reducing cost but also providing better care, more personalized for individual patients avoiding in many cases the need for injections.

Most common human diseases, such as coronary heart disease, diabetes, cancers, asthma and Alzheimer's disease have a complex genetic architecture. The decreasing cost of sequencing has facilitated to investigate the predictive relationships between individual's genotypes and complex diseases by analyzing thousands of individual's genomes. Such studies, known as Genome Wide Association Studies (GWAS), have allow identifying hundreds of genomic positions (known as single nucleotide polymorphisms – SNPs) associated with risk of a large number of complex diseases [MacArthur, 2017]. For instance, in the case of type 2 diabetes (T2D), one of the most studied complex diseases, over 80 genomic loci have been identified in association with T2D susceptibility and related traits [Boehnke, 2017].

Unfortunately, barring a few exceptions like the variants in the BRCA1/2 genes in the case of breast cancer [Narod, 2004], the impact on the disease risk of the SNPs identified in GWAS is generally small, and not clinically actionable due to their limited predictive power [McCarthy, 2008]. Even the cumulative signal of all identified loci fails to explain most of the known heritable complex diseases [Manolio, 2009]. This has recently led to a proposal of an "omnigenic model" for complex diseases, where every single gene expressed in the cell types relevant for the disease would have a small but overall impactful contribution to the risk, due to the interconnectivity of cell regulatory networks [Boyle, 2017].

It is thus becoming increasingly clear that the study of SNPs in isolation might only be able to provide limited insight on disease susceptibility, and that a better understanding might be achieved by considering those genomic variants collectively. The current trend is therefore towards a more direct and operational approach to disease prediction problems: learn a predictive function that, from the input of an individual's genotypes (e.g millions of SNPs), predicts the risk of developing a given disease. Predictive disease problems are clearly well suited for standard machine learning methods which are able to decipher the overall genetic architecture of complex diseases.

Offit [2011] mentioned that genomics will play a central role in the development of personalized medicine. Personalized medicine uses traditional medical techniques combined with advanced genetic of the patient providing more individualized care. There are many current examples of the application of genomics in the field of personalized medicine such as for instance in the reduction of adverse reaction on drugs [Odekunle, 2017]. An interesting article in this regard is [Archer, 2013] in which the author discussed the application of genomic in Parkinson's medication. The authors also studied genomics applications of drug induced Parkinson's as a promising field of research. Another interesting application of omics techniques in personalized medicine is on cardiovascular diseases [Chatterjee, 2013]. Among other applications [Chatterjee, 2013] used a genomics technique to assess effectiveness of cardiovascular drugs as well as dosage implications. This article is based on the observation that for some patients some cardiovascular drugs are not efficient even at high dosages.

Jinek [2015] followed perhaps more systematic analysis of personalized medicine and identifies three broad areas of personalized medicine: 1) improved target discovery and confirmation, 2) somatic gene therapy and, 3) tissue engineering. To be fair doctors have been using genetic analysis for decades but the omics revolution is shifting those techniques from single gene to a more systematic genome wide approach. This is possible, as previously mentioned by the improvements in sequencing, which can be done now in a fast and relatively inexpensive way. These technical improvements have in turn generated large amount of data and have come simultaneously with an explosion in the field of big data analysis, It should be mentioned however that there are some ethical considerations particularly us some the omics techniques could accurately forecast increased likelihoods of getting some illnesses for a certain gender or racial ethnicity. There are also some privacy considerations to take into account. Nevertheless, the potential advantages of personalized medicine through the use of omics techniques is undeniable with potential applications in fields such as aging, cancer and diabetes research as we outline in the next subsections.

1.1.1 Aging

Aging is the natural process of decay or more precisely described a “process of steady physiological deterioration” [Linford, 2013] and it is related to the concept of lifespan. Different species have very different lifespans and they can vary drastically. For instance, the Drosophila melanogaster has an average life expectancy of approximately 50 days [Paaby, 2009, Linford, 2013, Khazaeli, 2005] while blue whales have a life expectancy approaching 90 years. The process of aging is clearly not yet well understood with [Karin, 2018] mentioning that it reasons behind the different aging rates in organism that are otherwise genetically identical are unclear.

There are many ways of assessing the aging process in for example humans and other organisms with indicators such white hairs, hair loss or wrinkles typically associated with the aging process. In this context biological clocks can be a useful tool. A biological clock is a toll that aims to estimate the biological age of an individual using empirically measurable biological data such as DNA methylation. There are two seminal articles in this areas [Horvath, 2013, Hannum, 2013]. These two authors created two of the most popular biological clocks based on DNA methylation. These clocks have been used to associate biological age with different diseases and mortality [Fransquet, 2019].

1.1.2 Cancer

Cancer is one of the major causes of death across the world. According to figures from the U.S. National Cancer Institute, cancer is as of 2016 the second cause of death in the U.S. accounting for 21.8% of all death, see figure 1.1. Heart disease was the only other factor causing more deaths than cancer with a total of 23.1% of the death. It should be noted that the proportion of deaths attributable to heart diseases has come down significantly over the last few decades with 37.8% of all death in the U.S. in 1975 attributable this causes. At the same time the proportion of death attributable to cancer has actually increased in the same period from an initial 19.2% in 1975 to the previously mentioned 21.8% in 2016. There is a large number of different types of cancers with very different incidence (1.2) and survival rates. Given the mortality rate of some of these cancers it is of clear importance to develop tool for cancer detection. One option is to use DNA methylation data for tissues or blood from a suspected patient in combination with techniques such as support vector machines, neural networks a K-nearest neighbors which are illustrated in the next sections.

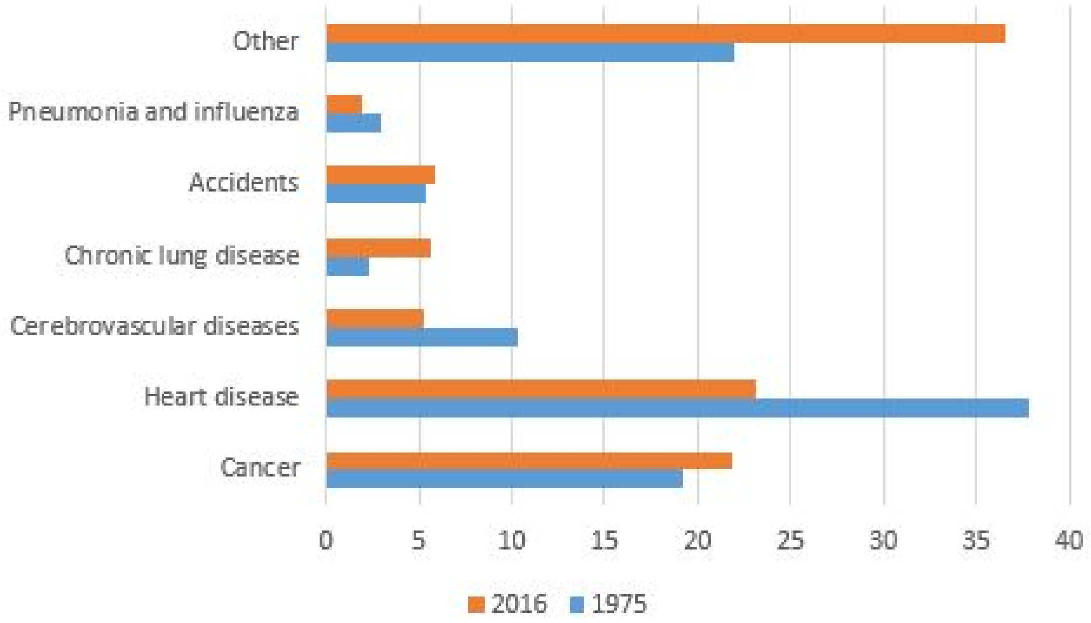


Figure 1.1: Causes of death (%) in the U.S. in 2016. Source:U.S. National Cancer Institute

There is an increased focus on applying statistical techniques in cancer research. This is partly due to the large amount of information made public by institutions such as the National Center for Biotechnology Information (NCBI). There is currently a significant amount of human DNA methylation data publically available from such source. DNA methylation is natural process in which a methyl group links with a cytosine in a specified location (carbon 5). This is a process that occurs naturally in humans and other species and has been linked to many biological processes and diseases such as cancer. Changes in methylation levels in tumors have been analyzed in many articles such as [Das-Partha, 2004, Phillips, 2008]. There is an increasing literature linking DNA methylation changes with several types of cancers [Varela-Rey, 2013, Warton, 2016, Davis, 2004, Oakes, 2013]. Given the efforts that have been made trying to study the relationship with DNA methylation and cancer it seemed reasonable to try to apply algorithm such as Neural Networks or K-Nearest Neighbors to DNA methylation data. It will be shown that using this techniques it is possible to classify, within a reasonable

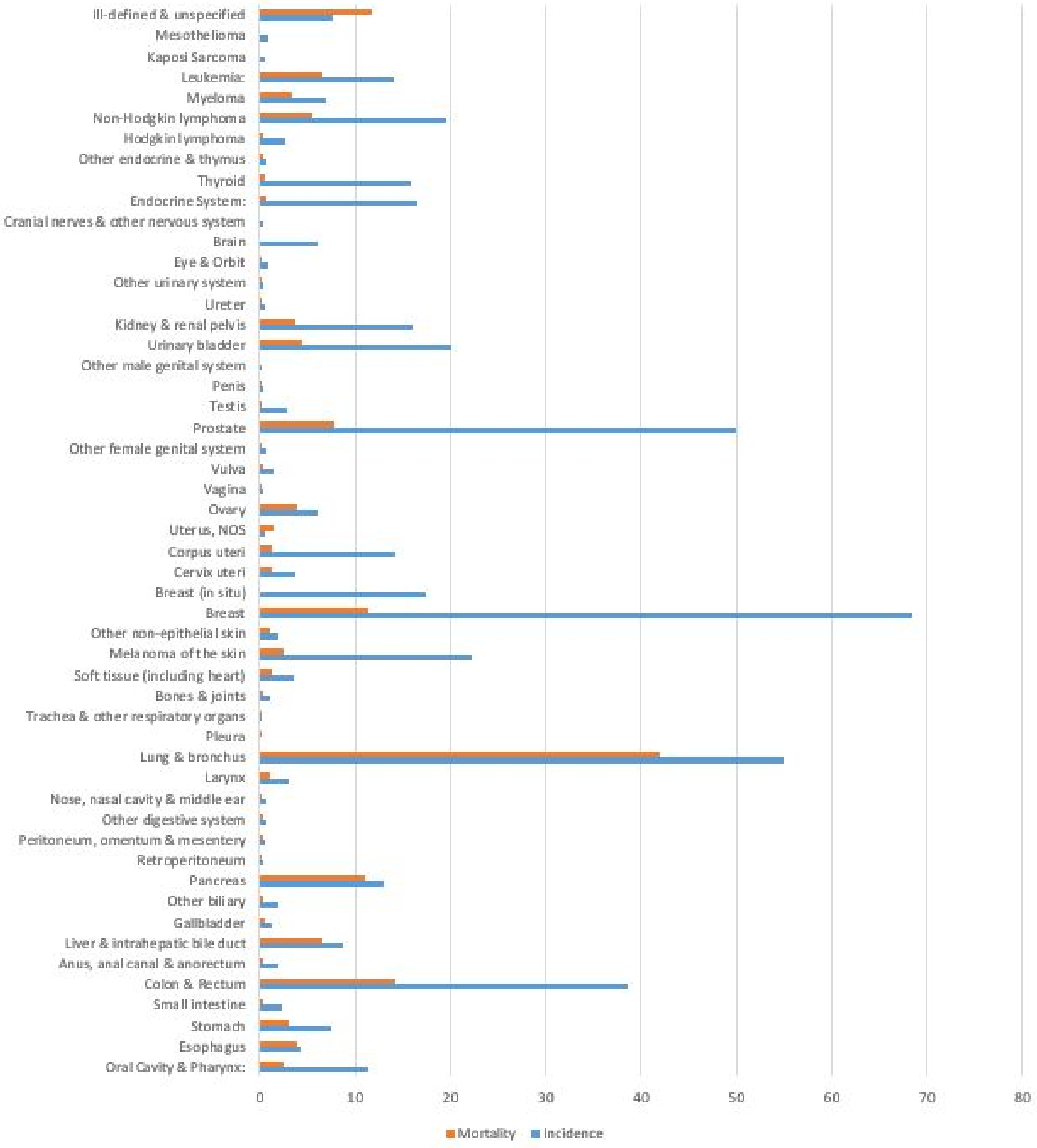


Figure 1.2: Cancer incidence and mortality rates (per 100,000 individuals) in the U.S. 2016. Source:U.S. National Cancer Institute

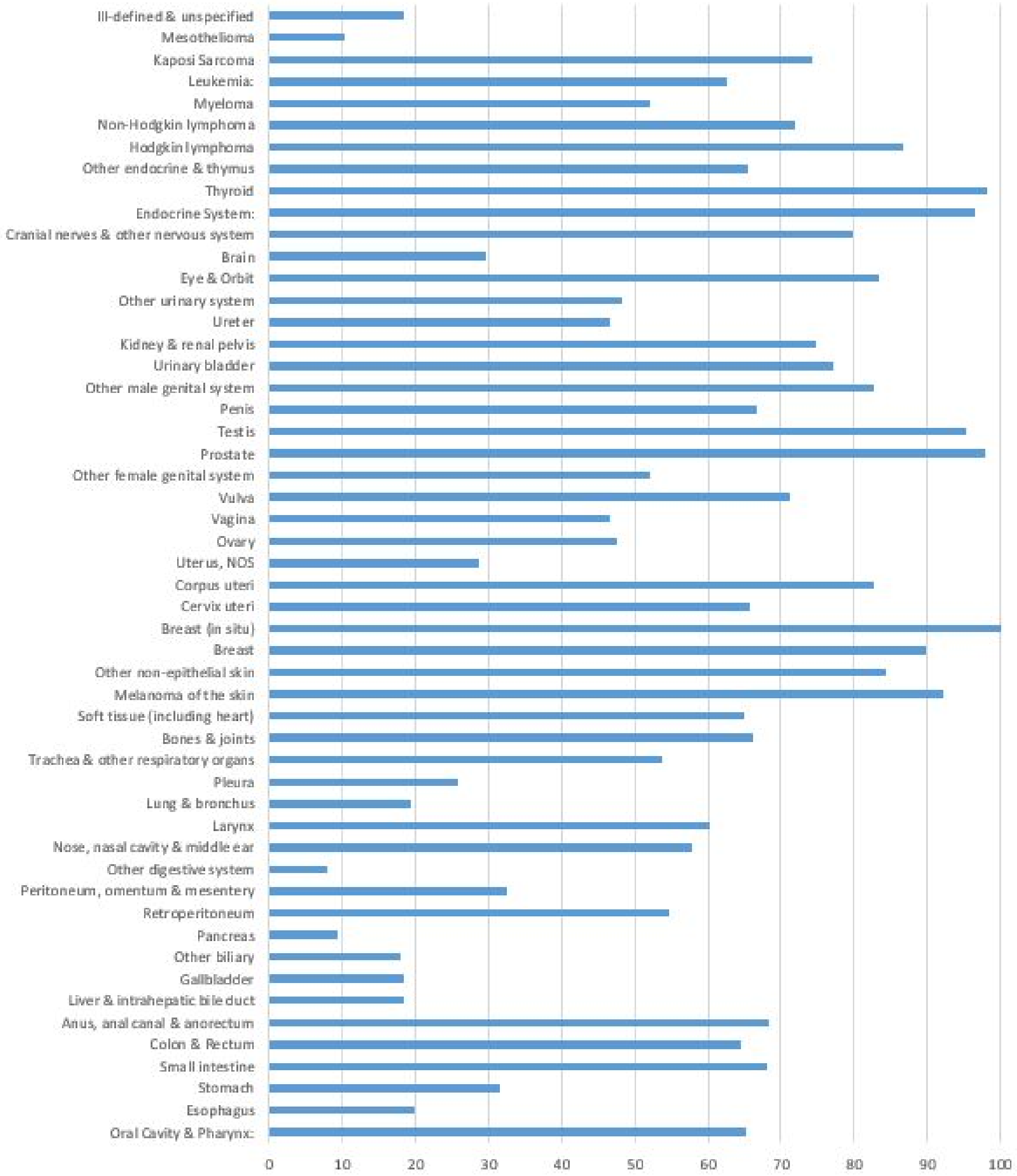


Figure 1.3: Cancer survival rate (%) in the U.s. in 2016. Source:U.S. National Cancer Institute

level of error, patients into two categories. One representing cancer patients and one representing healthy control patients.

1.1.3 Diabetes

Type 2 diabetes is a relatively common but serious illness that impacts life expectancy. According to [Cecil, 2017] type 2 diabetes impacts approximately 8% of the population in the United States with the percentage as high as 25% in high risk populations such as individuals above 65 years old. Type 2 diabetes is the more frequent form of diabetes accounting for 90% to 95% of all the cases [CDC, 2017]. In 2017 diabetes was the 7^{th} cause of death in the United States with 83,564 related death, accounting for 25.7 death per 100,000 people [CDC, 2017]. The trend has remained basically constant over that least few years with the number of days being 20.8, 21.3 and 21.0 per 100,000 individuals in 2010, 2015 and 2016 respectively. Type 2 diabetes has a genetic components and is also related to several lifestyle factors such as weight and physical exercise. Another important risk factors are the smoking and alcohol consumption. Nevertheless, body weight is considered the most important risk factor.

Type 2 diabetes impacts the ability of the patient to produce insulin. This ability to produce insulin declines as the illness progress. The previously mentioned factors of body weight and physical exercise are a critical factor in the regulation of insulin [Cecil, 2017]. One of the most common complications of diabetes are microvascular accidents that can generate lesions that require amputations such as fore example finger amputations. Patients with Type 2 diabetes have an anomalous reaction to insulin i.e., their reaction to an injection of insulin is much smaller than the reaction of an otherwise healthy patient to same dosage of insulin. According to figure from the International Diabetes Federation by 2050 there will be in excess of 500 million individuals worldwide with diabetes with, as previously mentioned, obesity playing a significant role in the spread of the disease. Morbid obesity accounts for approximately 55% of all the cases [CDC, 2017]. Some authors have also related this spread to living and eating habits in Asia and other developing areas becoming more westernized with an increased in the consumption of fast food with high caloric concentration. Given the significant mortality rate in the population and the relatively large percentage of the population at risk of contracting type 2 diabetes, particularly as individuals age, having reliable techniques to identify the illness in a quickly, and hopefully inexpensive way, is of clear importance.

1.2 OMICS

The word omics in general terms refers to the collection of large amount of data, generally related to molecular biology. In fact, the medical dictionary defines omics as thee “analysis of large amounts of data representing an entire set of some kind, especially the entire set of molecules, such as proteins, lipids, or metabolites, in a cell, organ, or organism” [Medical-dictionary, 2007]. As just mentioned one of the characteristics of these fields is that they are interrelated to the field of big data. These research fields generate enormous amounts of data and with them numerous issues such as data analysis and interpretation. The advances in data storage with the development of cloud services to store massive amounts of data at relatively inexpensive rates and high speed connections has also helped the expansion of the field of omics. Clearly, the development of tools for processing those large volumes of information is becoming a critical area of research [Horgan, 2011].

The field of omics is a vast an expanding one with different multiple applications that encompasses a large set of research fields such as genomics, transcriptomics and epigenomics. Genomics could be defined as a field of genetics that attempts to understand the information contained in the genome. The U.S. national library of medicine define genome as the “organism’s complete set of DNA, including all of its genes” [Horgan, 2017] and contains all the genetic information necessary to construct such organism. The World Health Organization defined genomics as “the study of genes and their functions, and related techniques” [WHO, 2002].

Transcriptomics in simplified terms is a field of research that focuses on the area of gene expression [Strachan, 2018]. A perhaps more complete way of describing it is “the study of the complete set of RNAs (transcriptome) encoded by the genome of a specific cell or organism at a specific time or under a specific set of conditions” [Kuchka, 2012]. Epigenomics, which is classified by some authors as a subfield of genomics, is a discipline that analyses “changes that modify the expression and function of the genetic material” [Merriam-Webster, 2017].

The fields of omics has benefited substantially from technical developments [Mullish, 2018] such as improvements in DNA sequencing [Bogyo, 2013, Ulloa, 2013, Gubb, 2009] lowering its cost and producing in turn a proliferation on the amount of datasets that would have been unthinkable a few decades ago. There is a very significant body of existing literature in omics with multiple journals focusing on this field. The study of the different subfields of omics is becoming so prevalent that some authors, such as [Abuasad, 2011], have described the twenty century as the era of omics. It will be shown later that one very successful are of application of omics techniques is in medicine [Mullish, 2018] with some authors, such as [Chris-Overall, 2011] mentioning that every disease has a genetic component. Other areas in which omics techniques have been successfully applied include fields as diverse as agriculture and environmental science [Debmalya-Barh, 2017, Canas, 2014].

1.2.1 Genomics

Genomics was actually the first of the omics disciplines to be developed and as previously described can be understood as the study of all the genes (genome) and their function [Genome-Website, 2018] of an organism. In order to understand the genes of an organism it is necessary to obtain the DNA from its cells and sequence it. The sequence of the organism analyzed can then be compared to the sequence from another organism from the same species to look for mutations (differences in base pairs) and see the functional impact that those mutations have on the organism. Mutations can have impact of a wide arrange of factors, from the color of the eyes of a person to how susceptible that individual is to get an illness. A human being has approximately 3.2 billion base pairs distributed in 23 chromosomes (humans have actually two sets of 23 chromosomes, one from the mother and one from the father). There are approximately 500 to 4,000 genes per chromosome. The vast majority of the genome is not protein coding with approximately only two percent of the genome being protein coding.

Clearly genomics is a huge field of research with vast number of applications in the biomedical space. A recent publication from the World Health Organization identified the following genomics applications/major projects in the field of healthcare in the following countries:

- Brazil. National program developing the use of genomics in healthcare
- Cyprus and Sardinia. National genomics program increasing awareness of thalassemias
- Iceland. National genomics program targeting 30 disease including Alzheimer's and emphysema
- Germany. New anti-malaria treatment
- Kenya. HIV treatment research. Collaboration between the University of Nairobi and the University of Oxford. A Kenyan population group seems to have relatively high resistance to contract HIV despite frequent exposure
- Mexico. National program. Existing drugs seem to treat some diseases seem to have reduced effectiveness in the population of indigenous background compared to the population of Spanish background.
- UK. Malaria research. New vaccine being developed. This new vaccines contains DNA from P. falciparum

According to the U.S. National Library of Medicine SNPs are among the most common genetic variations and account for approximately 90% of sequence differences [Collins, 1997]. SNPs are differences in DNA at the single nucleotide level [Horgan, 2017]. It is not well understood the causes of SNPs [Wang and Moult, 2000]. Some authors have mentioned that the majority of SNPs have no functional impact [MIT, 2013] but a subgroup of SNPs, according to [Albert, 2010] are responsible for a large amount of hereditary human individuality including how a particular individual reacts to a certain drug treating an illness. It has been mentioned by many scholars the complexity of finding relevant SNPs among the many available directly related to illnesses [Florez, 2008]. The field of SNPs research and their impact on the understanding of some diseases is rapidly expanding but it is not without issues as there is an increasing need to development better and more efficient techniques to analyze them [Helyar, 2011]. Nevertheless, the field is in clear expansion with applications not only as biomarkers for diseases but also in many other fields such as for instance crop genetics [Tabassum-Jehan, 2006]. SNPs can be classified in two major categories according to what nucleotide substitution occurred [Smith, 2002]:

- Transition
- Transversion

Transition occurs when the substitution is between either two purines (adenine and guanine) or two pyrimidines (cytosine and thymine). Transversion happens when the substitution is between nucleotides from the different groups i.e., between a purine and a pyrimidine. [Smith, 2002] also mentioned that SNPs are not homogeneously distributed in the genome with more concentration of SNPs in the coding areas than in the non-coding areas. It is also impotent to mention the difference between a synonymous substitution and a non-synonimous substitution. In a synonymous substitution the amino acid encoded does not change while in a non-synonymous substitution the amino acid encoded does change [Smith, 2002].

Similarly to the case of DNA methylation, the amount of SNPs data available has expanded enormously in recent times on the back of technological improvements such as the development of high- throughput SNP Arrays [Shahid-Raza, 2016] and that has created an increasing need to develop techniques that are able to analyze the very large amount of information generated in experiments.

1.2.2 Epigenomics

The term “epi” in the word epigenomics is a Greek expression meaning above, so literally epigenomics means above the genome. The field of epigenetics studies the external modifications of DNA that while not altering the actual DNA sequence it impacts gene expressions. While the field has experienced substantial growth in recent years it origins (in modern form) dates back to the mid twenty century with the term epigenetics first used by Waddington in 1942 [Waddington, 1942]. The difference in terminology between epigenetics and epigenomics is that in the first case one gene is analyzed while in the second case all genes are analyzed. The field is also closely related to the process of the tight folding of the genome. Tighten folded genes might make the process of gene expression for those genes mode difficult (inducing gene suppression) while loser folding can potentially have the opposite effect, potentially making gene expression easier.

There are many studies in this field using identical twins [Craig, 2015] as they have identical DNA sequence but they present different characteristics throughout their life as environmental factors impact DNA expression. These epigenetic changes can produce for instance that a twin develops a cardiovascular disease while the other twin does not. Authors such as [Craig, 2015] found that epigenetic variation in early life is directly related with the onset of several diseases later in life. Epigenomic changes occur when a chemical tag attaches to the DNA impacting the way that the DNA expressed. There are four major groups of epigenetic modifications [ProteinTech, 2017]:

- DNA methylation
- Non coding RNA (ncRNAa)
- Covalent histone
- Non-covalent

One of the best understood epigenetic modifications is DNA methylation that could be described as the addition of a methyl group to a CpG.

While there is an extensive body of research in epigenomic mechanisms many of the underlying processes remain not fully understood [Pinello, 2014]. Nevertheless, relatively recent developments, such as for instance, alignment-free techniques seem to be yielding interesting results [Pinello, 2014]. This and other computational advancements are likely to further expand the field of epigenomics as more data and better statistical techniques to analyze it become available.

DNA methylation is an epigenetic modification [Caifa, 2004, Ju-Yeon, 2012, Schuebeler, 2012, Serrano, 2018], based on the addition of a methyl group in the fifth carbon (C5) in CpG dinucleotides. [Patterson, 2011a] mentioned that the primary sequence target for mammals related to methylation is the 5-CpG-3 adding that methylation tends to be concentrated in regions that are commonly described as CpGs islands. A methyl group is just a group formed by a carbon and three hydrogens (CH_3) as it is one of the most frequent organic compounds. The reaction of DNA methylation is catalyzed by DNA methyltransferases Dnmt3a and Dnmt3b [Paige-Bommarito, 2019, Grant-Challen, 2014] and preserved by Dnmt1. The chemical expressions for cytosine and methyl cytosine can be seen in figures 1.4 and 1.5 respectively. DNA methylation is a naturally occurring process and can alter the function of the DNA. DNA methylation is conserved when a cell divides so it has being described by some scholars as a cell memory mechanism [Paige-Bommarito, 2019]. DNA methylation is known to have an impact on gene expression [Razin, 1991, Brandeis, 1993, Jacob, 2000, Paige-Bommarito, 2019].

DNA methylation has been related with several processes such as aging as well as illnesses including multiple types of cancer, Chron's disease and diabetes and has being used as a biomarker. Biomarker are "naturally-occurring characteristics by which a particular pathological process or disease can be identified or monitored" [Mikeska, 2014]. Perhaps one of the areas in which there is more existing academic literature in the topic of DNA methylation is in the field of cancer research. For instance [Das-Partha, 2016] highlighted the applicability of DNA methylation for early detection of tumors and even mentioned the ongoing efforts to develop drugs that reverse methylation, such as for instance 5-azacytidine and decitabine.

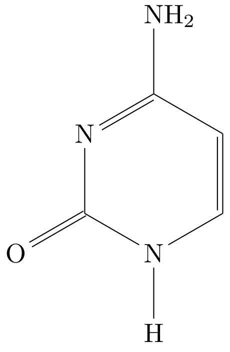


Figure 1.4: Cytosine

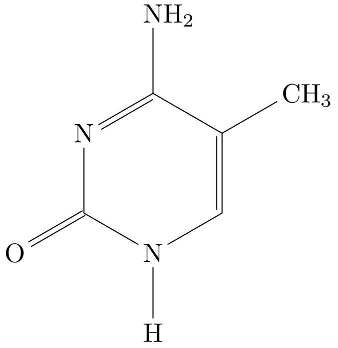


Figure 1.5: Methyl Cytosine

The causes behind the methylation process are yet not well understood with several authors, such as for instance [Horvath, 2013, Hannum, 2013] linking it, at least to some degree to the aging process. There is also extensive literature relating various types of cancer and DNA methylation levels such as for instance: breast cancer [Magzoub, 2016, Varela-Rey, 2013], ovarian cancer [Magzoub, 2016, Woloszynska-Read, 2007], colorectal cancer [Magzoub, 2016, Carmona, 2013], upper aerodigestive track cancer [Varela-Rey, 2013], liver cancer [Varela-Rey, 2013, Yonghong-Zhang, 2018], lung cancer [Leygo, 2017, Qinghua-Feng, 2008, Tibor-Rauch, 2012, Nikolaidis, 2012, Chandrika-Piyathilake, 2002, Pfeifer, 2017, Malcolm-Brock, 2008], bladder [Xylinas, 2016], Kidney cancer [Niraj-Shenoy, 2015].

1.2.3 Transcriptomics

As previously mentioned transcriptomics is the field of gene expression. This could also be understood as the analysis of all the RNA existing in a cell (transcriptome), including coding and non-coding RNA. The amount of non-coding RNA is in fact larger than the coding RNA. Interestingly, some authors such as [Hrdlickova, 2014] have found that non-coding RNA play an important role in several diseases and play a significant regulatory function. [Hrdlickova, 2014] classified codding and non-coding RNA according to the following classification:

1. Coding genes (20,330)
2. Non-coding genes
   (a) Small non-coding RNA
      i. MicroRNAs (3,086)
      ii. Other small ncRNAs (5,992)
   (b) Long non-coding RNA
      i. LincRNAs (6,020)
      ii. Atisense IncRNAs (4,589)
      iii. Sense intronic IncRNA (674)
      iv. Other IncRNA transcripys (1,909)
      v. Sense overlapping IncRNAs (141)

Gene expression refers to the type and amount of proteins present in the cells of an organism. Changes in the amount of RNA generated can change the amount of protein present changing the behavior of cells. Factors such as the type of tissue or environmental conditions lay also a major role in the amount of protein present. The process of gene expression is clearly extremely complex. Some authors, such as [Dong-Zhicheng, 2013] have mentioned that transcriptomics is one of the best developed fields in the omics space with a large amount of research published in the topic.

1.3 Machine learning methods in personalized medicine

In recent years there has been an increased interest in applying machine learning algorithms to biomedicine ranging from the work of [Horvath, 2013] on biological clocks using elastic net regression to cancer identification techniques using neural networks with various degrees of accuracy [Gorynski, 2014]. The expansion of the sector is likely related to a combination of factor such as substantial increases in computational power as well as more data available from experiments using technological developments such as high- throughput arrays.

It will be shown later in this dissertation that combining some techniques such as for instance elastic net and neural network generate better results than using directly neural networks on DNA methylation data when trying to estimate the biological age of an individual. This approach will be shown to be also better than applying linear regression models.

Neural networks have been used in this field for some time now. For instance, neural networks have been used for automatic identification of lung cancer using mammography [Tariq, 2018]. Another interesting article on this regard is [Miles-Jefferson, 2000]. In this article the authors used neural networks to predict the outcome of breast cancer surgeries using as input data such as the expected size of the tumor and location. A similar approached was followed by [Chang, 2018] but in this case with gliomas.

Oustimov [2014] highlighted the increasing importance of neural networks in the field of cancer genomics and biomarkers. Two of the most important biomarkers in the detection of illnesses such as cancer and diabetes are DNA methylation and SNPs. These are obvious fields of research because there is ample existing literature linking DNA methylation with cancer [Das-Partha, 2004]. There have been some studies applying neural networks to cancer identification using DNA methylation such as for instance [Coppede, 2015] on colorectal cancer. Perhaps one of the commonalities of many of these articles is that they focus on a specific subset of cancer, such as for instance in this case colorectal cancer. It should be noted that in this dissertation the focus was on trying to develop more generic algorithm that could potentially be used as a wide screening tool. Genetic scores are another biomarker frequently used as tools to estimate the predisposition to get certain types of illnesses such as diabetes and cancer [Udler, 2019].

Another interesting application for cancer identification are support vector machines. For instance, [Kim, 2016] applied them in the field of breast cancer detection. Perhaps, one of the issue with support vector machines, similar to some degree of neural networks is the out-of-sample forecasting error, particularly when applied to different types of cancers. One more algorithm, among the many, that deserves to be mentioned are k-nearest neighbours. This is a less complex algorithm than neural networks or support vector machines but it remains nevertheless very useful in many applications in this field. Some interesting articles describing the application of this technique in the field of cancer identification are [Bhuvanesuari, 2015, Xianglin-Zhang, 2019].

There field of machine learning applied to genomic data has seen a significant expansion in recent times as more data became available. [Libbrecht, 2015] for instance mentioned the current use of machine learning techniques to identify transcription start sites (TSS). Other applications of machine learning identified by [Libbrecht, 2015] include splice sites, promoters and enhancers. The author also compared the typical machine learning approach with the scientific way. In machine learning, an algorithm is first identified or developed. Then the algorithm is trained with a training data set and then the data is tested out of sample with new data. This, in the opinion of [Libbrecht, 2015] is not too dissimilar from the scientific was of developing and hypothesis and then testing it with experiments.

Chapter 2

Hypothesis and objectives

2.1 Hypothesis

The general hypothesis of this dissertation is that the of use advanced statistical methods based on machine learning joint with a dimensional reduction of the large amount of omic data can improve the predictive capacity of predictive models. In particular, we propose to:

1. Reduce the dimensionality of the omic data using Elastic Net regression.

2. Use techniques with high predictive level such as Bayesian Neural Networks techniques

This general hypothesis can be summarized in these specific hypothesis:

1. DNA methylation can be used to generate forecast profiles for individuals assessing the risk of contracting complex illnesses such as cancer. While reversible DNA methylation changes are stable and have a significant impact on gene expression having therefore an impact on cancer.

2. Chronological age can be estimated using DNA methylation data. Some evidence of anti-aging interventions, such as calorie restriction, having a sig-

nificant impact on epigenomic clock (and DNA methylation) support this hypothesis.

3. Genetic factors inherited through genetic polymorphism predispose individuals to contract illnesses such as diabetes.

2.2 Objectives

The overall objective of this dissertation is to show that the proposal of modeling in two steps, by first reducing the dimensionality of the data followed by the application of advanced models such as Bayesian techniques, is a viable approach to generate accurate predictive models. To this end, we propose two address this issue by covering three main objectives representing three different areas where personalized medicine have provided some advances. We aim to create:

1. An accurate biological clock using DNA methylation data.
2. Models to classify individuals with cancer using DNA methylation data.
3. Models to predict individuals at high risk to suffer diabetes using genomic data.

Chapter 3

Methods

In this chapter it is shown the methods applied for the analysis of aging, cancer and diabetes.

3.1 Aging

3.1.1 Age estimation using neural networks and DNA methylation levels

There are several techniques available to determine the biological age of a patient by analyzing the DNA methylation levels of some of their cells. In this section the forecasting accuracy of neural networks is compared to the k-nearest neighbors ("KNN") technique. The accuracy of the forecast is related to the sample size. For smaller datasets the KNN provide some moderately accurate results, with an average error of approximately 10 years. When the sample size increase the KNN does not appear to work properly (for a dataset of 720 samples) and neural networks start to provide better results. While the amount of samples in each dataset varied the number of CpGs per case was constant at approximately 27,000. Several simulations were performed randomly reducing the number of CpGs in the samples. It was found that typically the best results were found not when using all the CpGs (27,000) but when using a relatively randomly selected subset of approximately 300 to 400 CpGs. While the sample size is too small to be conclusive the results seem to indicate that for an age forecasting point of view a significant fraction of the CpGs methylation data might add mostly noise. It is clearly required further work to determine is this intuition is actually correct. DNA methylation is a normal process that is impacted by environmental factors [Lucia, 2012] and life style. Indications of the central role that methylation has in many biological process is known since 1979, after the highly influential paper [McGhee, 1979]. Since then there has been an increasing amount of literature involving methylation and many specific processes from diseases [Daniels, 2008] to aging [Mansego, 2015]. From a biochemical point of view methylation occurs when a methyl group links to a base (either C or G). The level of methylation changes from tissue to tissue and it is different if the individual has some disease, such as cancer. Age is also a factor impacting methylation. There are indications that newborn methylation levels are impacted by maternal smoking during pregnancy [Joubert, 2016] and even might have an impact on memory [Day, 2010]. Currently it is relativity straightforward obtaining methylation data from many different cells, such as sperm [Cassuto-Nino, 2016] or colon cells [Fernandez, 2012] with the majority of the sample publically available being of whole blood. One widely accepted technique to determine DNA methylation levels is bisulphite modification [Patterson, 2011b]. Thank you to this technique and similar approaches the accuracy of DNA methylation measurements has increased substantially over the last decade. While there has been a large amount of research regarding methylation there continuous to be many questions remaining such as the exact role the methylation has in the aging process or if methylation changes can be induced to prevent certain illnesses.

Methylation has been mentioned in a multitude of research reports as an aspect influencing the aging process in humans [Rowbotham, 2014]. Changes in methylation levels related to aging have been measured not only in humans but also in some other species such as mice [Sabine, 2017], salmon [Marc, 2015, Berdishev, 1967] or great apes [Hernando-Herraez, 2013]. There seems to be a consensus in the literature with the existence of some type of relationship between DNA methylation levels and aging but less of a consensus of how the aging process actually occurs or if changes in DNA methylation can actually increase life spans [Marc, 2015]. Abnormal methylation levels do appear to be related with premature aging and some illnesses. It should be noted that currently it is possible to induce changes in DNA methylation and that this is an active area of research. Methylation alteration has been mentioned as an easier way to modify DNA than through mutations [Daniels, 2008]. There currently exist accurate multi tissue clocks, such as [Horvath, 2013] , that can predict biological age of a person using methylation levels from several different types of tissues with an error of only a few years. All these indications points towards some type or relationship between methylation levels and aging and warrant doing further research on what statistical applications to use. In this section neural networks and the k- nearest neighbor approach were followed to link those DNA methylation levels with the patient age.

Neural networks are a statistical application that has proven valuable for signal fitting. It is biologically inspired and similar to many other machine learning application does not require theoretical knowledge of the relationship between the input and the output. The first theoretical steps in the neural network space track back to the late 50^{th} early 60^{th} but these techniques only became popular several decades later with the development of computers.

A neural network is composed of artificial neurons, which could be described as a transfer function that generate an output as a result of a given input. Biological neurons (3.2) inspired researcher to create the model of an artificial neural network as a self learning tool. Artificial neurons are typically arranged in layers, a typical neural network structure with two hidden layers is shown in figure 3.3. There are many different types of different neurons, some of the most frequently can be seen in equations 3.1, 3.2, 3.3, 3.4 :

1. Sigmoid:
$$\zeta(x) = \frac{1}{1+e^{-x}} \tag{3.1}$$

2. Tansig:
$$\zeta(x) = \frac{2}{1+e^{-2x}} - 1 \tag{3.2}$$

3. Linear:
$$\zeta(x) = x \tag{3.3}$$

4. Radial:
$$\zeta(x) = e^{-x^2} \tag{3.4}$$

The output of these neurons can be seen in figure 3.1.Each neuron in a neural network has an associated weight. In this section supervised learning is used. Supervised learning is an algorithms that trains the network using a training data set in which the output is known i.e., if the patient has cancer or not. After the network is trained it can be used to estimate if a new patient has cancer. There are many types of different training algorithms with different levels of sophistication and accuracy. In simplifies term what the training algorithm does is changing the weights on the neurons so the output generated by the model is as close as possible to reality.

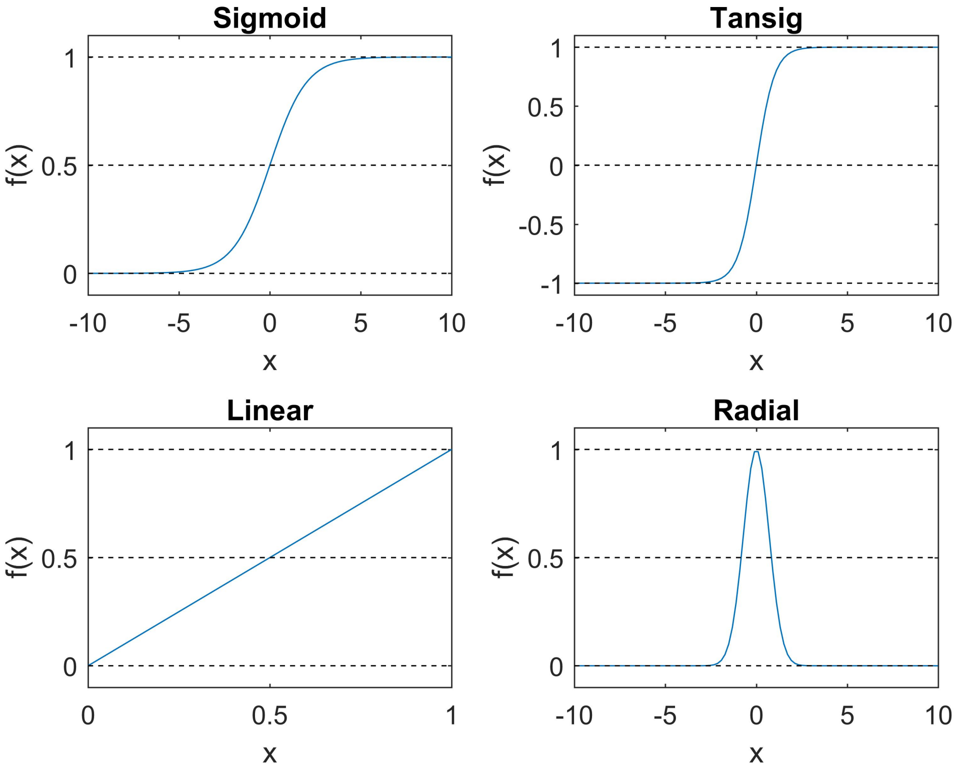


Figure 3.1: Output of frequently used neurons

One of the most successful applications of neural networks was in the field of supervised learning. For supervised learning applications a neural network composed of a number of neurons is trained to replicate an actual output as closely as possible by adjusting the relative importance of the value of those neurons. Then the network is typically tested with new data to try to identify its generalization power. There is a huge amount of different neural networks. Some of the main differences are the network structure, the type of neurons used and the training algorithm. Neural networks have been applied in some areas of medical research, such as forecasting of growth of staphylococcus in milk [Orawan, 2016] or medical diagnosis [Kadhim, 2011]. These techniques are typically used when the underlying relationship between the exogenous and endogenous variables is not known or when such relationship is too complicated to model explicitly.

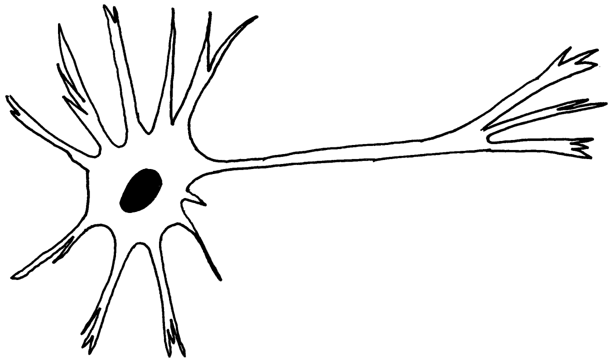


Figure 3.2: Biological neuron

The following databases were used: 1) (GSE56606) containing methylation information of CD14+ monocytes for 100 patients with diabetes as well as control subjects from [Rakyan, 2011], 2) (GSE34035) containing methylation data for saliva of 197 patients with different alcohol consumption from [Liu, 2010], 3) (GSE24884) methylation data of subcutaneous adipose tissue of 56 patients from [Arner, 2015], and 4) (GSE41037) database of 720 patients suffering from schizophrenia as well control subjects from [Horvath, 2012]. All the datasets are publically available. In a first instance, a neural network (backpropagation) with 10 neurons was applied to all the three datasets. As expected the results were

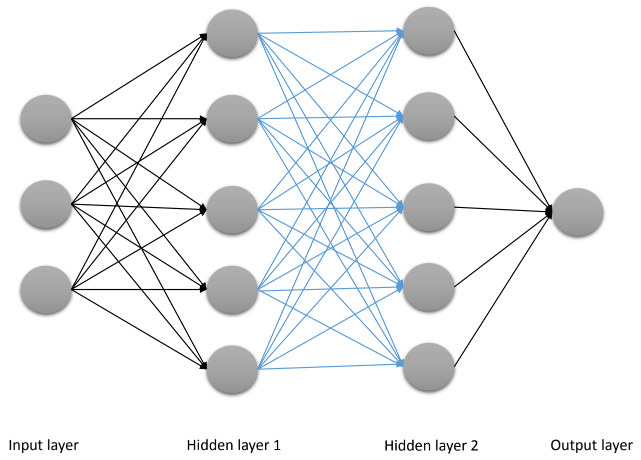


Figure 3.3: Biologically inspired artificial neural network

considerably more accurate for the large data set than for the smaller ones. No meaningfully prediction was obtained for the smaller datasets using neural networks.

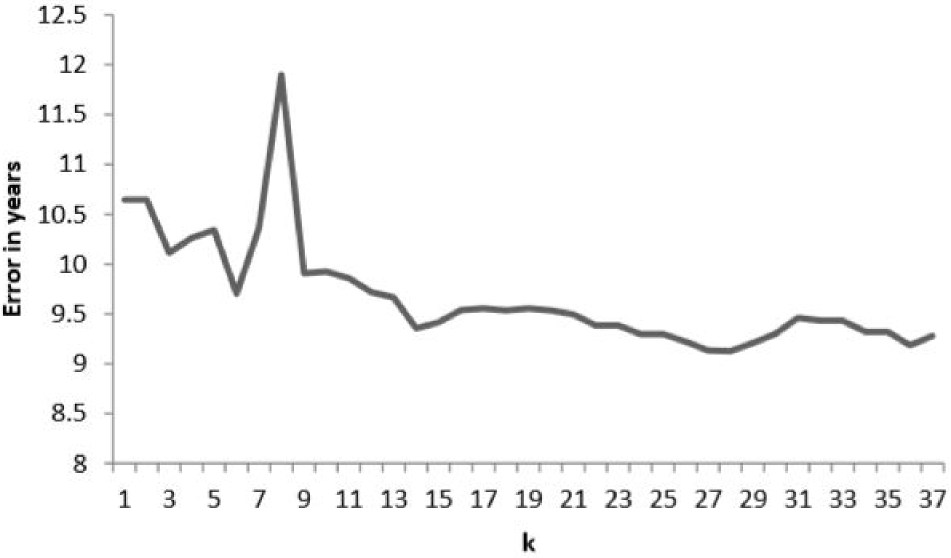


Figure 3.4: Mean error for out of sample values - KNN (GSE34035)

Then the KNN technique was used. The KNN approach was followed using a series of values of k, from 1 to 35 as, well as for several distance measures, such as Euclidean, Cityblock, Correlation and Cosine. The results, for some of the smaller datasets, can be seen in figures 3.4 and 3.5. The GSE41037 dataset was then sliced into smaller subsets to see if as the number of samples increased the accuracy of the KNN improved (figure 3.6) but this was not observed (perhaps due to the limited sample size). The regressions for two subsets of GSE41037 can be seen in figures 3.7 (75 cases) and 3.8 (100 cases) and some more details in table 3.1.

N. of samples	**75**	**100**	**575**
P1	[0.0042, 0.1969]	[0.0467, 0.1806]	[0.0779, 0.2377]
P2	[14.89, 24.33]	[16.02, 24.17]	[19.11, 28.90]
R-square	0.07279	0.08735	0.1142

Table 3.1: Regression model $f(x) = P1 * x + P2$ for various amount of samples (GSE41037) at 95% confidence

The sensitivity of the results regarding the number of CpGs included in the

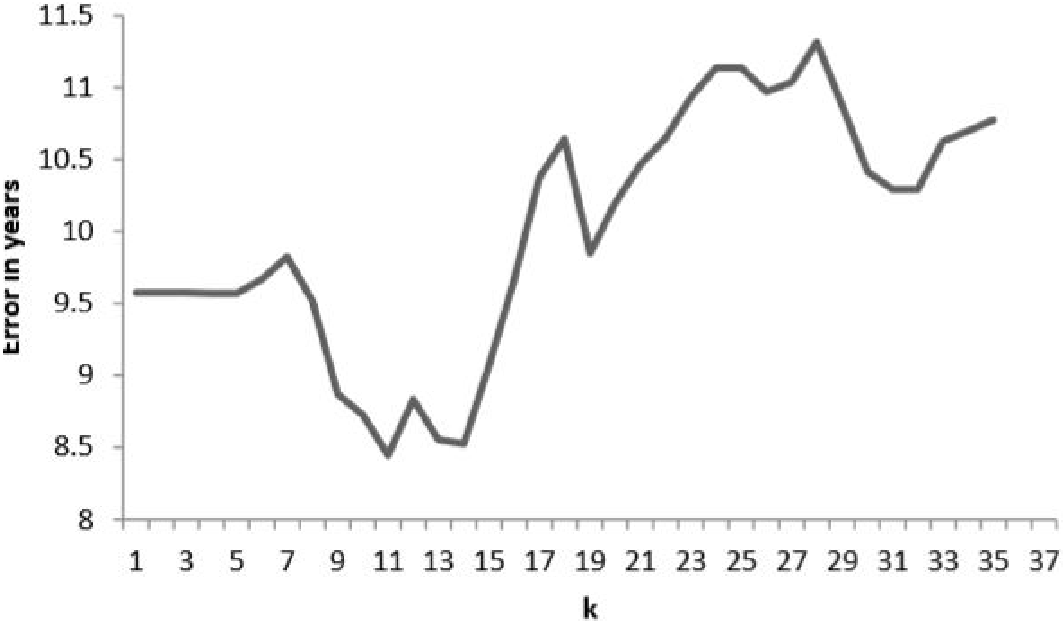


Figure 3.5: Mean error for out of sample values - KNN (GSE4996)

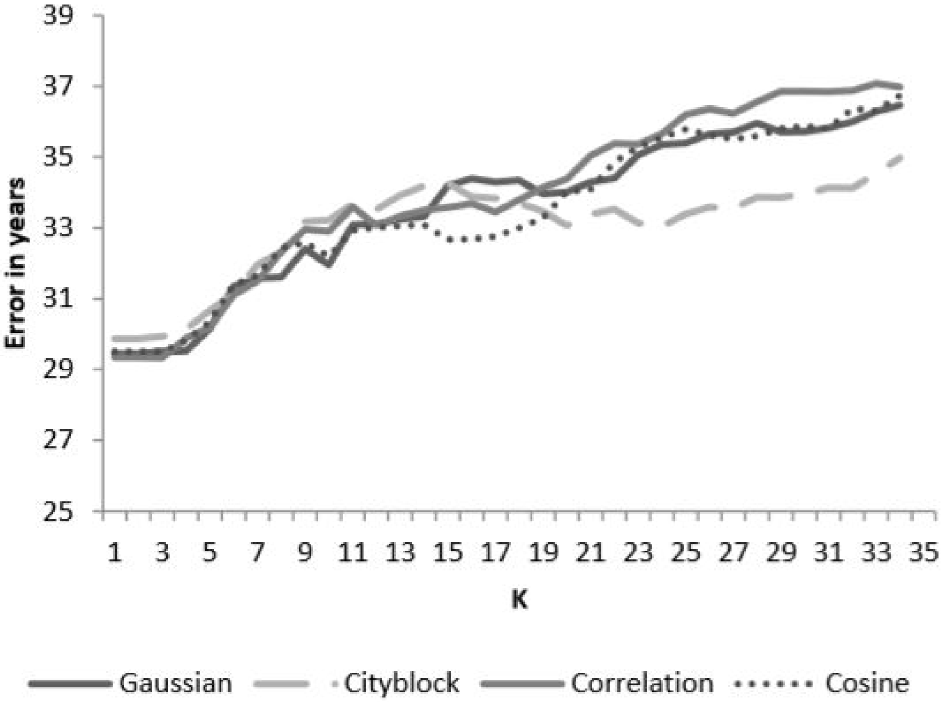


Figure 3.6: Mean error for out of sample values - KNN (GSE41037)

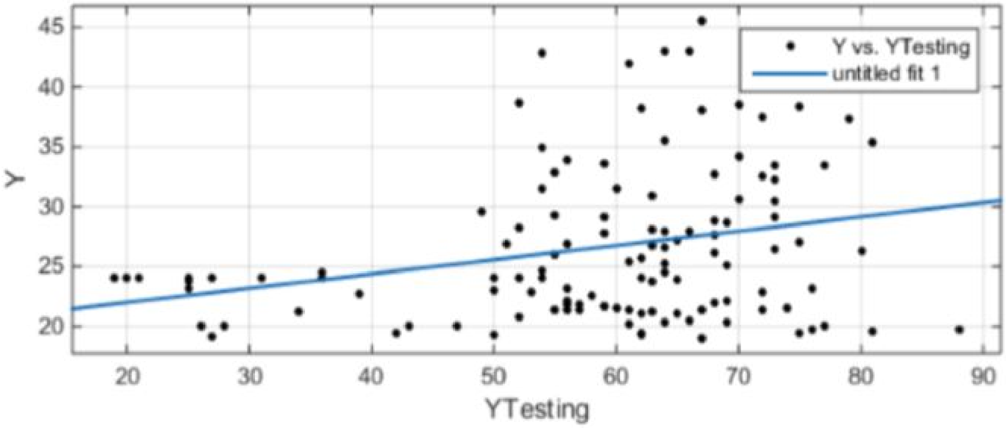


Figure 3.7: Linear regression with 75 cases (GSE41037 subset)

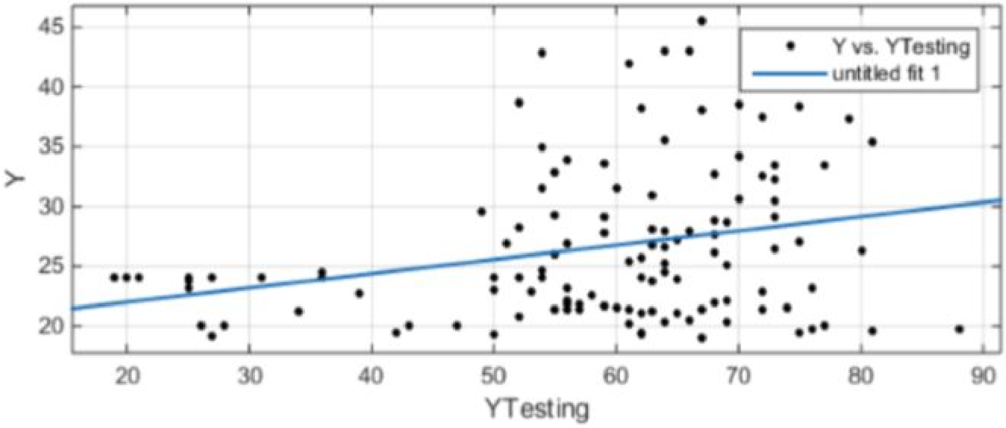


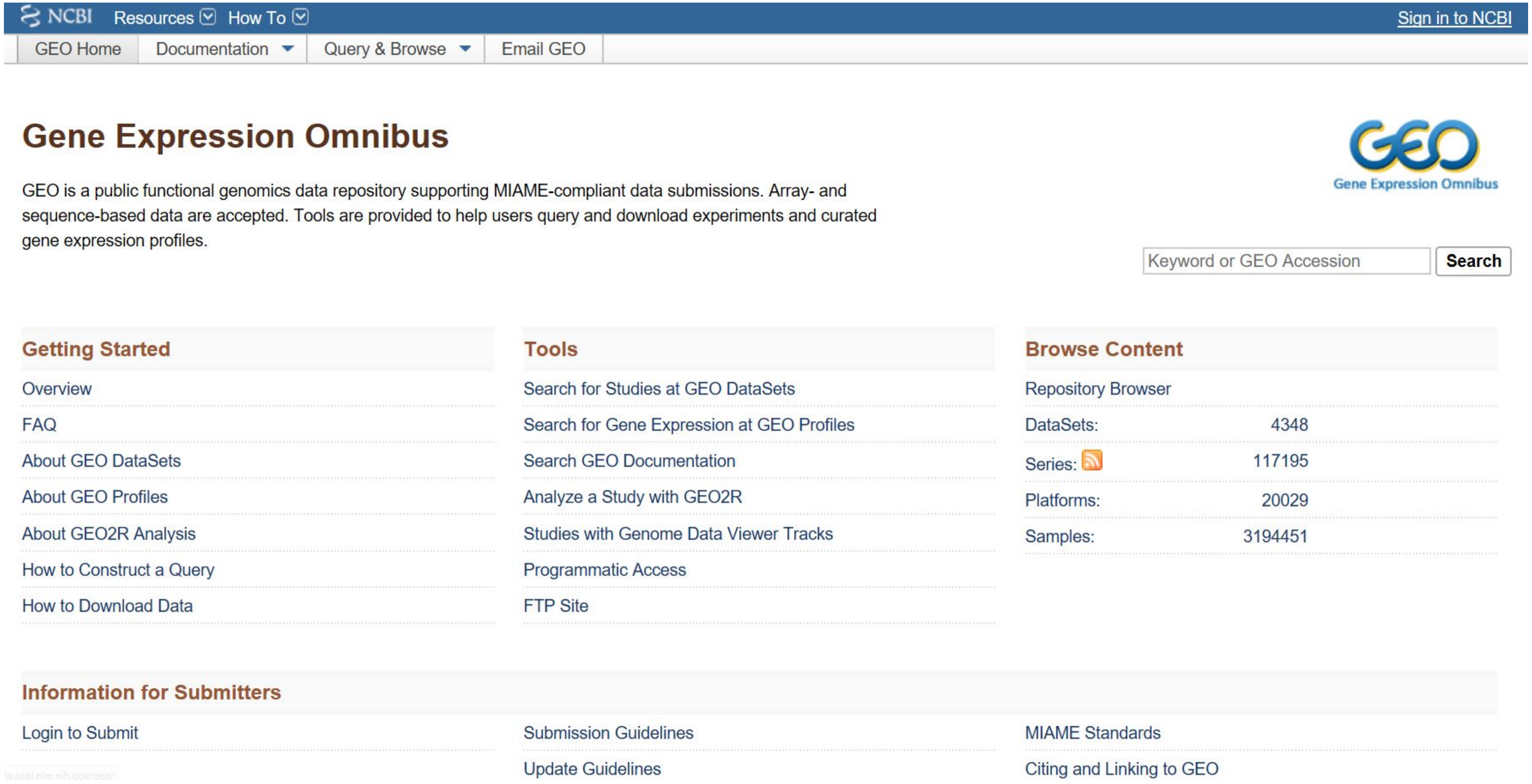
Figure 3.8: Linear regression with 100 cases (GSE41037 subset)

analysis was also tested. The dataset GSE24884 was first analyzed with neural networks including all the CpGs in the dataset. 15 neural networks were performed in order to determine a median value for R and a probability distribution. As expected, due to the small amount of sample this approach did not produced an accurate forecast, with the mean value coming at -0.120. Then 50 % of randomly selected CpGs were deleted from the data set and the process repeated reducing the amount of CpGs by approximately 50% in each step. This was performed iteratively until only approximately 100 CpGs were left. The best forecast obtained in this way was when using approximately 300 CpGs randomly selected, with a mean value of 0.292. This process was repeated 20 times generating each times a different subset of CpGs to be deleted. For all the 20 subsets, except one, the best combination was when using approximately 300 CpGs.

3.1.2 Bayesian neural networks for the optimisation of biological clocks in humans

DNA methylation is related to aging. Some researchers, such as Horvath or Hannum, have managed to create biological clocks using epigenetic data. Both of these authors used Elastic Net methodology to build age predictors that had a high prediction accuracy. In this section, I propose to improve their performance by incorporating an additional step using neural networks trained with Bayesian learning. It will be shown that this approach outperforms the results obtained when using Horvath's method, neural networks directly,or when using other training algorithms,such as Levenberg Marquardt's algorithm. The R-squared value obtained when using our proposed approach in empirical (out of sample) data was 0.934, compared to 0.914 when using a different training algorithm (Levenberg Marquard), or 0.910 when applying the neural network directly (e.g. without reducing the dimensionality of the data). The results were also tested in independent datasets that were not used during the training phase. Our method obtained better R-squared values and RMSE than Horvath's method in the datasets (R-squared values ranging from 0.40 to 0.70). We demonstrate that building an age predictor using a Bayesian based algorithm provides accurate age predictions. This method is implemented in an R function, which is available through a package created for predicting purposes and is applicable to methylation data.This will help to elucidate thereof DNA methylation age in complex diseases or traits related to aging.

In recent years there has been an increased interest in studying the impact of methylation levels on aging. Some researchers, such as [Horvath, 2013, Hannum, 2013], have developed biological clocks that are able to estimate the age of a patient by analyzing the level of methylation of blood cells and also tissues such as brain, breast or colon matter. At a chemical level, methylation is the addition of a group methyl to cytosine base at 5-CPG-3location, see [Caifa, 2004, Patterson, 2011a, McBryan, 2014, Cerchietti, 2017, Yuval, 2018]. CPG represents the link between a Cytosine and a Guanine base by a phospodiester bond. It has been theorised by authors such as [Lim, 2010] that methylation has an important role in regulating gene expression. However, the role of DNA methylation in the process of aging in humans remains unclear. Authors such as [Jones, 2013] have noticed that "certain CPGs sites are highly associated with age, to the extent that prediction models using a small number of these sites can accurately predict the chronological age of donors", referring to biological clocks such as those created by Horvath or Hannum. The impact of methylation on mortality has also been analyzed [Marioni, 2016].[Horvath, 2014] found that methylation is impacted by some environmental and lifestyle factors, such as obesity. In particular, they found that obesity increases aging in the liver. Aging is clearly a complex process with many intertwined factors. For instance, the impact of telomere shortening in aging has been the object of several studies [Takubo, 2014, Tsuji, 2002, Epel, 2004]. There have also been extensive studies regarding the link between methylation and cancer. For instance, [Hashimoto, 2016] used methylation as a biomarker for gastrointestinal tumors and [Pouliot, 2015] used a similar approach for breast cancer. Although the reported prediction accuracy in both Horvath and Hannum's age predictors was high, it can be improved by using better statistical methods. Both methods used Elastic Net (EN) technique, which aims to reduce the dimensionality of data. EN is based on a regularised regression that linearly combines the L1 and L2 penalties of the lasso and ridge methods. We hypothesise that using neural networks (NN) after dimensionality reduction seems a logical step because there is no indication that the level of methylation and the chronological age of the patient should follow a linear relationship. In addition, reducing the dimensionality of the data before applying neural networks is beneficial because it likely decreases the possibility of issues with local minimums, which is a frequently mentioned drawback of neural networks. It will also decrease the computational costs of the calculations required to train it. Therefore, our proposed method combines EN and NN to predict chronological age based on methylation data. In particular, the dimensionality of the data is reduced by doing an EN regression,as proposed in Horvath's paper, and then Bayesian learning is applied. In this section I illustrate that this approach yields results that are on average superior to directly using NN on the data, as well as using EN combined with NN using other training algorithms, such as Levenberg-Marquard's algorithm. The proposed approach is validated using real datasets that were obtained from GEO repository (https://www.ncbi.nlm.nih.gov/geo/), see figure 3.9. This method is implemented in an R function, which available through a package called methylclock. A sample, showing the structure of a methylation file for a patient can be seen in figure 3.10.



Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

Tools

- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- Studies with Genome Data Viewer Tracks
- Programmatic Access
- FTP Site

Browse Content

Repository Browser

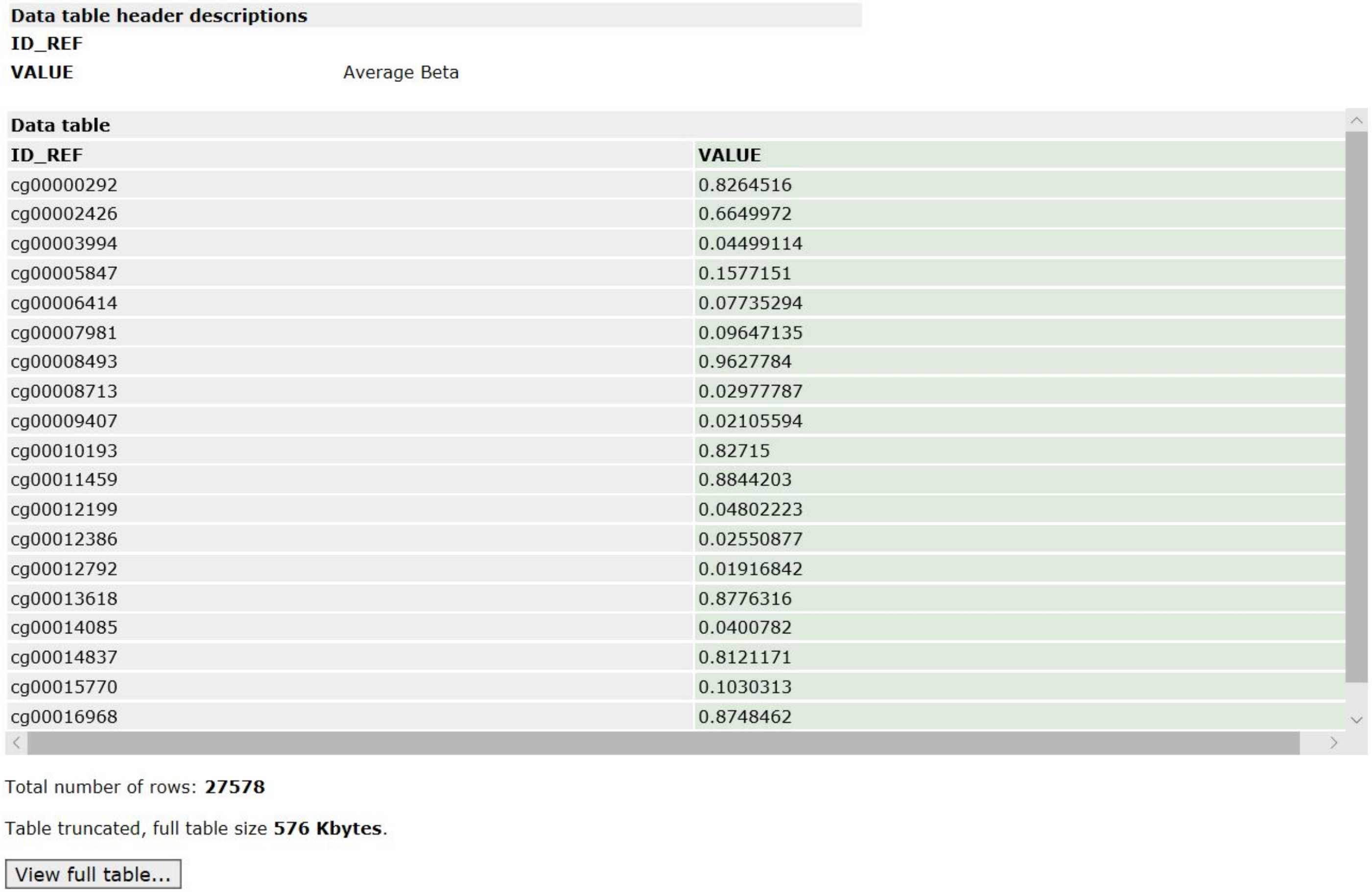
DataSets:	4348
Series:	117195
Platforms:	20029
Samples:	3194451

Information for Submitters

- Login to Submit
- Submission Guidelines
- MIAME Standards
- Update Guidelines
- Citing and Linking to GEO

w.ncbi.nlm.nih.gov/geo/

Figure 3.9: Geo Database



Data table header descriptions

ID_REF	
VALUE	Average Beta

Data table

ID_REF	VALUE
cg00000292	0.8264516
cg00002426	0.6649972
cg00003994	0.04499114
cg00005847	0.1577151
cg00006414	0.07735294
cg00007981	0.09647135
cg00008493	0.9627784
cg00008713	0.02977787
cg00009407	0.02105594
cg00010193	0.82715
cg00011459	0.8844203
cg00012199	0.04802223
cg00012386	0.02550877
cg00012792	0.01916842
cg00013618	0.8776316
cg00014085	0.0400782
cg00014837	0.8121171
cg00015770	0.1030313
cg00016968	0.8748462

Total number of rows: **27578**

Table truncated, full table size **576 Kbytes**.

Sample.JPG

Figure 3.10: Geo Database

Herein, we describe the statistical methods that we used in the two steps performed to predict chronological age using methylation data. The method assumes that input data are CpGs probes containing beta values obtained from methylation DNA arrays (Illumina 27K or 450K).

Elastic net

The elastic net is a robust algorithm, [Hui-Zhang, 2005], for linear regression that has the interesting property of making some of the coefficients equal to zero reducing in practice the amount of independent variables in the model. If we assume that we have a dependent variable (y) and several independent variables (X), then an elastic net regression is done by finding the estimator that minimises this equation [Hui-Zhang, 2005]:

$$\hat{\beta} = argmin\left\{L(a, b, \beta)\right\} \tag{3.5}$$

$$L(a, b, \beta) = |y - X\beta|^2 + b|\beta|^2 + a|\beta| \tag{3.6}$$

By doing this, the dimensionality of the problem is reduced. This helps us to deciding which independent variables to keep in the regression. This is particularly useful when there is a large number of independent variables and it is not clear which ones are relevant for the regression. However, keeping too many independent variables could cause the obtained expression to generalise poorly when new data is used.

Levenberg Marquardt

Levenberg-Marquardt [1963, 1994], "LM" is a commonly used training algorithm in neural networks, [Hagan, 1994, Smaoui, 2003, Bahram, 2003, Basterrech, 2011], that avoids calculating the Hessian matrix and has the goal of minimising a non-linear function. In mathematical terms, the goodness of the estimated $\hat{y}$ values and the actual y values can be described using a chi distributed error that can be computed using the formula proposed by [Gavin, 2011]:

$$\chi^2 = \sum \left\{ \frac{y - \hat{y}}{\sigma_y} \right\} \tag{3.7}$$

With the LM model commonly used as a training algorithm solving this equation [Gavin, 2011]:

$$\left\{ J^T M J + diagonal(J^T M J) \right\} = J^T M(y - \hat{y}) \tag{3.8}$$

Where $\hat{y}$ are the predicted values, J is the Jacobian and M is a diagonal matrix with each diagonal component equal to the inverse of the variance. The LM is a well-tested algorithm with application in fields as diverse as character recognition, as shown by [Badi, 2013], or to model exponential increases of viruses, as shown by [Novella, 1995]. For a more in depth description of the algorithm, we point the reader to the work by [Chen-Yu, 2014].

Bayesian regularisation

Another possible training algorithm for neural networks, which seems to have obtained better results when applied to the case of DNA methylation, is Bayesian regularisation. The purpose of Bayesian regularisation is, as described by [Mackay, 1992], to "minimise a linear combination of squared errors and weights." It is also designed to have good generalisation properties (it should be noted that this is the reference description given by some commercial software as Matlab). The issue of over-fitting is important while analyzing methylation in cells because the large amount of data makes it easy to fall due to the over-fitting issue. Over-fitting is, in simple words, an issue that arises where a neural network matches very closely the output of the training data but then produces poor results when applied to other data, which are not seen by the algorithm. In other words, the network does not generalise well.

It is common practice in this type of problem to try to minimise the sum of the squared errors:

$$\phi_{error} = \sum (y_i - \hat{y}_i)^2 \tag{3.9}$$

In Bayesian regularisation an additional term, [Hagan, 1997], the sum of the squares of the weights ($\phi_{weights}$) is added, with the expression becoming:

$$\psi = a\phi_{error} + b\phi_{weights} \tag{3.10}$$

With the basic algorithm being [Hagan, 1997]:

1. Initialise parameters a and b

2. Calculate ψ

3. Using the approximation:

$$H \approx 2aJ^T J + 2bI \tag{3.11}$$

   Estimate:

$$\delta = N - 2bTr(H)^{-1} \tag{3.12}$$

4. Estimate the value of the parameters:

$$a = \frac{n - \delta}{2\phi_{error}} \tag{3.13}$$

$$b = \frac{\delta}{2\phi_{weights}} \tag{3.14}$$

5. Repeat iteratively (starting after step 2)

For a more detailed explanation please see the original article [Hagan, 1997].

Horvath's age predictor was based on 8,000 samples from different tissues and cell types. Probes of these samples were generated from the Illumina 27K DNA methylation arrays. The age prediction was based on 353 CpGs that were selected using EN. Using these CPGs as the first step, we then trained a Bayesian Neural Network using data from 720 individuals, who were used to study the effect of ageing on methylation (GEO accession number GSE41037, https://www.ncbi.nlm.nih.gov/geo/). Methylation data was obtained from Illumina HumanMethylation27 bead chip that provides methylation levels across approximately 27,000 CpGs measured in blood. We used 15 percent of the data for internal testing purposes. Following standard procedures, these data were not used during the training phase to evaluate how the network generalised to new data and, more importantly, to assess the issue of over-fitting. We then used 8 different GEO datasets measured in different tissues and platforms to assess the external validity of the proposed method as well as to compare model's performance with Horvath's method (table 3.2).

DNA origin	**Platform**	**n**	**age range**	**samples**	**GEO number**
Whole blood	27k	172	(33, 80)	Healthy	GSE58045
Blood	27k	214	(42, 93)	Healthy	GSE19711
Blood	450k	16	(21, 32)	Healthy	GSE65638
Whole blood	450k	43	(47, 59)	Healthy	GSE53128
Blood PBMC	27k	91	(24, 45)	Healthy	GSE37008
Blood CD4+C14	27k	50	(16, 69)	Healthy	GSE20242
Whole blood	450k	231	(34, 72)	Cancer/Control	GSE49996
Whole blood	450k	214	(51, 82)	Alcohol	GSE112987

Table 3.2: GEO datasets used for external validating purposes

We computed two statistics in the different GEO validation sets to evaluate the performance of our age predictors: (1) the correlation between predicted age and chronological age (R-squared); and (2) the Root Mean Square Error (RMSE). We then computed the standard error of the R-squared and RMSE using the formulas provided in [Hagan, 1997], respectively. We meta-analyzed the difference of the R-squared and RMSE between our proposed method and Horvath's approach using both fixed and random effect models implemented in meta R package.

3.2 Cancer

3.2.1 K-nearest neighbours and neural networks as tool to study DNA Methylation, cancer and concentration of information on CpGs

There is an increased focus on applying statistical techniques in cancer research. This is partly due to the large amount of information made public by institutions such as the National Center for Biotechnology Information (NCBI). There is currently a significant amount of human DNA methylation data publically available from such source. DNA methylation is natural process in which a methyl group links with a cytosine in a specified location (carbon 5). This is a process that occurs naturally in humans and other species and has been linked to many biological processes and diseases such as cancer. Changes in methylation levels in tumors have been analyzed in many articles such as [Das-Partha, 2004, Phillips, 2008]. There is an increasing literature linking DNA methylation changes with several types of cancers [Varela-Rey, 2013, Warton, 2016, Davis, 2004, Oakes, 2013]. Given the efforts that have been made trying to study the relationship with DNA methylation and cancer it seemed reasonable to try to apply algorithm such as Neural Networks or K-Nearest Neighbors to DNA methylation data. It will be shown that using this techniques it is possible to classify, within a reasonable level of error, patients into two categories. One representing cancer patients and one representing healthy control patients.

A frequently used tool for classification problems is the K-Nearest Neighbors algorithm. The K-Nearest Neighbors algorithm tries to find similarities between the case that is currently analyzed, such as a new sample, and a dataset of already categorized samples. Similarly to neural networks this technique has been successfully applied in many fields [Khamis-Hassan, 2014, Parvin-Hamid, 2010]. There are also some interesting articles applying this technique to cancer classification. A good example would be [Chaoli, 2012]. The authors in [Chaoli, 2012] used this technique to classify metastasis in a gastric cancer analysis. As previously mentioned the algorithm is relatively simple. The basic idea is comparing the distance between the data and trying to group the samples accordingly. There are however several factors to take into account. For instance, there are multiple distance measures. The practitioner needs also to select the amount of neighbors to be used for comparing purposes. There is no general rule to select these attributes. The distance metrics used can be found in equations 3.15, 3.16, 3.17 and 3.18. All the calculations were performed in Matlab and the nomenclature used for defining the distance metrics is also Matlab nomenclature.

1. Euclidean:

$$(\varphi_i - \phi_j)(\varphi_i - \phi_j)' \tag{3.15}$$

2. Cosine:

$$\frac{(\varphi_i\varphi_i'\phi_j\phi_j')^{1/2} - \varphi_i\phi_j')}{(\varphi_i\varphi_i'\phi_j\phi_j')^{1/2}} \tag{3.16}$$

3. Seuclidean:

$$(\varphi_i - \phi_j)\Phi^{-1}(\varphi_i - \phi_j)' \tag{3.17}$$

4. Hamming

$$\varphi_i = \phi_j \rightarrow 0; \varphi_i \neq \phi_j \rightarrow 1 \tag{3.18}$$

5. Jaccard

$$\frac{max(\varphi_i, \phi_j) - min(\varphi_i, \phi_j)}{max(\varphi_i, \phi_j)} \tag{3.19}$$

6. Minkowski

$$(\sum |\varphi_{il} - \phi_{jl}|^q)^{1/q} \tag{3.20}$$

7. Chebychev

$$max(|\varphi_{il} - \phi_{jl}|) \tag{3.21}$$

Neural Networks are a set of algorithms frequently used for forecasting, classification and clustering purposes. Perhaps the most typical approach is using them for supervised learning. In other words, trying to obtain an output, given a series of inputs as close as possible to the actual output. This is a tried and tested technique in many fields. A neural networks is composed by artificial neurons which in this context should be understood just as a mathematical formula giving an output when it receives an input. The output is typically limited to values between zero and one but there are other potential options. There are some weights associated with these neurons. The network learns by modifying the weights until the obtained output is close enough to the actual target. There are several articles in the literature analyzing the issue of cancer detection using neural networks. For instance, [Ganesan, 2010] applied this technique to cancer diagnosis using demographic data. In an interesting article [Menendez-Alvarez, 2012] neural networks were successfully used for breast cancer screening. There are also some articles applying this technique to methylation data, such as [Copede, 2015]. The author used this technique in a study of colon cancer. It will be shown that neural networks can be applied, for cancer detection purposes, to two different types of cancers. It will be also shown that subdividing the data into smaller dataset neural networks can be also used as a tool to see how the cancer information is spread among different CpGs.

As previously mentioned analyzing DNA methylation data for cancer identification purposes is a growing field of research. It is important not only identifying what algorithms are successful on differentiating between cancer and non-cancer samples but also to have a better understanding of how the information is spread among CpGs. In this section two algorithms, Neural Networks and K-Nearest Neighbors, are applied to two different types of cancer. Both techniques seem to provide relatively accurate classifications. When using the K-Nearest Neighbors approach it was found that the distance metric used had a substantial impact on the error rate with for instance the Minkowski and Chebychev metrics giving accurate forecast over a relatively large amount of configurations. While more testing is required the results would seem to support the idea that for the two types of cancer analyzed the cancer information seemed to be spread among a large number of CpGs with forecasting remaining accurate even when using a small randomly selected subsets of CpGs. This result seemed to be consistent when using any of the two previously mentioned algorithms with any of the two datasets analyzed.

Two datasets obtained for the GEO Database were analyzed. The first data set comes from a prostate cancer article [Kobayashi, 2011] and has the GEO Database code GSE26126. It contains 193 samples. There are 83 healthy tissue samples in this data set. There are also 12 cultured cell samples. The methylation data was obtained using an Illumina Infinium 27K Human Methylation Beadchip. Using this machine it is possible to obtain methylation information for 27,578 CpGs. The second dataset analyzed was obtained from a carcinoma article [Warton, 2016]. It is a slightly smaller dataset containing 120 samples. The Geo Database code is GSE57956. The sample include carcinomas and adjacent non-tumor samples. All these databases are publically available at GEO Database which is a database supervised by the National Center for Biotechnology Information in the U.S..

Two techniques were used to analyze the data: 1) Neural Networks and 2) K-Nearest Neighbors. Both of these techniques are flexible tools frequently used for classification purposes in many fields. Both techniques are relatively flexible in the sense that no knowledge of the underlying process is assumed and tend to provide, depending on the specific application, acceptable results.

A neural network with two layers (one hidden) and 10 neurons was used as a classification algorithm. The output data using to classify the neural network was a binary state with the value 1 identifying a cancer patient and the value 0 identifying a healthy individual. 15% of the data were used for testing purposes. These data were not seen by the neural network during training. The out-of-sample (testing data) error was calculated for the neural network. This process was performed 30 times in order to obtain a probability distribution. One of the objectives is trying to analyze how spread the cancer information is among the CpGs. In an attempt to do this the data set was sliced into several subsets. One subset contained the entire amount of CpGs information (27,578) and the following subset dropped 10% of the CpGs data at the time. The selection of which CpGs to drop for every new subset was done randomly. This was done until having only 10% of the data. Smaller subsets containing 5%, 1%, 0.1% and 0.01% of the original data were also obtained and their out-of-sample classification error determined (see tables 3.3 and 3.4). The idea behind this approach is trying to see which amount of CpGs data is enough to have a reasonably good indicator about the presence of cancer.

Subset containing (%) **of original data**	**Mean**	**Median**	**St. Dev**
100%	0.1000	0.0862	0.0597
90%	0.1000	0.1034	0.0507
80%	0.1345	0.1032	0.1006
70%	0.1000	0.0690	0.0523
60%	0.1092	0.1034	0.0559
50%	0.1333	0.1379	0.0606
40%	0.1057	0.1034	0.0404
30%	0.1322	0.1379	0.0647
20%	0.1264	0.1207	0.0692
10%	0.1103	0.1034	0.0539
5%	0.1126	0.1207	0.0620
1%	0.1287	0.1207	0.0520
0.1%	0.1759	0.1724	0.0709
0.01%	0.4115	0.4138	0.0859

Table 3.3: Out-of-sample error (neural network) - GSE26126

The out-of-sample errors obtained for each data subset were formally compared to the errors obtained using the entire dataset using a Wilcoxon test (see 3.5 and 3.6). A Kruskal-Wallis test was also used. The Kruskal-Wallis test compares all the data sets simultaneously and determines if they come from the same distribution. This was done iteratively for each data set. In other words, in the first step all the subsets were analyzed using the Kruskal-Wallis test. Then the smaller subset, containing only 0.01% of the initial CpGs information, was excluded and

Subset containing (%) **of original data**	**Mean**	**Median**	**St. Dev**
100%	0.0954	0.0690	0.0625
90%	0.1172	0.1034	0.0484
80%	0.1093	0.1034	0.0534
70%	0.1057	0.1032	0.0593
60%	0.1366	0.1034	0.0769
50%	0.1172	0.1034	0.0638
40%	0.1080	0.1034	0.0494
30%	0.1391	0.1379	0.0650
20%	0.1183	0.1034	0.0583
10%	0.1437	0.1379	0.1195
5%	0.1149	0.1034	0.0499
1%	0.1828	0.1379	0.1655
0.1%	0.1713	0.1724	0.0776
0.01%	0.4241	0.4483	0.1029

Table 3.4: Out-of-sample error (neural network) - GSE57956

the test was performed again. Then the 0.1% subset was excluded and so on. This approach was repeated until the hypothesis that the error distributions obtained from the neural networks come from the same distribution for all the subsets cannot be rejected is obtained (table 3.7). As previously mentioned, the lower limit for the detection of cancer would depend on two factors. The first factor is the accuracy of the algorithm employed for the detection. The second factor is the data. Regardless of the accuracy of the algorithm the classification cannot be better than the information contained on the underlying data. The algorithm can however fail to produce a good classification of the patients (cancer and non-cancer) even if there is enough information in the underlying data to successfully perform such classification.

Subset containing (%) of original CpGs	**p**	**h**
90%	0.85580	0
80%	0.10500	0
70%	0.98160	0
60%	0.47960	0
50%	0.02600	1
40%	0.39040	0
30%	0.05530	0
20%	0.09530	0
10%	0.29300	0
5%	0.30030	0
1%	0.02370	1
0.1%	0.00016	1
0.01%	2.74E-11	1

Table 3.5: Wilcoxon test comparing the out-of-sample errors (entire data compared to subsets). GSE26126

The K-Nearest Neighbors technique was used to classify the data according to the presence of cancer. 15% of the data were excluded for the training process

Subset containing (%) **of original CpGs**	**p**	**h**
90%	0.06110	0
80%	0.27010	0
70%	0.51730	0
60%	0.02060	1
50%	0.27120	0
40%	0.32240	0
30%	0.00650	1
20%	0.11010	0
10%	0.03250	1
5%	0.08620	0
1%	0.00290	1
0.1%	0.00012	1
0.01%	3.31E-11	1

Table 3.6: Wilcoxon test comparing the out-of-sample errors (entire data compared to subsets). GSE57986

Data	**GSE26126**	**GSE57956**
All data sets	3.18705e-18	3.268e-17
Excluding 0.01% dataset	0.0005	0.0011
Excluding 0.01% and 0.1% dataset	0.1331	0.0607

Table 3.7: p-values obtained from the Kruskal-Wallis test from the two datasets

and used only to estimate the out-of-sample error rate. Seven different types of distance metrics were analyzed. Those seven metrics were: Euclidean, Seuclidean, Cosine, Hamming, Jaccard, Minkowski and Chebychev. Another factor to take into consideration when using this technique is the number of neighbours. A range from 1 to 160 was used. After than level, for both datasets the classifications do not appear to be accurate. In order to try to understand the distribution of information among the CpGs several subsets of data were analyzed. The first subset contains all the CpGs data. The second subset contained only 90% of the CpGs, selected randomly. This process was repeated until having only 10% of the original data. Then subsets with 5%, 1%, 0.1% and 0.01% of the data were also analyzed. The same approach was followed when using neural networks. The results from all the simulations can be found in figure 4.4 and in the appendix. As an example please see figure 3.11, a sample result for the GSE26126 dataset containing all the data. It should be noted that as the number of data decreases there would be a point in which the amount of information is just not enough for the algorithm to work properly, regardless if the CpGs contain cancer information or otherwise.

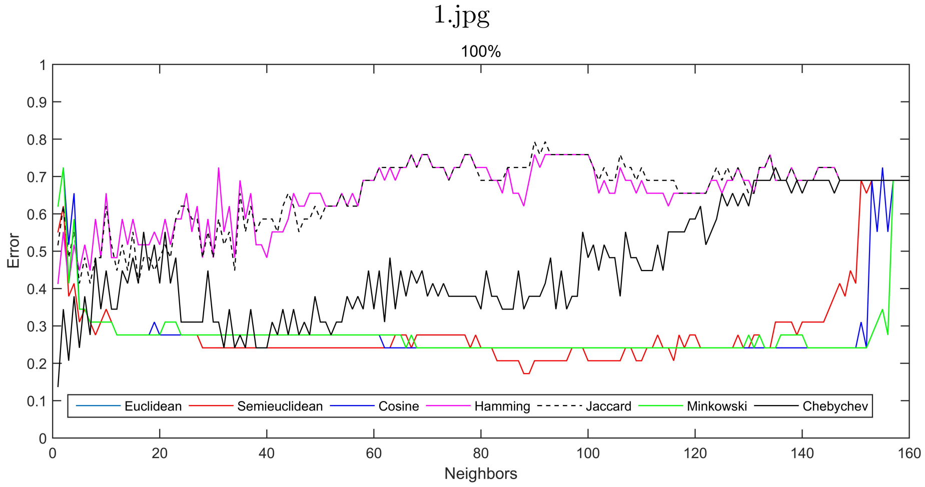


Figure 3.11: K-Nearest Neighbor. GSE 26126 (entire set)

3.2.2 Cancer detection using support vector machines trained with linear kernels

DNA methylation remains a very active area of research due to its suspected effect in areas as diverse as development [Baxter, 2012], aging [Horvath, 2013] and cancer [Charles, 2017]. DNA methylation remains an area not well understood, likely due to its very high level of complexity, and it seems intertwined with many biological processes. As a biological marker DNA methylation has proven a very useful technique and it is likely to generate a large amount of research in years to come. Technological advancements have made available an increasing amount of methylation data from patients that undergo procedures or that volunteer for research. This increase in data availability has put pressure to developed better and more efficient statistical models to try to understand these processes. This increase in data availability is almost certain to continue in the future. In this section I attempt to utilize a well know statistical tool called Support Vector Machines ("SVM") to the task of differentiating healthy tissue from tissue with cancer using methylation data. SVMs are a general statistical tool that can, and has, been applied to a multitude of different problems. It is likely that in the near future SVMs will continue finding new areas of application as the amount of data created in many scientific and engineering disciplines increases and simultaneously computing power, which allows such enormous amount of data to be processed, also continues to increase. SVM use the concept of separating data into the different sides of a hyperplane in order to categorize such data. It is a remarkably flexible technique and of relatively simple use. A SVM needs, in the context of this section, the methylation levels for each CpGs, which is a number ranging from 0 to 1 and a binary identification, defining if the sample comes from a tissue with cancer or from a healthy tissue. Currently is possible to obtain thousands of CpGs methylation data quickly from a patient sample using relatively affordable techniques. This creates a mismatch between the number of samples in studies, typically from a few dozens to a few hundreds, and the thousands of data points available for each patient. In this context SVM attempt to categorize the methylation for each patient into two categories: 1) caner and 2) no cancer. CpGs are just a bond between two bases, a Cytosine and a Guanine and they have proven rather important in several biological processes receiving a considerable amount of interest by researchers. Having a quantifiable indicator of cancer could be useful for the doctors making diagnosis as well as a potential tool for confirmation of such diagnosis. It will be shown that training the SVM with a linear kernel for the three tissues analyzed (liver, lung and cervix) produced more accurate results than using other kernels, such as polynomial or Gaussian. These results were rather consistent among the three data sets with direct tissue data (no blood samples). The approach of using an SVM trained with a linear kernel seems also to produce results more accurate than using a simple backpropagation neural network trained with 10 neurons. It will also be shown that the results are less accurate when the analysis is performed on blood samples, rather than using directly methylation data from lung, liver or cervix. The results regarding what type of kernel to use in this case are less conclusive. This last point is likely a good area for further research.

In this section a brief description of support vector machines is presented, for readers interested in a more mathematically detailed explanation of SVM please see [Zhouyan, 2012, Shawe, 2000, Ashfaq, 2013]. These are all very good articles and they go into details into formal mathematical issues. The mathematical formalism for support vector machines is not particularly simple and getting into its details is outside of the scope of this dissertation, which focuses on applying such techniques to the specific case of detecting cancer though SVM using methylation data as an input. Plainly speaking a SVM tries to create a boundary (hyperplane) between the two sets of data which is trying to classify. This boundary should be as far away from the data as possible while containing all of them. This clearly leads to a Lagrange multiplier type of situation in which a function needs to be maximized while certain constraints must be kept [Baxter, 2012]. There has been a lot of interest both theoretically [Sung-Hoon, 2007] as well as regarding practical applications of SVM [Dushicang, 2015]. There are some articles in the literature applying this technique for imaging processing (radiology). For instance, [Edriss, 2012] applied this technique to breast cancer data and [Kohad-Rashmee, 2014] applied it to lung cancer data. Imaging processing is clearly a natural candidate for application of SVM as it removes, at least to some degree, the subjectivity of the radiologist when examining MRI images to determine the presence of cancer. This process clearly depends heavily on the experience of the radiologist with some degree of subjectivity when analyzing unclear images or cancer in early stages. This is an area in which a great deal of automation could be applied and in fact it is currently a vibrant area of research. Perhaps less attention has received the application of support vector machines using methylation data as inputs. One interesting article in this regard is [List, 2014], which successfully applied the technique to breast cancer. In this article the input data used were not only methylation levels but also gene expression data. In another interesting article [Hosseinzadeh, 2014] used neural networks as a classification for differentiation between healthy tissue and lung cancer. The literature in this regard is expanding rapidly due to the clear practical applications of these techniques and the ever increasing amount of data available.

Support vector machines are used for determining is cancer is present in lung, liver and cervix tissue using multiple kernels. The results indicate that linear kernel in this regard seems to be a better approach than using polynomial or Gaussian kernels. It was also found that using support vector machines trained with a linear kernel seems to also produce more accurate results than using a backpropagation neural network with 10 neurons. The accuracy of classification decreases when methylation in blood samples is analyzed, rather than direct tissue samples, to determining the presence of cancer.

In this section I utilize a well know statistical tool called Support Vector Machines ("SVM") to the task of differentiating healthy tissue from tissue with cancer using methylation data. SVMs are a general statistical tool that can, and has, been applied to a multitude of different problems. It is likely that in the near future SVMs will continue finding new areas of application as the amount of data created in many scientific and engineering disciplines increases and simultaneously computing power, which allows such enormous amount of data to be processed, also continues to increase. SVM use the concept of separating data into the different sides of a hyperplane in order to categorize such data. It is a remarkably flexible technique and of relatively simple use. A SVM needs, in the context of this paper, the methylation levels for each CpGs, which is a number ranging from 0 to 1 and a binary identification, defining if the sample comes from a tissue with cancer or from a healthy tissue. Currently is possible to obtain thousands of CpGs methylation data quickly from a patient sample using relatively affordable techniques. This creates a mismatch between the number of samples in studies, typically from a few dozens to a few hundreds, and the thousands of data points available for each patient. In this context SVM attempt to categorize the methylation for each patient into two categories: 1) caner and 2) no cancer. CpGs are just a bond between two bases, a Cytosine and a Guanine and they have proven rather important in several biological processes receiving a considerable amount of interest by researchers. Having a quantifiable indicator of cancer could be useful for the doctors making diagnosis as well as a potential tool for confirmation of such diagnosis. It will be shown that training the SVM with a linear kernel for the three tissues analyzed (liver, lung and cervix) produced more accurate results than using other kernels, such as polynomial or Gaussian. These results were rather consistent among the three data sets with direct tissue data (no blood samples). The approach of using an SVM trained with a linear kernel seems also to produce results more accurate than using a simple backpropagation neural network trained with 10 neurons. It will also be shown that the results are less accurate when the analysis is performed on blood samples, rather than using directly methylation data from lung, liver or cervix. The results regarding what type of kernel to use in this case are less conclusive. This last point is likely a good area for further research.

All the data used are publically available in the GEO database and come from other research reports. There are the dataset containing methylation sample from cancer and control cases. The first data set contains cases with liver cancer (GSE57956) and comes from [Mah, 2014] article. There are 120 samples. The second dataset is from a lung cancer study [Lenka, 2017] contains 88 cases and has the GEO data base code (GSE49996). Half of the samples (44) are from lung tissue with cancer and the other half from healthy lung tissue. This dataset is from a cervical cancer article [Zhuang, 2012] and has the accession code (GSE30759) in the Geo Database. These are the three datasets containing methylation information from organs. A fourth dataset was used, in this case, rather than having sample from organs the methylation data was extracted from blood samples. This information was obtained from bladder cancer research published by [Scott, 2015] with the GEO database code (GSE50409), 120 samples. All the dataset contain DNA methylation information of patients obtained with the Illumina 27K. There are in excess of 27,000 CpGs methylation data points for each patient present in the dataset as well as an indicator representing if the data comes from a cancer sample or otherwise. All the data used is publically available and obtained from the Geo Database (www.ncbi.nim.nih.gov). The algorithm used to detect cancer was a support vector machine, trained with three different kernels: liner, polynomial or Gaussian. The objective is to obtain the smallest, out of sample, classification error possible. The three previously mentioned kernels can be defines as follows (equations 3.22, 3.23 and 3.24):

1. Linear:
$$kernel = \zeta_1 * \zeta_2 \tag{3.22}$$

2. polynomial:
$$kernel = (\zeta_1 * \zeta_2 + 1)^{\alpha} \tag{3.23}$$
where a is the degree of the expression

3. Gaussian
$$kernel = e^{-|\zeta_1 - \zeta_2|^2} \tag{3.24}$$

Deciding which kernel to use is of clear importance and can potentially have a substantial impact on the accuracy of the data classification. This decision, of what type of kernel to use, depends on the specific application. It is not easy, in principle, to decide a priori without actually comparing the results of different kernels which one to use. As an additional step and comparison purposes the results from the SVM were also compared with the results from a simple neural network with one hidden layer, 10 neurons and trained using backpropagation. The same process was applied for all the four data sets, regardless if the methylation data came from organs of from blood samples. 100 simulations were performed on each case to obtain a probability distribution. Then a Wilcoxon test was performed comparing the results obtained using SVM, with linear, polynomial and Gaussian kernels, as well as with neural networks. All the calculations were performed using the commercially available software package Matlab.

3.3 Diabetes

3.3.1 Diabetes detection using Bayesian neural networks and SNPs

It is possible to follow an approach similar to the one deployed for the analysis of cancer detection using DNA methylation data but applied instead to the detection of type two diabetes using SNPs. Neural networks seem to be an ideal candidate for this classification problem as there would not appears to be any reason sustaining the hypothesis that the relationships between using SNPs as biomarkers and a patient having type 2 diabetes being a linear one. Furthermore, there could be a very complex underlying process linking SNPs with diabetes that could be rather complex to model with traditional linear models.

The diabetes training set was obtained from Genetic Epidemiology Research on Aging (GERA) containing a large initial amount of SNPs. In this case every SNPs was analyzed doing a single association study. GWAS was run filtering out all the SNPs that did not meet a predefined threshold [Juan-Gonzalez, 2019]. Only those SNPs with p-values smaller r equal to 0.01 were included in the analysis. The final amount of SNPs was 8,574 which is very significant reduction from the original dataset. This analysis was performed using the function snp.rhs.tests in the GWAS package in R.

As there is no widely accepted model that describes in simple but accurate terms the relationship between SNPs and diabetes it seems reason to follow a highly non-linear approach to model this potentially complex process as for instance neural networks. When applying neural networks to a complex classification problem such as this one there are many parameters to take into account, such as for instance but structure of neural network to use, what type of neuron, how many neurons, what type of learning algorithm to use and how to divide the data between the training and testing phases.

A neural network with 150 neurons with Bayesian learning was used as a supervised learning algorithm to distinguish between patients suffering from type 2 diabetes and healthy individuals. The data consisted of three different datasets D_1, D_2 and D_3 with 8,574 SNPs per individual All the training data came from D_1. 20 times cross validation was performed in an attempt to achieve the best generalization power possible and avoid the overfitting issue. Out of sample validation was performed for D_1, D_2 and D_3. The methodological steps were as follows:

1. Define the structure of the neural network (NN)
2. Define the learning algorithm (Bayesian) ($BNN_Learning$)
3. Define the maximum classification acceptable error of the neural network
4. Define the size of neural network (NN).
5. Train the Bayesian neural network (BNN) with the training algorithm ($BNN_{Learning}$) and cross validation
6. Repeat step 6 until the out of sample error using the testing data $D_{1-testing}$ is smaller or equal than the error defined on step 3. If there is no further improvement change the size of the neural network (back to step 5)
7. If out-of-sample error using testing subset $D_{1-testing}$ is equal or smaller than the maximum acceptable error then chose that Bayesian neural network (BNN) as a valid one and store it in memory.
8. Estimate the out-of-sample error using external datasets (D_2 and D_3) as validation sets.

All the analysis was performed using Matlab on a small computer luster composed of five CORE i5 8^{th} generation computers. The external validation datasets D_2 and D_3 were obtained from public available sources. D_2 was obtained from Gene Environment-Association Studies "GENEVA" while D_3 was obtained from Finland-United States Investigation of NIDDM Genetics "FUSION".

Chapter 4

Results

The results obtained in this dissertation are based on the results in two published articles in journal plus an additional article that it is undergoing its second review by a journal. The articles are as follows:

- Gerardo Alfonso, Juan R Gonzalez. Bayesian neural networks for the optimisation of biological clocks in humans. *Royal Society Open Science*. Undergoing second review by journal

- Gerardo Alfonso. Age estimation using neural networks and DNA methylation levels. International *Journal of Genetic engineering and biotechnology*. Volume 6. Number 1. 2017. 1-12.

- Gerardo Alfonso. Cancer detection using vector machines trained with linear kernels. *International Journal of Science and Research*. 2017. Volume 6. Issue 7.

The method proposed in the first paper is implemented in an R/Bioconductor package (under evaluation) whose developmental version is available at: `https://github.com/isglobal-brge/methylclock/`.

Another paper is being written which cover the use of our proposed approach on genomic data and diabetes:

- Gerardo Alfonso, Jonatan Gonzalez-Rodriguez, Juan R Gonzalez. Bayesian neural networks accurate predict diabetes risk based on genomic data. *In preparation.*

4.1 Aging

4.1.1 Bayesian neural networks for the optimisation of biological clocks in humans

We first applied NN to the data without reducing the dimensionality using EN (e.g. direct neural network approach). The network was trained 100 times using LM and another 100 times using Bayesian regularisation in both cases the results were less than optimal with substantial dispersion in out-of-sample values obtained for the R-squared of the regression comparing the predicted age estimates with the chronological age of the patients with the network in several cases, stopping after 1,000 iterations without reaching the target error. The input for the network in both cases was the methylation levels of the 27,000 CpGs, with no further transformation. The target output was the age of the patients. The network had two layers (only one hidden) and had 10 neurons in the hidden layer. The low R-squared values obtained from the simulations indicate that the model is not capturing too much variability (median: 0.34, range: 0.22, 0.83). These results suggests that applying a neural network directly to the data is not the best approach because the variance that is explained by this approach is too modest. This is likely to happen because of the issue of local minima in neural networks, which are of particular importance when, as in this the case, there are a large number of input data (CPGs) and a smaller number of samples, see [Bo-Liu, 2017]. Reducing the dimensionality of the data can substantially increase the accuracy of the forecasts.

Our two steps approach aims to overcome this difficulty. We first reduce the dimensionality of the data by using EN and then apply a better predictor methodology (e.g. NN). The EN step is performed by using the CpGs obtained in [Horvath, 2013] that were obtained after regressing chronological and age on about 27,000 CpGs from 8,000 individuals. The CPGs selected in that paper have proven to be useful for age predicting purposes. Consequently, the dimensionality of the problem is reduced from 27,000 variables per individual to 353 variables. The methylation data for these CpGs were used as the input for NN. The NN chosen has two layers with 10 neurons in the hidden layer. The neural networks were then trained using both the Levenberg-Marquardt algorithm and Bayesian regularisation. The results of a neural network will change every time that it is trained because we used different initialisation parameters to avoid sub-optimal local convergence. We run 100 simulations for LM algorithm and another 100 for Bayesian regularisation. Figure 4.1 shows the distribution of R-squared values of each simulation for the two different training methods.

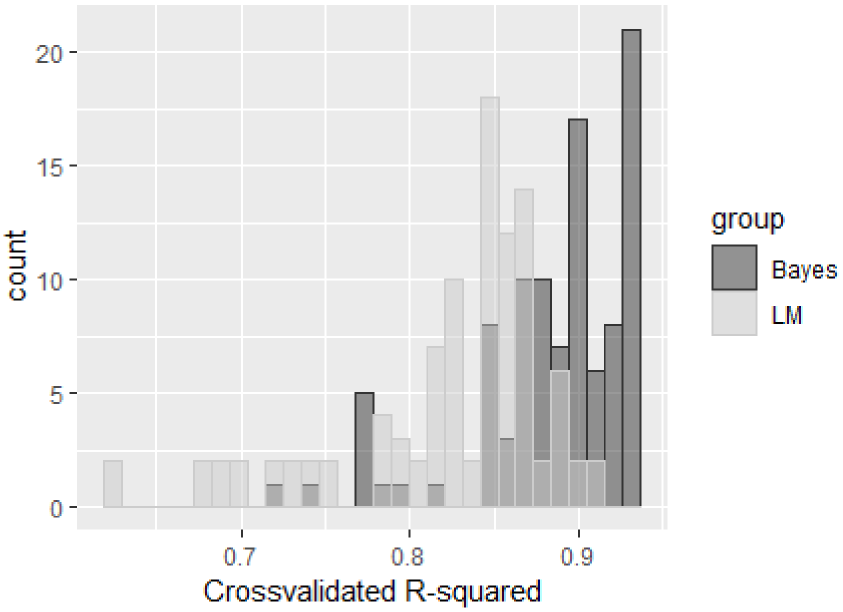


Figure 4.1: R-squared histograms for LM and Bayesian learning methods of simulated datasets.

We observed that Bayesian training clearly outperforms the LM method since the median R-squared for LM is 0.85 (range: 0.63 to 0.92) while for Bayesian training is 0.90 (range: 0.71 to 0.93) (table 4.1). These observed differences are statistically significant after applying Wilcoxon test ($p = 6.4 \times 10^{-53}$). The same conclusion is obtained when comparing our two stage proposed method (EN + NN with Bayesian training) with one based on a Bayesian NN applied to the entire dataset (Wilcoxon test, $p = 6.3 \times 10^{-28}$). This indicates that reducing the dimensionality actually (statistically significantly) improved the forecasts.

	R-squared	
Algorithm	**Median**	**Range**
LM	0.8492	[0.6274, 0.9137]
Bayesian	0.900	[0.7152, 0.9340]

Table 4.1: Median and range of R-squared values obtained from two Neural Networks trained using two methods different methods (Levenberg-Marquardt and Bayesian) after dimension reduction of the dataset GSE58045 in 100 simulated datasets.

We applied our proposed method to the internal test dataset (15% out-of-sample of GEO number GSE41027 not using when creating the predictive model, n=108). The regression of the predicted values against the chronological age can be seen in figure 4.2. We observe a good performance with an R-squared equal to 0.94, being 0.97 in the training data (85% of GSE41027 dataset, n=612), which obviously over-estimates the model's predictive value.

Bayesian neural network implementation

The Bayesian neural network was trained using Matlab. The first step was to create a neural network structure of two layers, one of which is hidden. The hidden layer has 10 neurons. This structure generated better results than more complex neural networks with larger amount of neurons or of hidden layers. Some of the more complex networks analyzed generated good results for the databases analyzed but generalised poorly when handling new databases. The database

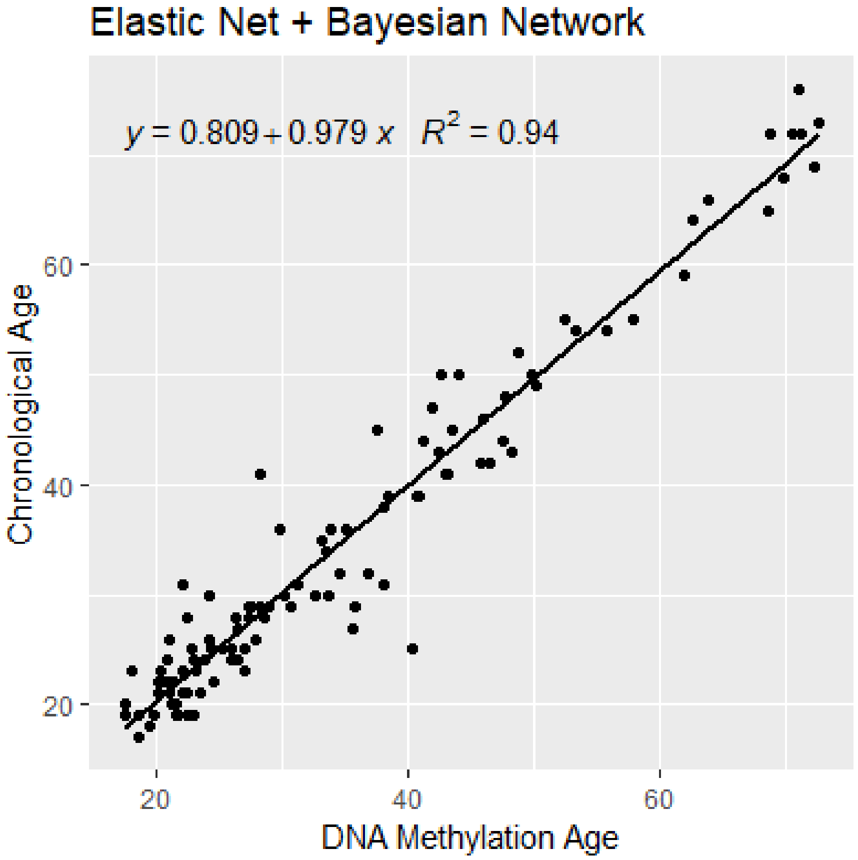


Figure 4.2: Regression of the predicted age against the chronological age using the internal test dataset (15% out-of-sample of GSE41037).

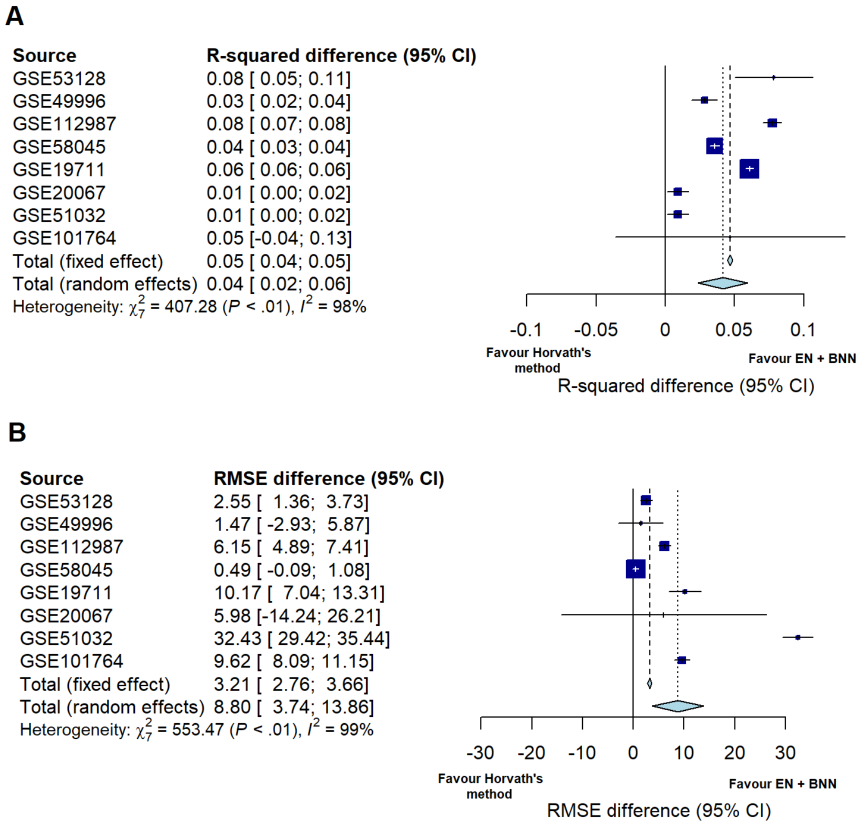
GSE41037 was used for training purposes with 15 percent of the data kept aside for testing purposes. The training algorithm used was Bayesian regularisation, using the package `trainbr` in Matlab. A ten times cross validation approach was followed for training purposes. Consequently, a net structure was generated in Matlab. For convenience that net structure was transformed into a function using the Matlab function `genFunction`. This function takes as an input a matrix of values in which each column represents one patient and each row represents the methylation value for each CPG, as previously mentioned we used the CPGs obtained by [Horvath, 2013]. The function then generates a vector output that is the obtained value for the age of each patient is then compared to the chronological age of the patient using simple linear regression. The function was then transformed into C++ code using the Matlab application "Matlab Coder". An R function that calls this C++ code was created to allow users predicting age using DNA methylation data that is available at `https://github.com/isglobal-brge/methylclock`.

Validation

Our proposed method was tested using 8 independent GEO datasets. The results obtained from our two stage approach (Elastic Net + Bayesian Neural Network) were compared with those obtained using Horvath's method. We selected some GEO datasets that have been published after Horvath's paper was published to avoid including individuals that were used to build Horvath's predictive model and, hence, avoid overfitting (table 3.2). Figure 3 shows the meta-analysis of Rsquared and RMSE obtained from our proposed method and Horvath's approach. We observe that, as expected, there is a large heterogeity in the results (p of heterogeneity <0.01). There are some GEO datasets were both methods are providing similar R-squared (GSE20067, GSE51032 and GSE101764) but, in general, a better performance when using our proposed method is achieved (figure 4.3).In summary,we observe that our method explained a 4% more variability of the chronological age than Horvath's method (CI95%: 2%-5%). Similar conclusion was achieved when comparing RMSE (figure 4.3B). This is not surprising considering that there could be some degree of non-linearity in the relationship between methylation levels and aging. For processes that are, at least to some degree, non-linear, neural networks should provide better results than linear regression.

4.1.2 Age estimation using neural networks and DNA methylation levels

The techniques showed in this section, both neural networks and KNN, need a certain minimum amount of data to function properly but the sensitivity to the actual number of sample appear to be rather different. If the data set is relatively small the results seem to show that KNN works moderately well, regardless of the



A

Source	R-squared difference (95% CI)
GSE53128	0.08 [0.05; 0.11]
GSE49996	0.03 [0.02; 0.04]
GSE112987	0.08 [0.07; 0.08]
GSE58045	0.04 [0.03; 0.04]
GSE19711	0.06 [0.06; 0.06]
GSE20067	0.01 [0.00; 0.02]
GSE51032	0.01 [0.00; 0.02]
GSE101764	0.05 [-0.04; 0.13]
Total (fixed effect)	0.05 [0.04; 0.05]
Total (random effects)	0.04 [0.02; 0.06]

Heterogeneity: $\chi^2_7 = 407.28$ ($P < .01$), $I^2 = 98\%$

B

Source	RMSE difference (95% CI)
GSE53128	2.55 [1.36; 3.73]
GSE49996	1.47 [-2.93; 5.87]
GSE112987	6.15 [4.89; 7.41]
GSE58045	0.49 [-0.09; 1.08]
GSE19711	10.17 [7.04; 13.31]
GSE20067	5.98 [-14.24; 26.21]
GSE51032	32.43 [29.42; 35.44]
GSE101764	9.62 [8.09; 11.15]
Total (fixed effect)	3.21 [2.76; 3.66]
Total (random effects)	8.80 [3.74; 13.86]

Heterogeneity: $\chi^2_7 = 553.47$ ($P < .01$), $I^2 = 99\%$

Figure 4.3: Validation results using different GEO datasets corresponding to our proposed method (Elastic Net + Bayesian Neural Networks - EN + BNN) and Horvath's method. The panel A shows the difference of the R-squared obtained after regressing chronological age with the age predicted using our proposed method and Horvath's approach. Panel B shows the same comparison for the RMSE (x1000).

distance metric used with mean errors of approximately 10 year, while for larger datasets the neural network approached seemed to work better for the analyzed cases. The mean error found using the KNN approach is not smaller than the one found by some other researchers but given the rather small sample size it is a reasonable result. The KNN approach seemed to produce values that were moderately sensitive to k within the specified range with a maximum difference in the error of approximately four years. For some datasets, such as GSE3403, increasing the value of k seemed to decrease the error but this was not a constant trend for all the datasets analyzed. In fact for some datasets, such as (GSE4996), the error seemed to increase after a certain value of k . The average errors (over all the k values) were 9.67, 10.01 and 9.87 years for the datasets (GSE34035), (GSE49996) and (GSE 24884). The error for the larger dataset (GSE41037) was actually large than for the smaller ones, coming at 33.61, 33.03,33.93 and 33.49 year using the Euclidean, Cityblock, Correlation and Cosine distance metrics respectively. For small data samples the neural network approach did not seem to produce accurate forecasts. Forecasting accuracy did increase as the number of samples increased with the R^2 value for the larger database using just 10 neurons coming at a reasonable average of 0.63 for the GSE41037 dataset. The accuracy of the neural network forecast, while changing the amount of CpGs included in the simulations, were analyzed. For the dataset analyzed the best results were obtained not when using all the CpGs but when using a relatively small amount of approximately 300 CpGs selected randomly. For instance, for the small dataset (GSE 24884) the best results obtained was for a subset of 354 CpGs with the 95 % confidence interval for R^2 being [0.0514, 0.5330], which was the only entirely positive interval for the combinations analyzed. It is important to keep in mind that a poor result for small datasets was expected. For the large dataset (GSE41037) the best result, such as the previously mentioned average 0.63 was obtained also with a subset of approximately 300 CpGs. More research is needed to explore this issue but the results seem to support the idea that a large amount of the CpGs might add mostly noise for age calculation purposes. It is also interesting that the results seem to be relatively consistent even when taking several randomly selected sets of 300 CpGs.

4.2 Cancer

4.2.1 K-nearest neighbours and neural networks as tool to study DNA Methylation, cancer and concentration of information on CpGs

The results obtained using neural networks seem to indicate that even when a substantial part of the CpGs are excluded from the analysis the network can still detect the presence of cancer. This seems to be the case until approximately only 1% of the original data is included. The data is this regard seems more conclusive for the GSE26126 case than for the GSE57956 case. The Wilcoxon test finds no statistically meaningful difference, at a 5% significance level, between the out-of-sample errors until only one percent of the original CpGs are actually used. There is one exception to this trend with the Wilcoxon test rejecting the hypothesis that the subset containing 50% of the data has a median error equal to the median error for the entire dataset. This rejection cannot be sustained at a 1% significance level. Besides this case, the majority of the subsets seem to generate out-of-sample error with medians in line with the mean error obtained for the entire dataset. While the data for GSE57956 is less conclusive, with some large dataset having means, according to the Wilcoxon test, different from the mean obtained with the entire dataset the trend seems to be similar with the errors becoming different for all the subsets using 1% or less or the original data. It is important to remember that these subsets were created randomly. These results seem to support the idea that cancer information is spread among a large number of CpGs. The results obtained using the Kruskal-Wallis approach also seem to support that idea that cancer information is spread among a large number of CpGs. From the Kruskal-Wallis results it can be inferred that when the subset representing 0.01% of the original data is excluded all the other subsets seem to come from the same distribution. This results seems consistent among the two data sets analyzed.

The K-Nearest Neighbors algorithm approached yielded accurate results, particularly for the GSE57956 dataset. The results differed according to the type of distance metric used, number of neighbors selected and number of CpGs used for training purposes but for a large range of these factors the K-Nearest Neighbors algorithm was able to provide very accurate out-of-sample forecast, in many cases with an error of zero. It is important to notice that given the limited number of testing data, 18 in the case of the GSE57956 dataset, the error rate should be taken with caution but nevertheless the approach seems promising. Four different distance metrics (Euclidean, Correlation, Cosine and Minkowski) obtained a zero out-of-sample classification rate (18 test samples) using only one neighbor. This zero error rate remained constant for these distance metrics using a wide range of data sets, form 100% to 10% of the original data. Below this level the error rate remained low but higher. Forecasting ability seems to remain relatively high for some distance metrics, such as Minkowski or Chebychev, even for small data subsets, such as the 0.1% of the original data subset. Forecasting accuracy is lost in the GSE57956 when using 0.01% of the data. There are noticeable differences between the distance metrics used and the accuracy of the classifications obtained with the Hamming and Jaccard metrics having worse that the other metrics. Merely increasing the number of neighbors does not necessarily increase forecasting accuracy. Forecasting accuracy seems to be very poor for all the distance metrics when more than 160 neighbors are used in the case of the GSE26126 dataset. In the case of GSE57956 forecasting accuracy for all distance metrics is severely impaired when 100 neighbors or more are used. The results were less accurate for the GSE26126 dataset. A sample figure can be seen in figure 4.4. All the other graphs are shown in the appendix.

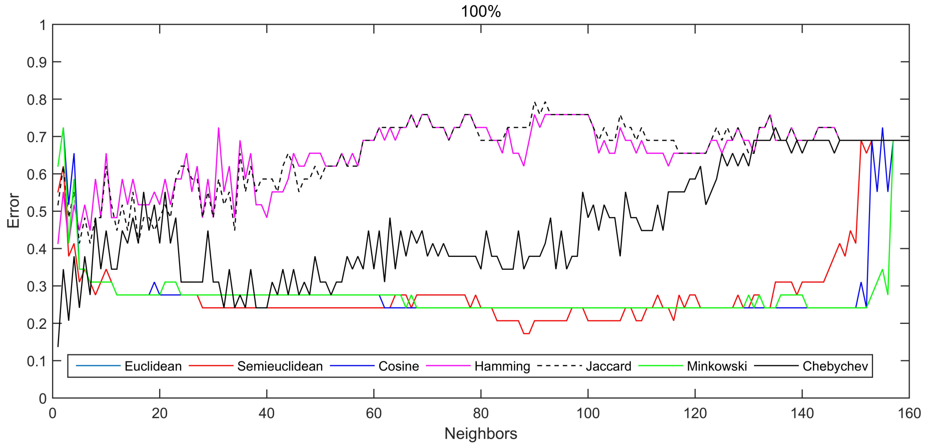


Figure 4.4: K-Nearest Neighbor. 100% of data. GSE26126

4.2.2 Cancer detection using support vector machines trained with linear kernels

Liver cancer

The lowest median error obtained using support vector machines for detection of cancer in liver tissue (out of sample data) in the 120 sample studied was obtained with a linear kernel (table 4.2). According to a Wilcoxon test the result was statistically significant at a 5 % significance level (table 4.3). The approach of using support vector machines with a linear kernel appeared to produce better results than using polynomial or Gaussian kernels. The linear SVM approach also produced a more accurate result than using a neural network with backpropagation and 10 neurons in the hidden layer. This NN approach generated a median error of 0.0556 with a standard deviation of 0.0329. All the simulations (for both SVM and NN) were repeated 100 times each. The error using SVM was statistically significantly smaller (table 4.4) for linear and polynomial kernels when compared to the NN approach but that was not the case when using a Gaussian kernel.

Statistic	**Linear**	**Polynomial**	Gaussian
Median	0.0250	0.0333	0.0833
Mean	0.0233	0.0347	0.0851
Standard deviation	0.0033	0.0032	0.0051

Table 4.2: Error rates for SVM using three different kernels

Model	**pl**	**h**
Linear - Polynomial	1.7e-39	1
Linear - Gaussian	3.6e-38	1
Polynomial - Gaussian	1.8e-38	1

Table 4.3: Results of Wilcoxon test for SVM using different kernels

Statistic	**Linear**	**Polynomial**	Gaussian
P	1.29e-5	1.19e-5	1.99e-26
H	1	1	1

Table 4.4: Comparison of NN results with NN (Wilcoxon)

Lung cancer

Similarly to the case of liver, the median error obtained using an SVM with a linear kernel is smaller (table 4.5) that the one obtained using either a polynomial or a Gaussian kernel. This hypothesis was tested with a Wilcoxon test (table 4.6). The median error obtained using backpropagation in a NN with 10 neurons was 0.1538 with and standard deviation of 0.0971. In this case, the error was statistically smaller using any of the three kernels and SVM when compared to neural networks (table 4.7). The confusion matrix and NN accuracy information can be seen in (figure 4.5) and (figure 4.6). All the compared errors were obtained using untrained data. In other words, data not used for training purpose by the algorithm.

Statistic	**Linear**	**Polynomial**	Gaussian
Median	0.1023	0.1364	0.1136
Mean	0.0972	0.1356	0.1198
Std. dev.	0.0067	0.0122	0.0085

Table 4.5: Error rates for SVM using three different kernels

Statistic	**Linear**	**Polynomial**
Median	4.2e-35	1
Mean	2.0e-33	1
Std. dev.	9.1e-19	1

Table 4.6: Results of Wilcoxon test for SVM using different kernels

Statistic	**Linear**	**Polynomial**	Gaussian
P	1.20e-3	7.80e-3	1.30e-3
H	1	1	1

Table 4.7: Comparison of NN results with NN (Wilcoxon)

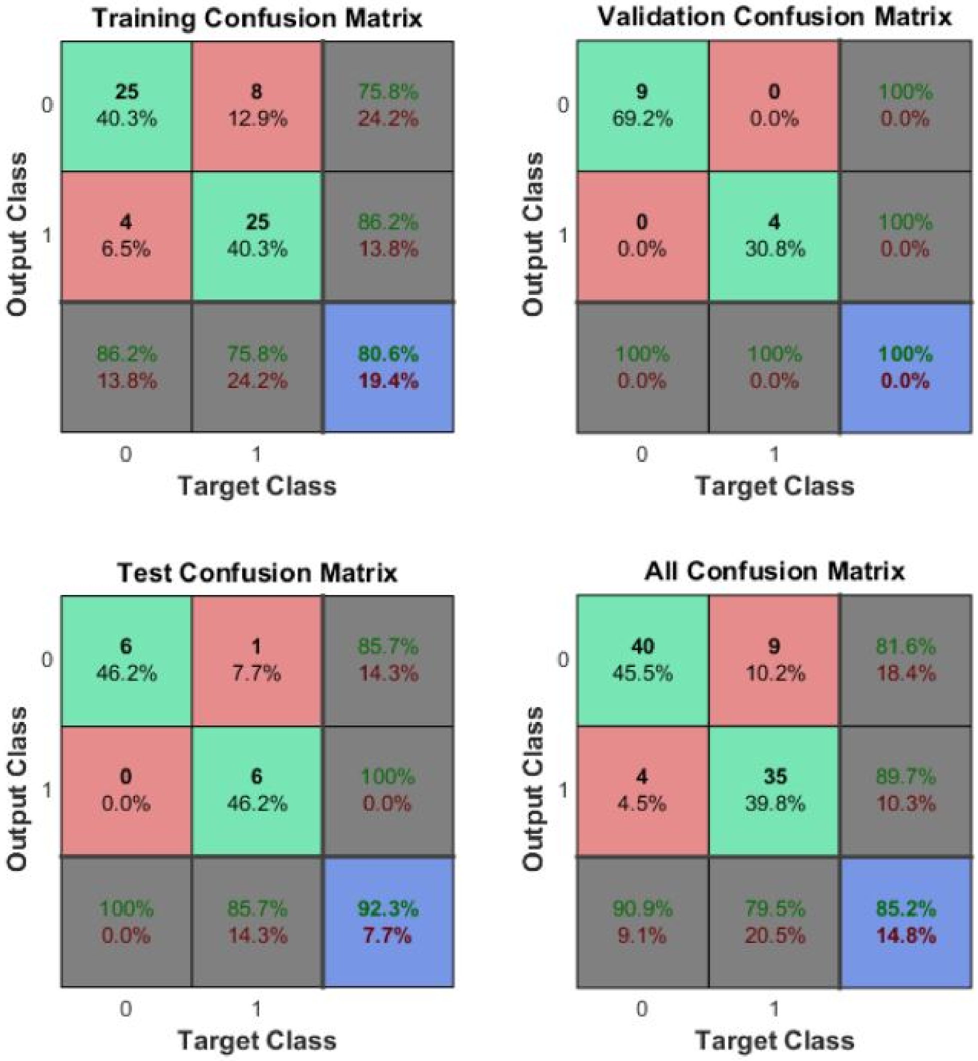


Figure 4.5: Confusion matrix sample obtained for a single lung cancer NN simulation

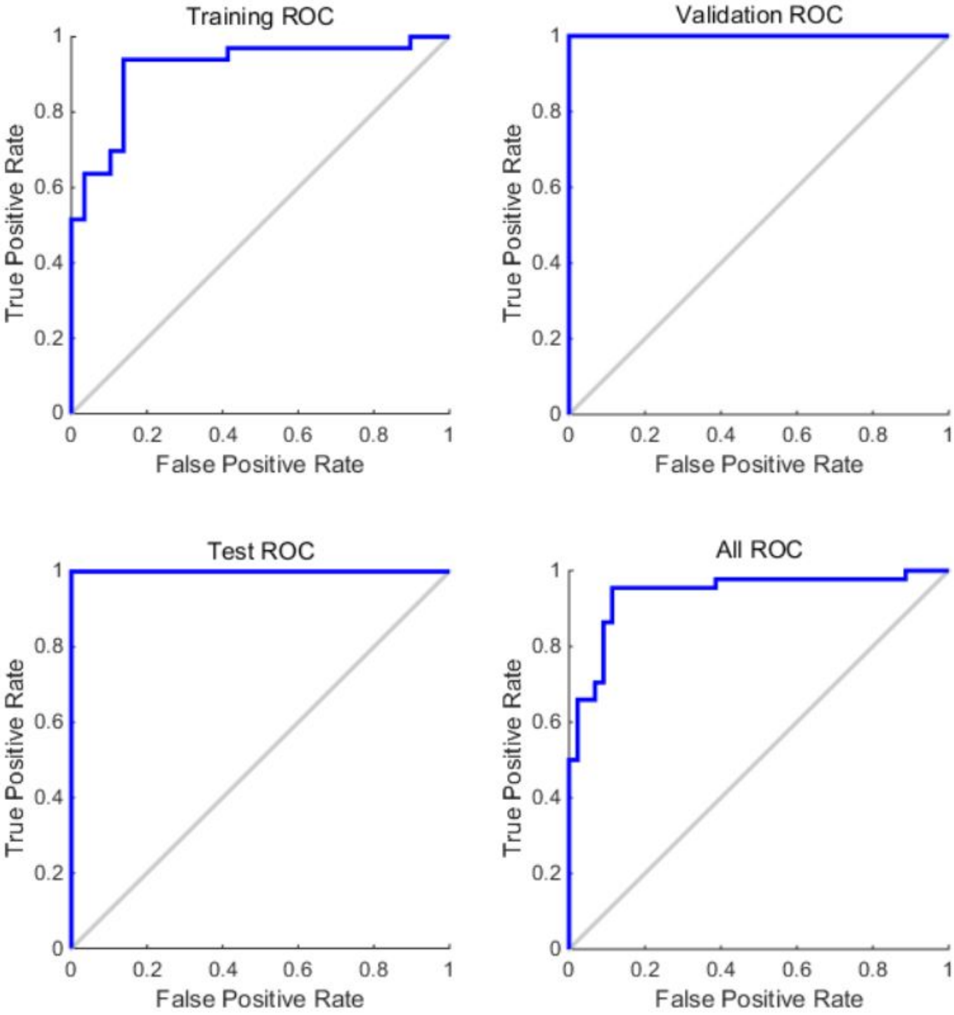


Figure 4.6: Lung cancer NN simulation

Cervical cancer

The results using tissue samples from the cervix (63 patients in total) are consistent with the ones obtained from lung and liver samples (table 4.8). The approach of using SVM with linear kernel seems to produce the smallest error and to be statistically significantly smaller than the median errors obtained using either polynomial of Gaussian kernels (table 4.9). The median result, after 100 simulations, obtained using backpropagation and a NN was 0.1111 with a 0.1008 standard deviation. Using, once more (table 4.10) a Wilcoxon test the values obtained using SVMs and NNs were compared. SVMs using linear and polynomial kernels had statistically significantly smaller errors than the NNs. The major difference with the previous cases is that for the cervical cancer data set the hypothesis that the medians for the error obtained using SMVs with Gaussian kernel and the NN being equal cannot be rejected.

Statistic	**Linear**	**Polynomial**	Gaussian
Median	0.0010	0.0159	0.0317
Mean	0.0006	0.0092	0.0263
Std. dev.	0.0005	0.0079	0.0120

Table 4.8: Error rates for SVM using three different kernels

Statistic	**p**	**h**
Linear – polynomial	3.2e-35	1
Linear – Gaussian	3.5e-33	1
Polynomial – Gaussian	1.5e-19	1

Table 4.9: Results of Wilcoxon test for SVM using different kernels

Statistic	**Linear**	**Polynomial**	Gaussian
P	0.0019	0.0008	0.0691
H	1	1	0

Table 4.10: Results of Wilcoxon test for SVM using different kernels

Bladder cancer

The approach used in the bladder cancer section was different from the previous three cases as the methylation data come from blood samples from the patients rather than from tissue samples from the area potentially affected by cancer. The idea was to see if the results can be extrapolated to analyzing the methylation of blood, which can be obtained with much less invasive techniques than organ tissue samples. The obtained median errors are substantially higher than in the previous cases (when using sample directly from the organs). In this case, the SVM with the smallest error (120 patients) is the one using a polynomial kernel (table 4.12). There appears also not to be a statistically significant difference when using neural networks compared to both a linear and a Gaussian kernel in a SVM. In table 4.13) the obtained values of a Wilcoxon test using SVMs and NNs are shown.

Statistic	Linear	Polynomial	Gaussian
Median	0.4167	0.3667	0.4166
Mean	0.4166	0.3682	0.4140
Std. dev.	0.0229	0.0217	0.0204

Table 4.11: Error rates for SVM using three different kernels

Statistic	p	h
Median	2.9e-26	1
Mean	3.2e-1	0
Std. dev.	2.1e-26	1

Table 4.12: Results of Wilcoxon test for SVM using different kernels

Statistic	Linear	Polynomial	Gaussian
P	0.0732	0.0214	0.0617
H	0	1	0

Table 4.13: Comparison of NN results with NN (Wilcoxon)

4.3 Diabetes

4.3.1 Diabetes detection using Bayesian neural networks and SNPs

A neural network (NN) with 150 neurons generated an out of same error of approximately 17.7% on the validation data $GERA_{testing}$, see table 4.14. This was done without using any additional SNPs filtering i.e., using all the 8,574 SNPs available. The model generated using $GERA_{training}$ and tested using $D_{1-testing}$ was then further tested using two external datasets $GENEVA$ and $FUSION$ with diverse accuracy of results. In the case of the data set $GENEVA$ the results were particularly accurate with the classification error rate being only 10.7%. The error histogram for the data set $GENEVA$ can be seen in figure 4.7. It should be mentioned that neither the $GENEVA$ data set or the $FUSION$ data set were used during the training phase. In other words, $GENEVA$ and $FUSION$ can be considered as pure external validation datasets. The results for the $GENEVA$ are less accurate with an error rate of approximately 46.8%. This discrepancy could be due to subtle experimental differences between the two experiments that generated the data sets $GENEVA$ and $FUSION$. Nevertheless, the approach of using neural networks with Bayesian learning for classification purposes seems to generate relatively accurate results for some of the analyzed data sets such as $GERA$ and $GENEVA$. Particular care was done on trying to avoid the issue of overfitting with the training phase employing 20 time cross validation. This was possible due to the relatively large amount of cases available for analysis.

Data set	n	Error (%)
$GERA$	28,426	17.72
$GENEVA$	1,673	10.7
$FUSION$	2,614	46.8

Table 4.14: Error rate for the validation data set

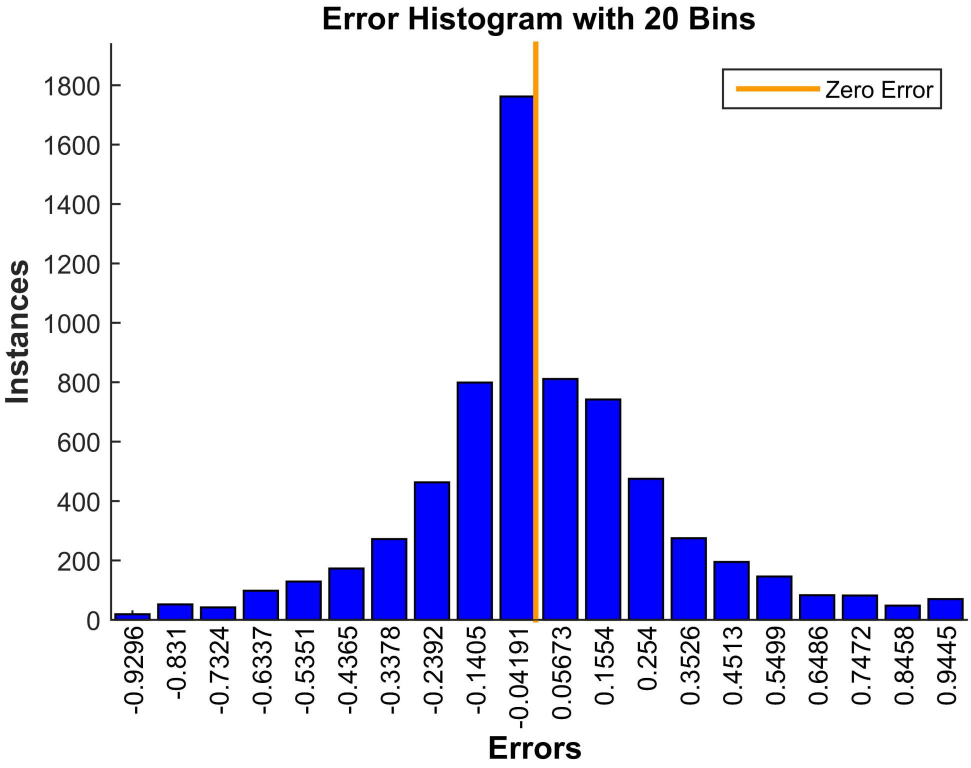


Figure 4.7: Error histogram for D_2

Chapter 5

Discussion

Technological advances have drastically increased the amount of data generated in experiments analyzing both DNA methylation as well as SNPs facilitating in turn the expansion of the OMICS field as well as several subfields. In this dissertation aging as well as two major diseases, namely cancer and diabetes, were analyzed using advanced statistical tools such as Bayesian Neural Networks (BNN), Support Vector Machines (SVM) and K-nearest Neighbours.

One of the main achievements of this dissertation was building a biological clock using Bayesian Neural Networks that generates accurate estimate of the age of an individual. It was also shown that reducing the dimensionality of the data through a technique such as elastic net before applying a neural network generates results that are more accurate. This is likely related to the issue of local minima in which the neural network gets stack in such local minima when facing a very large amount of inputs and stops finding a global minimum. This is practical term then translates to the issue of overfitting in which the biological clock generates very accurate forecast for the training data but produces poor results when facing new data. In addition to reducing the dimensionality of the data a technique that is frequently used to try to minimize the issue of overfitting is cross validation, which was used in the process of building the biological clock. The biological clock was built using as an input DNA methylation data. There are increasingly large amount of accessible DNA methylation data. Given this large amount of data available and the observation that there is no indication that the relationship between DNA methylation and age needs to be a linear one it seemed reasonable to use non-linear big data techniques, such as Bayesian Neural Networks, for the analysis.

Biological clocks built using EN regression followed by NN trained with Bayesian regularization applied to blood methylation data seem to produce better results than using a single step method (either using EN or NN to the entire methylation data). Our conclusion is based, first, on the results obtained when comparing the estimated age with the chronological age of the patients using only EN or neural networks with other training algorithm, such as Levenberg- Marquardt's algorithm. The second piece of evidence in favor of our proposed method relies on the comparisons performed with Horvath's method in different real datasets. One of the base assumptions is that there is some non-linearity in the relationship between methylation levels and aging that cannot be captured when using linear models, as in the case of Horvath's method that uses elastic net regression models. Meanwhile, it is likely that the reason why reducing the dimensionality of the data before applying neural networks produces good results is related to the issue of local minima.

In conclusion, the results support the hypothesis that there is some degree on nonlinearity in the relationship of methylation levels and age. Some existing linear models predict age with a reasonable level of accuracy but neural networks, particularly after reducing the dimensionality of the data, generate better results. This suggests that there is some level on non-linearity in the process. Having an R implementation of our predictive model will help biomedical researchers to incorporate an epigenetic biomarker to assess its impact in age-related complex diseases, such as cancer or Alzheimer's, among others.

Another of the results obtained was that application of K-Nearest Neighbors and Neural Networks to DNA methylation for age forecasting purposes seems to produce relatively accurate values. When there is only a limited sample space, i.e., small number of patients with DNA methylation available, the KNN technique seems to generate better results. For the larger dataset analyzed the neural networks approaches generated more accurate results. Reducing the dimensionality of the data (using a subset of all the CpGs rather than all the CpGs available) increased the forecasting accuracy. This is likely related to the problem of overfitting in which the neural network generates very accurate values for the training set but has poorer out of sample results. It is interesting that the results seem to be relatively consistent even when taking several randomly selected sets of 300 CpGs

In the second part of the dissertation I focused on the detection of patients with severe illness such as cancer and diabetes using DNA methylation data and SPNs data. Neural Networks and the K-Nearest Neighbors algorithms seem to be generate accurate classification results when applied to some cancer and control samples databases identifying between healthy and sick patients using DNA methylation data as input. Even when a substantial part of the CpGs are excluded from the analysis a neural network can still detect the presence of cancer. This seems to be the case until approximately only 1% of the original data is included

Another important conclusion is that the distance metric used in the K-Nearest Neighbor algorithm is a very important parameter to ensure accurate results with the Minkowski and Chebychev metrics generating some of the best results. While clearly more research is necessary the analysis also seems to suggest that cancer information seems to be spread among a large number of CpGs with small randomly selected subsets of CpGs providing accurate classifications.

For the data sets analyzed, the results indicate that when using DNA methylation data from the liver, lung or cervix, to determine the presence of cancer using a support vector machine a linear kernel training generates results that are more accurate, than using other training kernels such as polynomial or a Gaussian kernels. The difference was statistically significant (tested with a Wilcoxon test). The results were also more accurate than the ones obtained using a simple backpropagation NN with 10 neurons. These results were also statistically significant. The dynamics seems to be rather different when the methylation analysis is performed on blood samples, rather than tissue from the previously mentioned organs. In this case the accuracy of the method seems to be substantially smaller and there appears to be less statistically significance differences between using SVM and NN.

Diabetes was the topic of the last section of the dissertation. In this case the input data were SNPs rather than DNA methylation data. Neural networks with Bayesian learning seem to produce relatively accurately results for type 2 diabetes detection in the majority of cases analyzed using SNPs. The results were less accurate for one of the external databases used for validation. This might be related to slight different experimental differences between the experiments. As previously mentioned the SNPs filtering was done in this case using single association studies rather than Elastic Net regression. Future work could include filtering the SNPs using Elastic Net and compare the results with those obtained when using single association analysis.

Chapter 6

Conclusions

Given the large amount of genomic and epigenomic data currently available it is increasingly important to find better approaches to analyze such data in applications in fields such as aging, cancer and diabetes. Based on the results obtained in this dissertation the main conclusions are as follows:

- DNA methylation can be used as the base data for biological clocks as well as for cancer detection.
- SNPs data can be used for diabetes detection
- Using a two-step approached composed of a data dimensionality reduction step followed by a Bayesian Neural Networks is possible to improve the accuracy of existing DNA methylation based biological clocks.
- Support Vector Machines can be successfully applied for cancer detection for a variety of different cancers including liver, lung, cervical and bladder cancer.
- Cancer can be detected using a small amount of CpGs DNA methylation and Support Vector Machines.
- For large datasets, containing many CpGs DNA methylation data, Support Vector Machines generate better results than K-nearest neighbors identifying

cancer. For very small datasets K-nearest Neighbors generated more accurate results.

- Bayesian Neural Networks produce accurate results when applied to SNPs data to detect patients suffering from type 2 diabetes.

Chapter 7

Future Work

Currently there is a new version of the Illumina machine, the Illumina 855k which greatly increases the amount of CpGs analyzed. While at the moment there are very few available datasets using this new powerful machine that will no doubt change in the future, allowing for further refining of the models. Another area of possible expansion is the usage of deep neural networks that use many hidden layers. The computational challenges of using deep neural networks in combination for instance with methylation data from an Illumina 855k machine (850,000 CpG data per patient) should not be underestimated but it is an area of clear potential interest. As the data available continuously increases another important are of future work are improvements in data dimensionality reduction techniques such as for instance the previously mentioned Elastic Net approach.

Appendix A

Additional cancer statistics

3.jpg

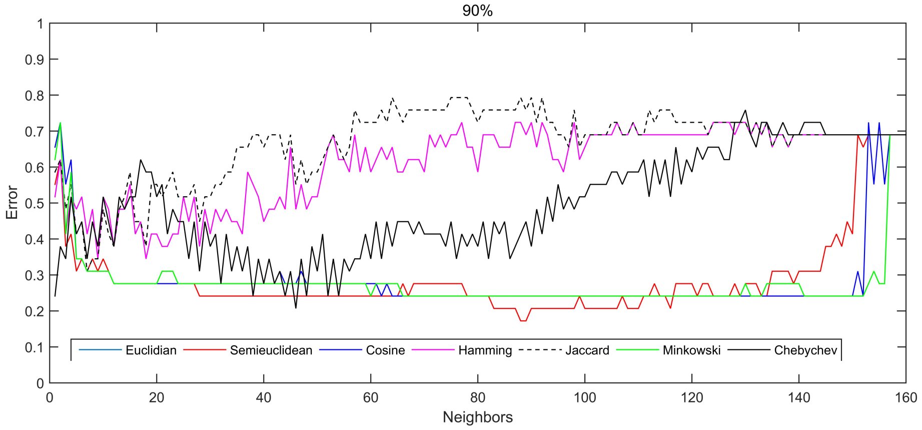


Figure A.1: K-Nearest Neighbor. 90% of data. GSE26126

4.jpg

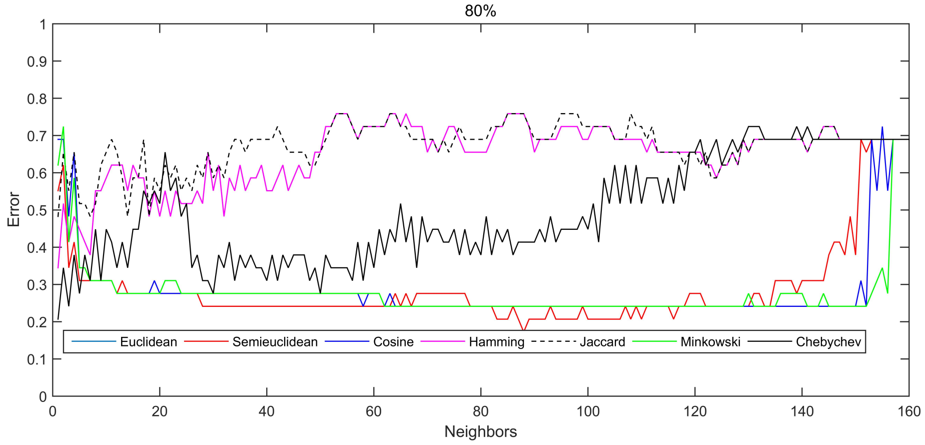


Figure A.2: K-Nearest Neighbor. 80% of data. GSE26126

5.jpg

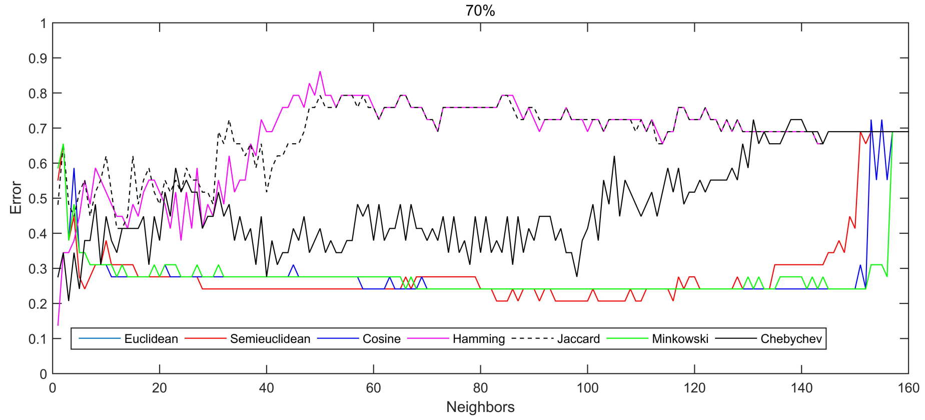


Figure A.3: K-Nearest Neighbor. 70% of data. GSE26126

6.jpg

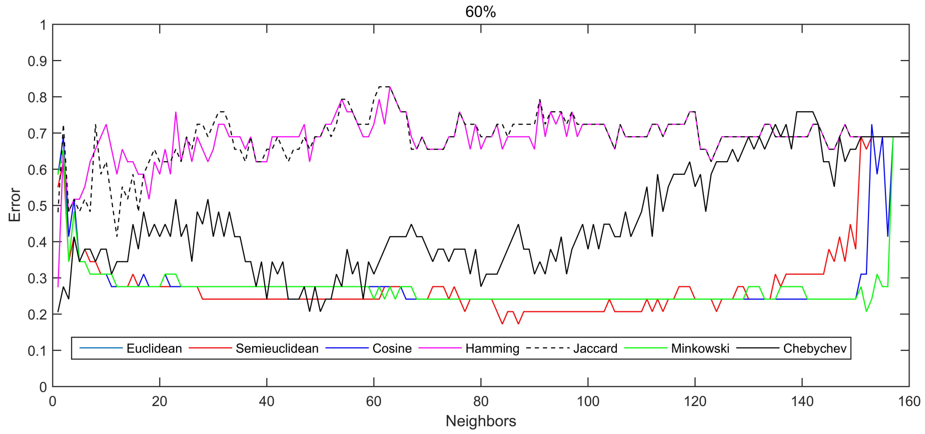


Figure A.4: K-Nearest Neighbor. 60% of data. GSE26126

7.jpg

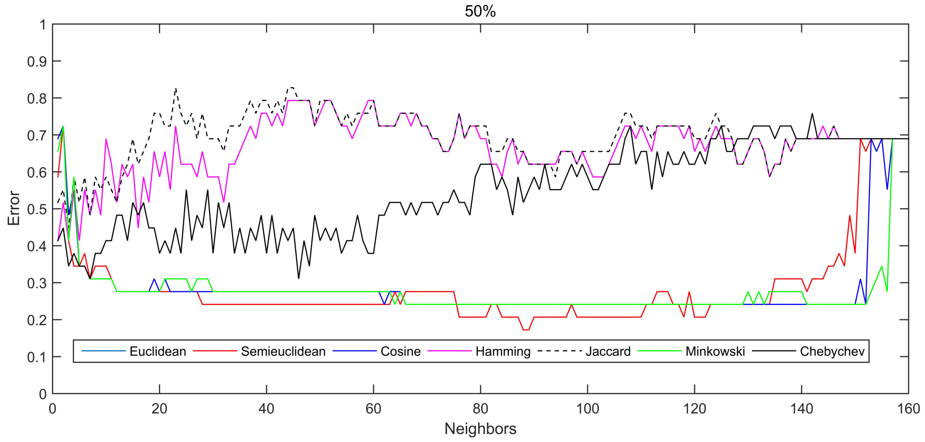


Figure A.5: K-Nearest Neighbor. 50% of data. GSE26126

8.jpg

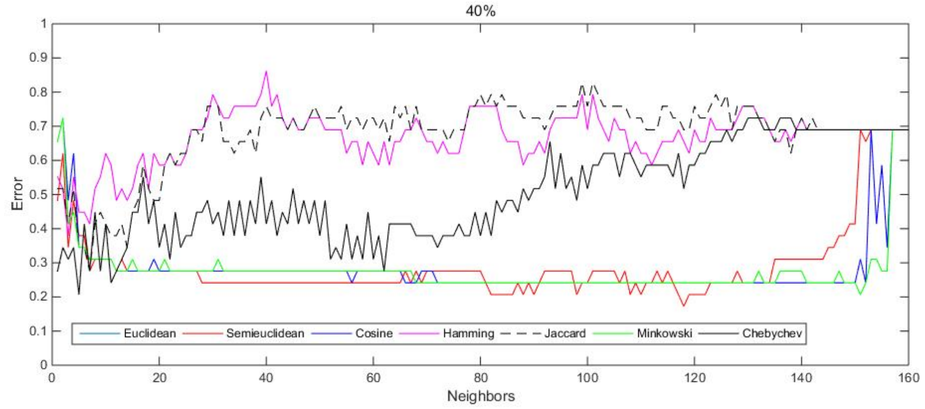


Figure A.6: K-Nearest Neighbor. 40% of data. GSE26126

9.jpg

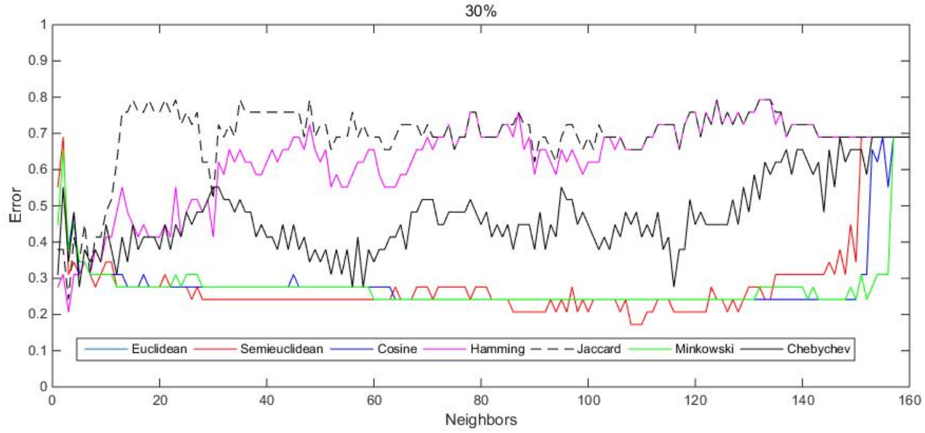


Figure A.7: K-Nearest Neighbor. 30% of data. GSE26126

10.jpg

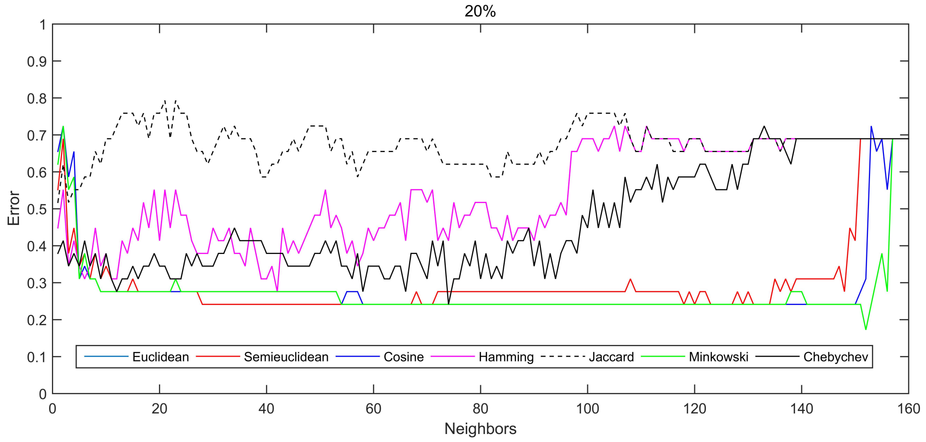


Figure A.8: K-Nearest Neighbor. 20% of data. GSE26126

11.jpg

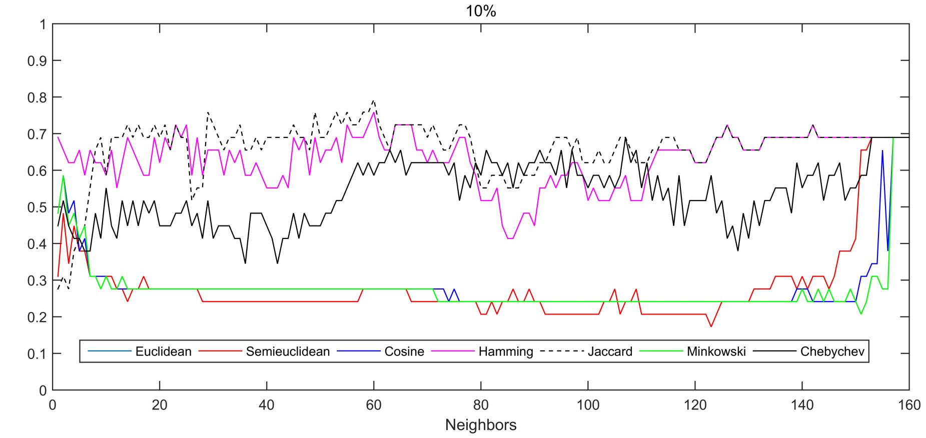


Figure A.9: K-Nearest Neighbor. 10% of data. GSE26126

12.jpg

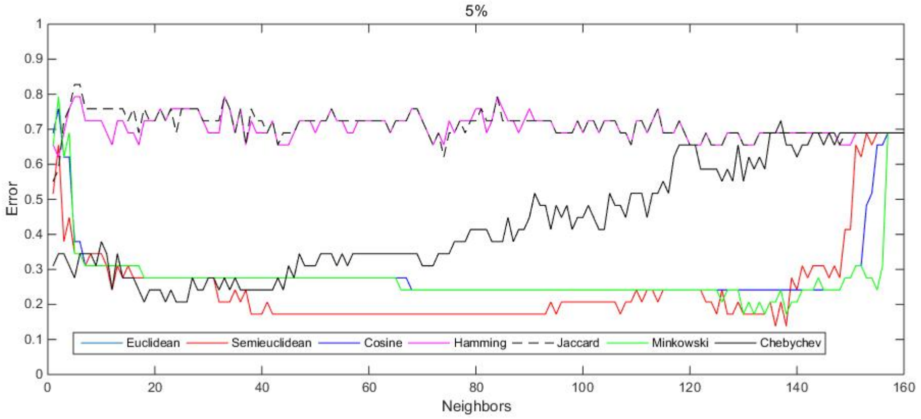


Figure A.10: K-Nearest Neighbor. 5% of data. GSE26126

13.jpg

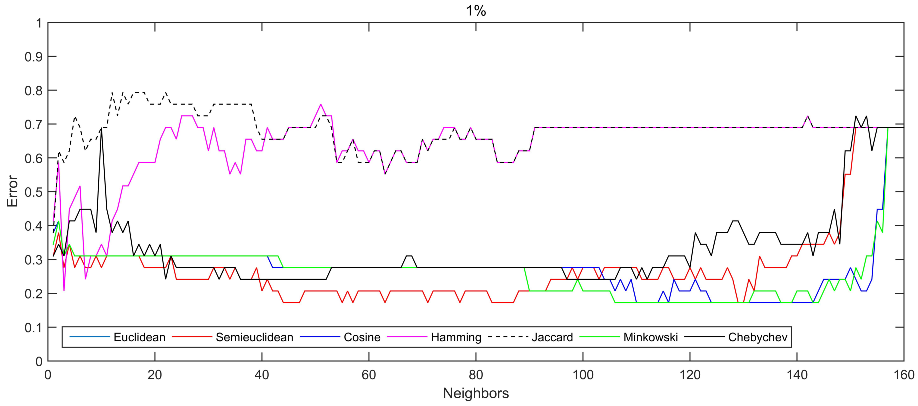


Figure A.11: K-Nearest Neighbor. 1% of data. GSE26126

14.jpg

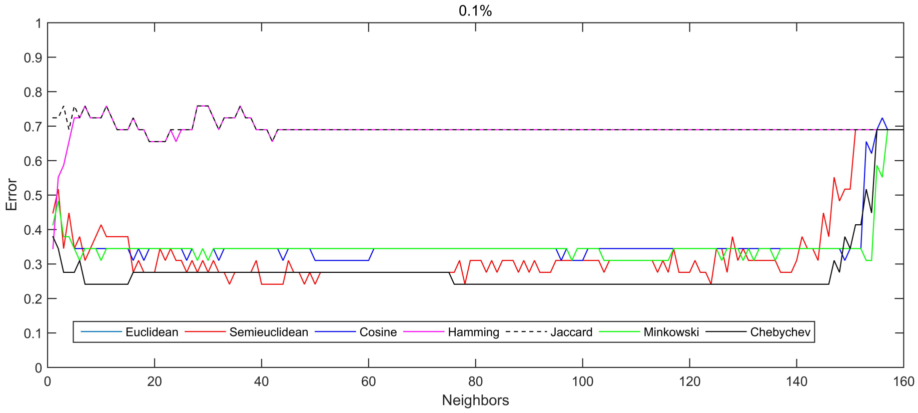


Figure A.12: K-Nearest Neighbor. 0.1% of data. GSE26126

15.jpg

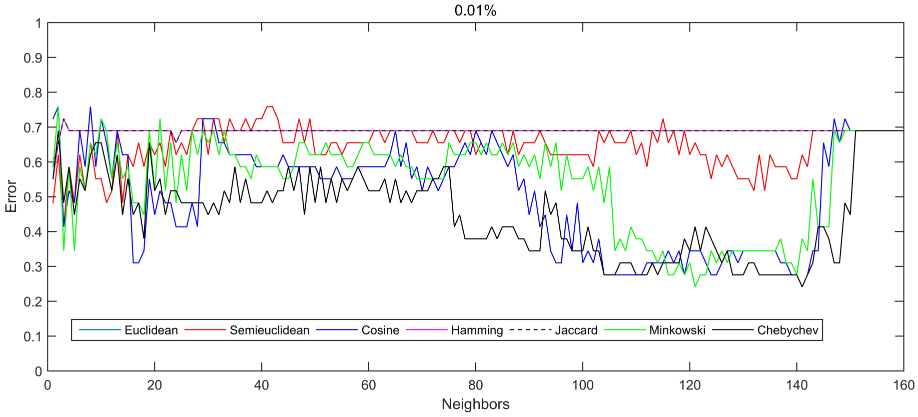


Figure A.13: K-Nearest Neighbor. 0.01% of data. GSE26126

16.jpg

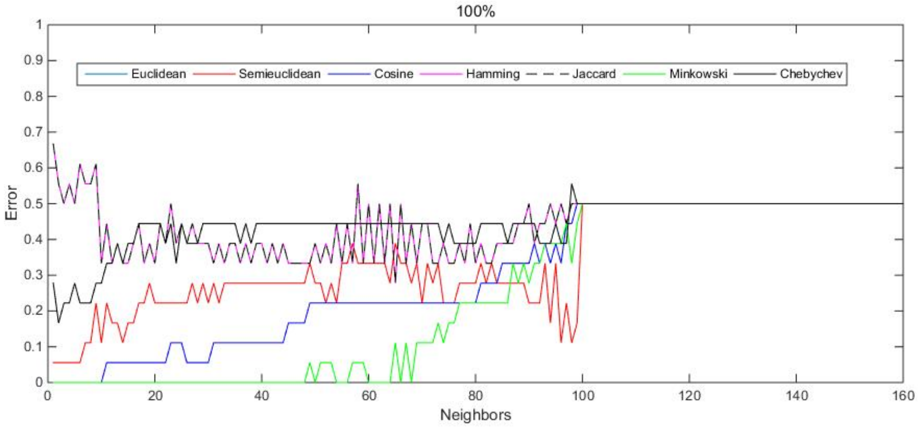


Figure A.14: K-Nearest Neighbor. 100% of data. GSE57956

17.jpg

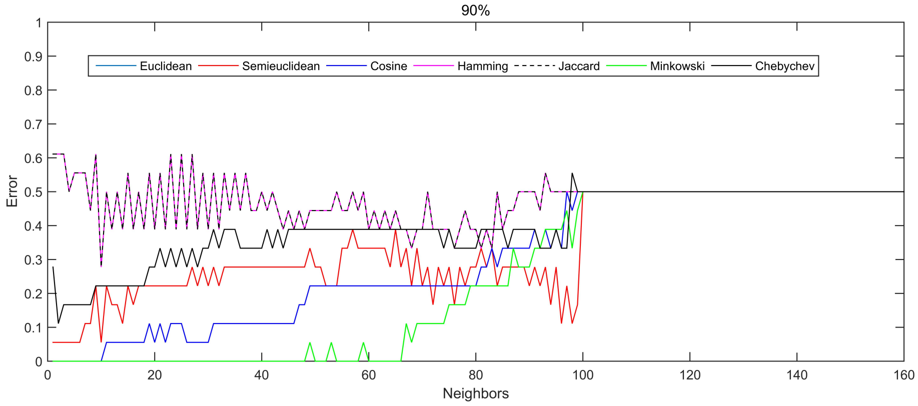


Figure A.15: K-Nearest Neighbor. 90% of data. GSE57956

18.jpg

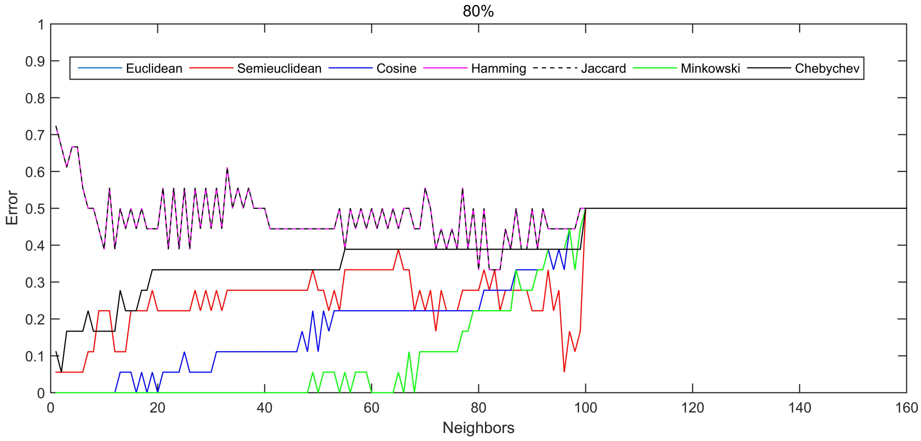


Figure A.16: K-Nearest Neighbor. 80% of data. GSE57956

19.jpg

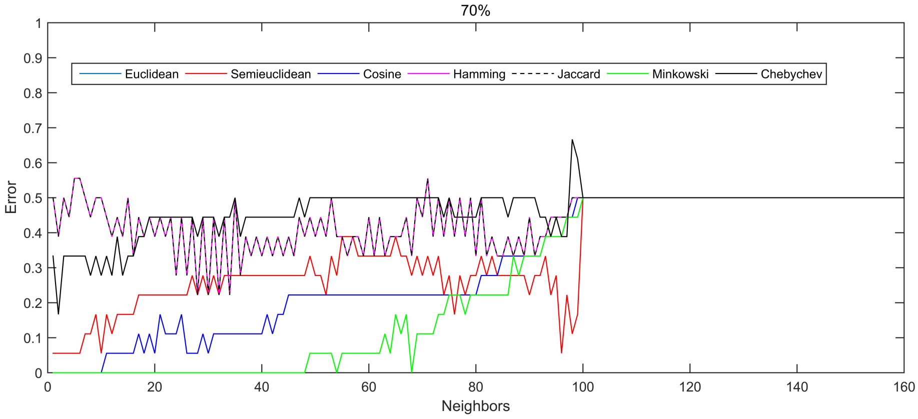


Figure A.17: K-Nearest Neighbor. 70% of data. GSE57956

20.jpg

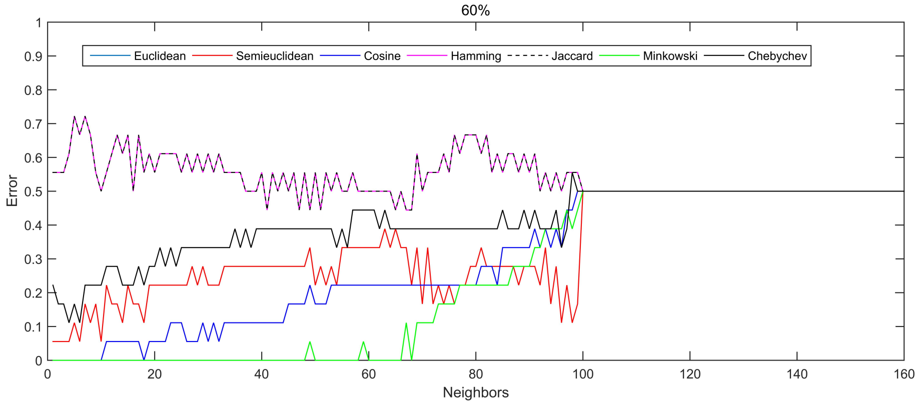


Figure A.18: K-Nearest Neighbor. 60% of data. GSE57956

21.jpg

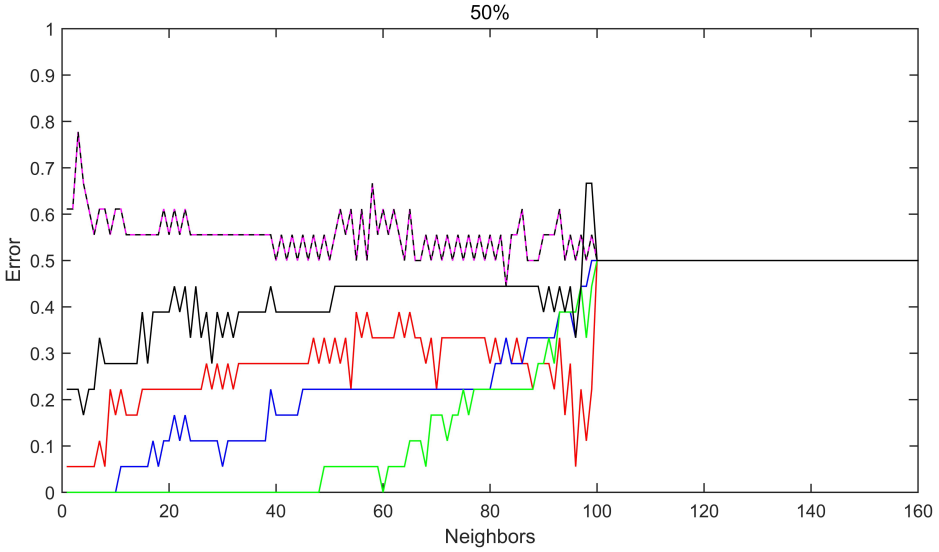


Figure A.19: K-Nearest Neighbor. 50% of data. GSE57956

22.jpg

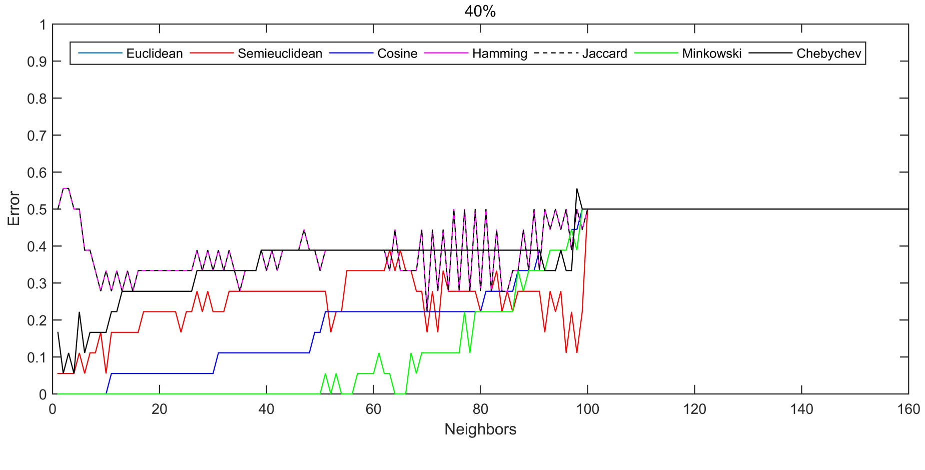


Figure A.20: K-Nearest Neighbor. 40% of data. GSE57956

23.jpg

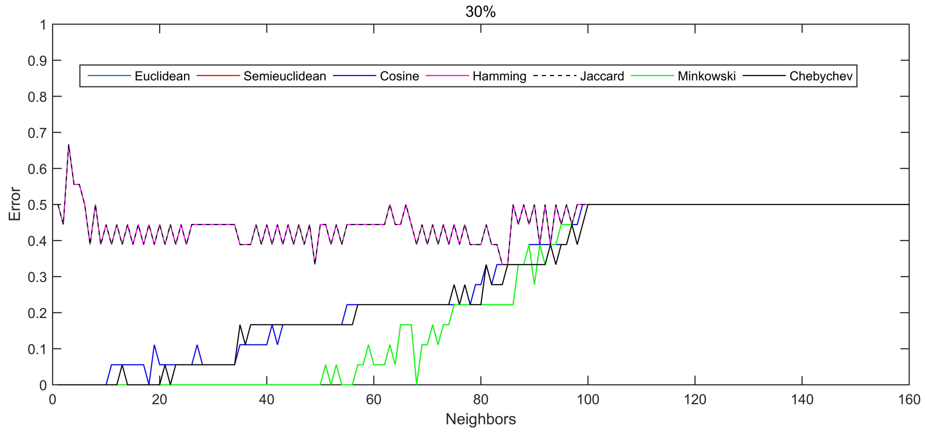


Figure A.21: K-Nearest Neighbor. 30% of data. GSE57956

24.jpg

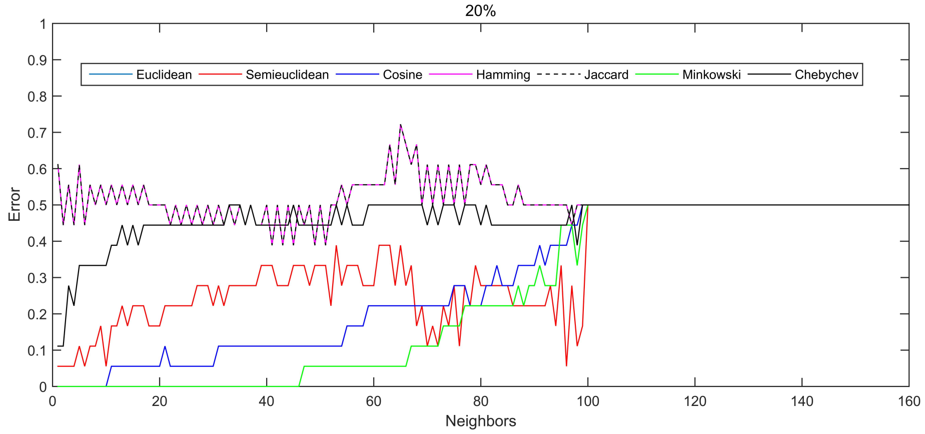


Figure A.22: K-Nearest Neighbor. 20% of data. GSE57956

25.jpg

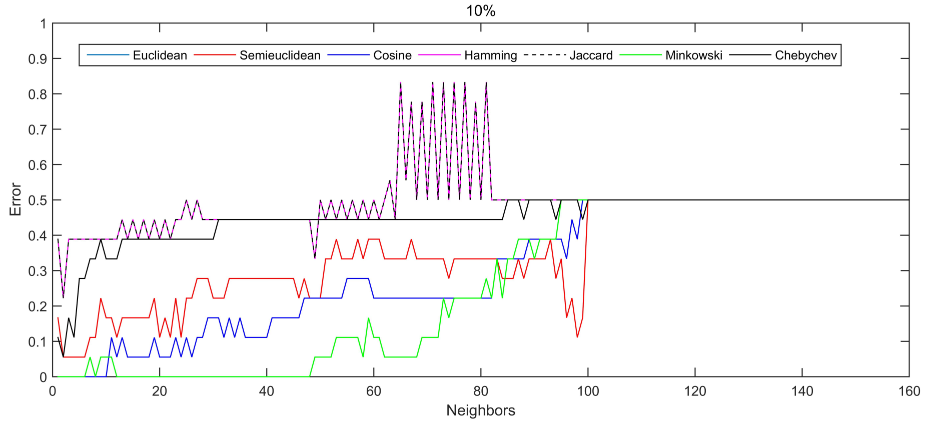


Figure A.23: K-Nearest Neighbor. 10% of data. GSE57956

26.jpg

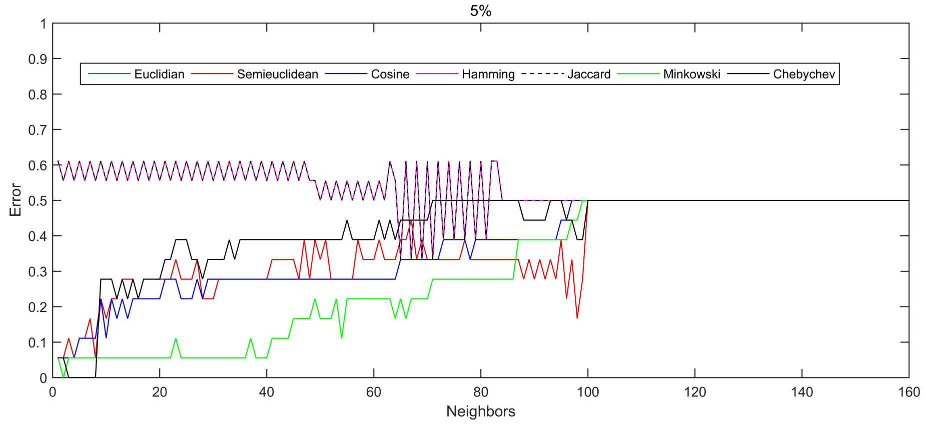


Figure A.24: K-Nearest Neighbor. 5% of data. GSE57956

27.jpg

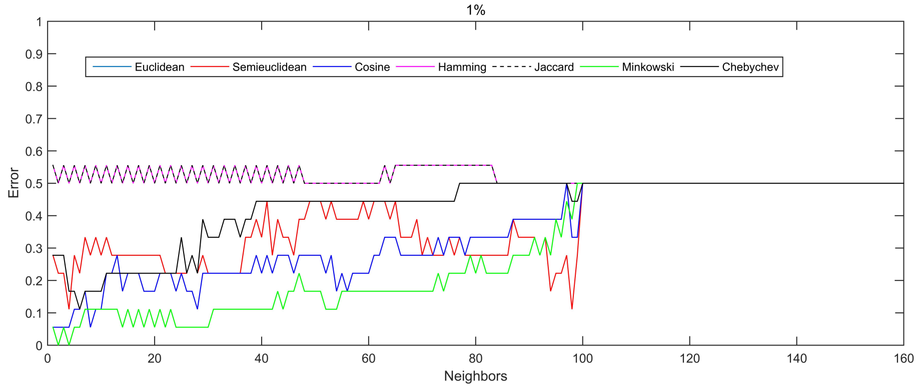


Figure A.25: K-Nearest Neighbor. 1% of data. GSE57956

28.jpg

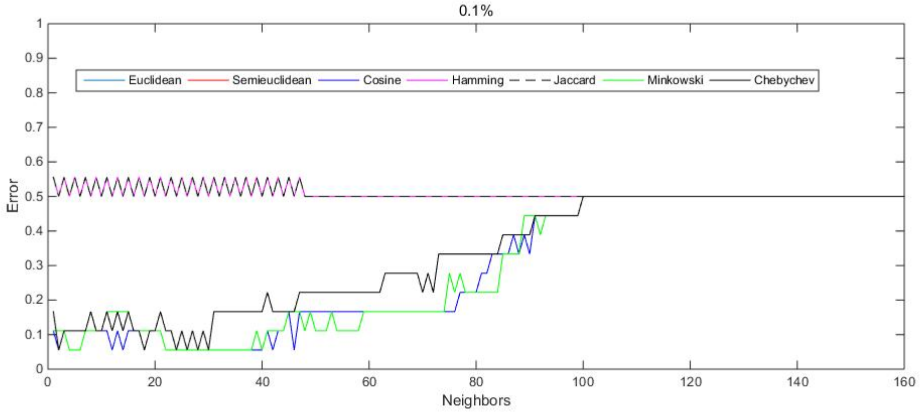


Figure A.26: K-Nearest Neighbor. 0.1% of data. GSE57956

29.jpg

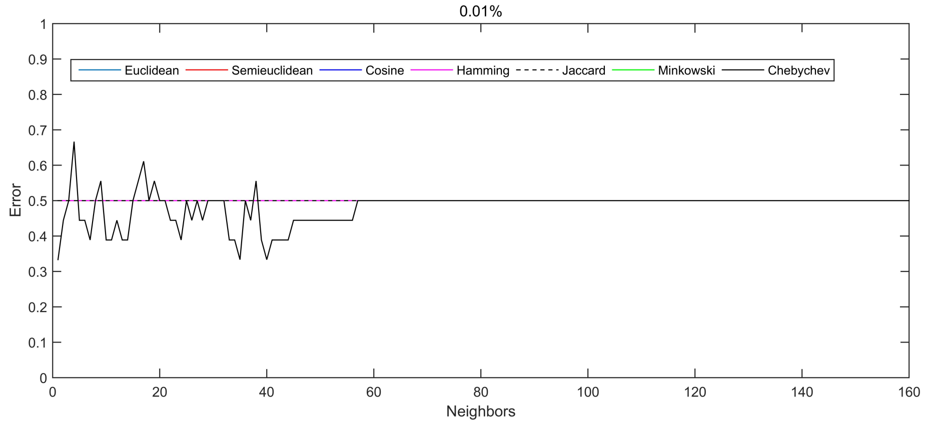


Figure A.27: K-Nearest Neighbor. 0.01% of data. GSE57956

Bibliography

Salvatore Alesci Abuasad, Mohamed Chaouchi. Biomarkers in the age of omics: time for a systems biology approach. *A journal of integrated biology*, 2011. URL `10.1089/omi.2010.0023`.

Albert. Biologia molecular de la celula. quinata edicion. *Ediciones Omega*, 2010.

Anders Fredriksson Archer. Pharmacogenomics and personalized medicine in parkinsonism. 2013.

Thorell A Ryden M Arner, Sinha I. The Epigenetic Signature of Subcutaneous Fat Cells is Linked to Altered Expression of Genes Implicated in Lipid Metabolism in Obese Women. *Clin Epigenetics*, 2015.

Sultan Aljahdali Ashfaq. Comparative Prediction Performance with Support Vector Machine and Random Forest Classification Techniques. *International journal of computer applications*, 2013.

Boutalline M Badi. The Neural Networks: Application and Optimization of LM Algorithm for Tifinagh Character Recognition. *International Journal of Science, Environment and Technology*, 2013.

Nagle T Bahram, Schiffman S. Performance of the Levenberg-marquardt Neural Network Training Method in Electronic Nose Applications. *Sensors and Actuators B: Chemical*, 2003.

Rubino G Mostafa S Basterrech, Samir M. Levenberg-marquardt Training Algorithms for Random Neural Networks. *The Computer Journal*, 2011.

Smith Baxter. Lagrange Multipliers Tutorial in the Context of Support Vector Machines. *University of Newfoundland*, 2012.

Boiarskikh G Berdishev, Korotaev G K. Nucleotide Composition of DNA and RNA from Somatic Tissues of Humpback Salmon and its Changes During Spawning. *Biokhimiia*, 38, 1967.

P Bhuvanesuari. Genetic risk scores for diabetes diagnosis and precision medicine. *Procedia Materials Science*, 2015. URL `https://doi.org/10.1016/j.mspro.2015.06.077`.

Yu Z Qiang Y Bo-Liu, Ying W. Deep neural networks for high dimension, low sample size data. *Proceeding of the Twenty-sixth International Joint Conference on Artificial Intelligence*, 2017.

Mohlke Boehnke. Recent advances in understanding the genetic architecture of type 2 diabetes. *Hum Mol Genet*, 2017.

Matthew Bogyo. New technologies and their impact on 'omics' research. *Current opinion in chemical biology*, 2013. URL `http://dx.doi.org/10.1016/j.cbpa.2013.01.005`.

Pritchard JK Boyle, Li YI. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 2017.

Brandeis. Dynamics of DNA Methylation During Development. *TUGBoat*, 14, 1993. URL `10.1002/bies.950151103`.

Zampiere P Caifa. DNA Methylation and Chromatin Structure: the Puzzling CpG Islands. *Journal of Cellular Biochemistry*, 2004. URL `https://doi.org/10.1002/jcb.20325`.

Virginia Canas. Applications of advanced omics technologies: From genes to metabolites. *Comprehensive analytical chemistry*, 2014. URL `https://www.`

sciencedirect.com/handbook/comprehensive-analytical-chemistry/vol/64.

Antonio Berenguer-LLergo Agustin Fernandez Carmona, Daniel Azuara. DNA Methylation Biomarkers for Noninvasive Diagnosis of Colorectal Cancer. *Cancer Prevention Center*, 2013. URL 10.1158/1940-6207.

Siffroi Jean-Pierre Bouret Dominique Cassuto-Nino, Montjean Debbie. Different Levels of DNA Methylation Detected in Human Sperms after Morphological Selection Using High Magnification Microscopy. *Biomedical Research International*, 2016.

CDC. National diabetes statistics report. center for disease control and prevention. *International journal of scientific research publications*, 2017. URL https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf.

Cecil. Tratado de medicina interna. volume 2. *Elsevier Saunders*, 2017.

Melnick A Cerchietti. DNA Methylation-based Biomarkers. *Journal of Clinical Oncology*, 2017.

Gary Johanning Chandrika-Piyathilake. Cellular Vitamins, DNA Methylation and Cancer Risk. *The Journal of Nutrition*, 132, 2002. URL https://doi.org/10.1093/jn/132.8.2340S.

J Grinband. Chang. Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. 2018. URL https://doi.org/10.3174/ajnr.A5667.

Pang Lifang Lam Kinman Chaoli, Zhang Shuheng. Using the k-nearest Neighbor Algorithm for the Classification of Lymph Node Metastasis in Gastric Cancer. *Computational and Mathematical Methods in Medicine*, 2012.

Massie Charles. The Importance of DNA Methylation in Prostate Cancer Development. *The journal of steroid biochemistry and molecular biology*, 166, 2017.

Bishwanah Chatterjee. Cardiovascular disease pharmacogenomics. *Springer*, 2013. URL 10.1007/978-81-322-1184-6_20.

Gao Y Chen-Yu. Levenberg-marquardt Method for the Eigenvalue Complementarity Problem. *The Scientific World Journal*, 2014.

Chris-Overall. Introduction of omics and bioinformatics. *University of North Carolina*, 2011. URL https://webpages.uncc.edu/~jweller2/pages/SummerCamp2011/SummerCamp2011_Presentations/Overall_intro_to_binf_hs.pdf.

Charkravarti A Collins. Variations on a theme: cataloging human dna sequence variation. *Science*, 1997.

Lopomo A Spisni R Copede, Grossi E. Application of Artificial Neural Networks to Link Genetic and Environmental Factors to DNA. *Epigenomics*, 7, 2015.

Angela Lopomo Coppede. Application of artificial neural networks to link genetic and environmental factors to dna methylation in colorectal cancer. *Epigenomics*, 2015.

Ricard Saffery Craig. The power of two: epigenetics and twins. *Twin research and human genetics*, 2015. URL 10.1017/thg.2015.90.

Simons Daniels. Epigenetic Influences and Disease. *Nature Education*, 1, 2008.

Rakesh Singal Das-Partha. DNA Methylation and Cancer. *Journal of Clinical Oncology*, 151, 2004.

Rakesh Singal Das-Partha. DNA Methylation and Cancer. *Journal of Clinical Oncology*, 22, 2016. URL 10.1200/JCO.2004.07.151.

Uthus Eric Davis. DNA Methylation, Cancer Susceptibility and Nutrient Interactions. *Experimental Biology and Medicine*, 229, 2004.

Sweatt Davis Day. DNA Methylation and Memory Formation. *Nature Neuroscience*, 13, 2010.

Vasudei Zambarem Vasco Azevedo Debmalya-Barh. Omics: Applications in biomedical, agricultural and environmental science. *CRC Press*, 2017.

Chen Yan Dong-Zhicheng. Transcriptomics: advances and approaches. *Science China. Special issue: non-coding RNAs*, 2013. URL `10.1007/s11427-013-4557-2`.

Xi Lifeng Dushicang, Liu Changping. A Selective Multiclass Support Vector Machine Ensemble Classifier for Engineering Surface Classification Using High Definition Metrology. *Journal of Manufacturing Science and Engineering*, 137, 2015.

Feng Wu Zhi Edriss. Breast Cancer Classification using Support Vector Machine and Neural Network. *International Journal of Science and Research (IJFR)*, 2012.

Elizabeth H Lin J Dhabhar F Epel, Blackbur E. Accelerated Telomere Shortening in Response to Life Stress. *Proceedings of the Pational Academy of Sciences*, 2004.

Balint B Fernandez, Martin Subero JI. DNA Methylation Fingerprinting of 1628 Human Samples. *Genome Research*, 22, 2012.

Jesus Florez. Farmacologia humana. quinta edicion. *Elsevier Manson*, 2008.

PeterF Fransquet. The epigenetic clock as a predictor of disease and mortality risk: a systematic review and meta-analysis. *Clinical Epigenetics*, 2019.

Rama M Ganesan, Venkatesh K. Application of Neural Networks in Diagnosis Cancer Disease using Demographic Data. *International Journal of Computer Application*, 1(26), 2010.

Gavin. The Levenberg Marquardt Method for Nonlinear Least Squares Curve Fitting Problems. *Duke University*, 2011.

Genome-Website. Glossary of genomics and bioinformatics terms. 2018. URL `http://www.ornl.gov/sci/techresources/Human_Genome`.

Krzysztof Gorynski. Artificial neural networks approach to early lung cancer detection. *Central European Journal of Nedicine*, 2014.

Allison Mayle Grant-Challen, Deqian Su. Dnmt3a and Dnmt3b have O and Distinct Functions in Hematopoietic Stem Cells. *Stem Cell*, 5, 2014. URL `https://doi.org/10.1016/j.stem.2014.06.018`.

Rune Matthiesen Gubb. Introduction to omics. *Bioinformatic methods in clinical research*, 2009. URL `https://link.springer.com/protocol/10.1007/978-1-60327-194-3_1`.

Foresee D Hagan. Gauss-newton Approximation to Bayesian Learning. *International Symposium on Neural Networks*, 1997.

Menhaj M Hagan. Training Feedforward Networks with the Marquardt Algorithm. *IEEE Transactions on Neural Networks*, 1994.

Guinnet J Hannum. Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Molecular Cell*, 2013. URL `10.1016/j.molcel.2012.10.016`.

Goel A Hashimoto, Zumwalt T. Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Epigenomics*, 2016.

Martin Taylor Hemmer Hansen Helyar, Dorte Bekkevold. Application of snps for population genetics of non-model organism: new opportunities and challenges. *Molecular ecology resources*, 2011. URL 10.1111/j.1755-0998.2010.02943.x.

FernandezCallejo Marcos Hernando-Herraez, Padro-Martinez Javire Padro. Garg Paras. Dynamics of DNA Methylation in Recent Human and Great Ape Evolutions. *PLos*, 2013.

Richard Horgan. Omic technologies: genomics, transcriptomics, proteomics and metabolomics. *Obstetrics gynecology*, 2011. URL https://doi.org/10.1576/toag.13.3.189.27672.

Richard Horgan. What is a genome? *U.S. National Library of Medicine*, 2017. URL https://ghr.nlm.nih.gov/primer/hgp/genome.

Ammerpoh W Schonfels O Horvath, Erhart S. Obesity Accelerates Epigenetic Aging of Human Live. *Proceedings of the National Academy of Sciences of the United States of America*, 2014.

Langfelder P Kahn RS Horvath, Zhang Y. Aging Effects on DNA Methylation Modules in Human Brain and Blood Tissue. *Genome Biol*, 2012.

Steve Horvath. DNA Methylation Age of Human Tissues and Cell Types. *Genome biology*, 2013. URL https://doi.org/10.1186/gb-2013-14-10-r115.

Ebrahimi Goliaei Hosseinzadeh. Classification of Lung Cancer Tumors Based on Structural and Physiochemical Properties of Proteins by Bioinformatics Models. *PLOS ONE*, 7, 2014.

Zuzanna Borek Sebo Withoff Hrdlickova, Rodrigo de Aleida. Genetic variation in the non-coding genome: Involvement of micro-rnas and long non-coding rnas in disease. *Biocimica et biophysica acta – molecular basis of disease*, 2014. URL https://doi.org/10.1016/j.bbadis.2014.03.011.

Hastie T Hui-Zhang. Regularization and Variable Selection via the Elastic Net. *Journal of the royal statistical society*, 2005.

Robert Jacob. Folate, DNA Methylation, and Gene Expression: Factors of Nature and Nurture. *The American Journal of Clinical Nutrition*, 72, 2000. URL `https://doi.org/10.1093/ajcn/72.4.903`.

Ines Fonfara Michael Hauerm Jennifer Dudna Emmanuelle Charpentier Jinek, Krzysztof Chylinski. A programmable dual-rna-guided dna endonuclease in adaptive bacterial immunity. 2015. URL `http://science.sciencemag.org/content/337/6096/816`.

Kobor M Jones, Fejes A. DNA Methylation and Gene Expression: Who is Driving and Who is a Long for the Ride. *Genome Biology*, 2013.

Yousefi Paul Bakulski Kelly Joubert, Janine Felix. DNA Methylation and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *The American Journal of Human Genetics*, 2016.

Tae-Hee Lee Ju-Yeon. Effects of DNA Methylation on the Structure of Nucleosomes. *Journal American Chemical Society*, 134, 2012. URL `https://doi.org/10.1021/ja210273w`.

Alejandro Caceres Juan-Gonzalez. Omic association studies with r and bioconductor. *CRC Press*, 2019.

Qeethara Kadhim. Artificial Neural Networks in Medical Diagnosis. *IJCSI International Journal of Computer Sciences Issues*, 8(2), 2011.

Ziv Porat Valery Krizhanovsky Uri Alon Karin, Amit Agrawal. Isogenic organisms age at different rates due to critical-slowing-down of damage removal. *bioRxiv*, 2018. URL `https://doi.org/10.1101/470500`.

Kinami Stephen Khamis-Hassan, Kiputro Cheruiyot. Application of k-nearest

Neighbor Classification in Medical Data Mining. *International Journal of Information and Ccommunication Technology Research*, 4, 2014.

James W Curtsinger Khazaeli, Wayne Van Voorhies. Longevity and metabolism in drosophila melanogaster. *Genetics*, 169, 2005.

Sunghwan Kim. Weighted k-means support vector machine for cancer prediction. *Springer Plus*, 2016. URL `10.1186/s40064-016-2677-4`.

Gulzar ZG Young SR Kobayashi, Absher DM. DNA Methylation Profiling Reveals Novel Biomarkers and Important Roles for DNA Methyltransferases in Prostate Cancer. *Genome Res*, 21, 2011.

Ahire Vijaya Kohad-Rashmee. Diagnosis of Lung Cancer using Support Vector Machine with ant Colony Optimization Technique. *International journal of advances in computer science and technology (IJACST)*, 13(11), 2014.

Kuchka. Bioscience in the 21 century. *Univeristy of Lehigh*, 2012. URL `https://www.lehigh.edu/~inbios21/PDF/Fall2012/Kuchka_11262012.pdf`.

Lin HC Hsiao JH Lenka, Tsai MH. Identification of Methylation-Driven, Differentially Expressed STXBP6 as a Novel Biomarker in Lung Adenocarcinoma. *Sci Rep*, 7, 2017.

Levenberg-Marquardt. An Algorithm for Least-squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, 1963.

Levenberg-Marquardt. A Method for the Solution of Certain Non-linear Problems in Least Squares. *Quarterly Journal of Applied Mathematics*, 1994.

Hong Chuan Jin Michael Chan Leygo, Marissa Williams. DNA Methylation as a Noninvasive Epigenetic Biomarker for the Detection of Cancer. *Disease Markers*, 2017. URL `https://doi.org/10.1155/2017/3726595`.

Maxwell Libbrecht. Machine learning in genetics and genomics. *Nat Rev Gen*, 2015.

Maher E Lim. DNA Methylation: a Form of Epigenetic Control of Gene Expression. *The Obstetrician Gynecologist*, 2010.

Jennifer Ro Scott Pletcher Linford, Ceyda Bilgir. Measurement of lifespan in drosophila melanogaster. *Journal of visualized experiments*, 71, 2013. URL `10.3791/50068`.

Tan Qihan Batra Richa List, Hauschild Anne-Christin. Classification of Breast Cancer Subtypes by Combining Gene Expression and DNA Methylation Data. *Journal of integrative bioinformatics*, 2014.

Hutchison K Calhoun VD Liu, Morgan M. A Study of the Influence of Sex on Genome Wide Methylation. *PLoS One*, 2010.

Hunter Fraser Lucia, Emberly Eldon. Factors Underlying Variable DNA Methylation in Human Community Cohort. *Proceedings of the national academy of sciences of the United States of America*, 109, 2012.

MacArthur. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic Acids Res*, 2017.

Mackay. Bayesian Interpolation. *Neural Computation*, 1992.

Brennan K Gevaert O Magzoub, Prunello M. The Impact of DNA Methylation on the Cancer Proteome. *PLos Comput Biol*, 15, 2016. URL `https://doi.org/10.1371/journal.pcbi.1007245`.

Chow PK Chung AY Mah, Thurnherr T. Methylation Profile Reveals Distinct Subgroup of Hepatocellular Carcinoma Patient with Poor Prognosis. *PLOS ONE*, 9, 2014.

Emi Ota-Machida Yu Han Malcolm-Brock, Craig Hooker. DNA Methylation Markers and Early Recurrence in Stage I Lung Cancer. *The New England Journal of Medicine*, 2008. URL `10.1056/NEJMoa0706550`.

Manolio. Finding the missing heritability of complex diseases. *Nature*, 2009.

Zulet Maria Angeles-Moreno-Aliaga Mari Mansego, Milagro Fermin. Differential DNA Methylation in Relation to Age and Health Risk Obesity. *International Journal of Molecular Sciences*, 2015.

Jung Marc. Aging and DNA Methylation. *BMC Biology*, 13, 2015.

Shah S Marioni, Harris R. The Epigenetic Clock and Telomere Length are Independently Associated with Chronological Age and Mortality. *International Journal of Epidemiology*, 2016.

Adams P McBryan. Handbook of Pharmacogenomics and Stratified medicine. *Journal of Cellular Biochemistry*, 2014.

McCarthy. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, 2008.

Ginder GD McGhee. Specific Methylation Sites in the Vicinity of the Chicken Beta Globin Genes. *Nature*, 1979.

Medical-dictionary. Omics. *The American heritage*, 2007. URL `https://medical-dictionary.thefreedictionary.com/omics`.

Lasheras Sanchez Menendez-Alvarez. Artificial Neural Networks Applied to Cancer Detection in a Breast Screening Programme. *Mathematical and Computer Modelling*, 52, 2012.

Merriam-Webster. Epigenomics. *dictionary*, 2017. URL `https://www.merriam-webster.com/dictionary/epigenomics`.

Jeffrey Craig Mikeska. DNA Methylation Biomarkers: Cancer and Beyond. *Genes (Basel)*, 22, 2014. URL `10.3390/genes5030821`.

Sam Lucas-Micahel Horan. Miles-Jefferson, Neil Pendleton. Comparison of a genetic algorithm neural network with logistic regression for predicting outcome after surgery for patients with nonsmall cell lung carcinoma. *Cancer*, 2000. URL `https://doi.org/10.1002/(SICI)1097-0142(19970401)79:7<1338::AID-CNCR10>3.0.CO;2-0`.

MIT. Mit genetics notes. 2013. URL `https://ocw.mit.edu/courses/biology/7-03-genetics-fall-2004/lecture-notes/lecture28.pdf`.

Benjamin Mullish. The implementation of omics technologies in cancer microbiome research. ecancer medical science. 2018. URL `https://doi.org/10.3332/ecancer.2018.864`.

Foulkes WD Narod. Brca1 and brca2: 1994 and beyond. *Nat Rev Cancer*, 2004.

Soultana Markopoulou-John R Gosney Julie Bryan Nikolaidis, Olaide Y Raji. DNA Methylation Biomarkers Offer Improved Diagnostic Efficiency in Lung Cancer. *Cancer Research*, 2012. URL `0.1158/0008-5472.CAN-12-2309`.

Yiyu Zou-Jose Nahum Galeas Makardhawj Shrivastave Caroline Hum Katalin Susztak Niraj-Shenoy, Nishanth Vallumsetla. Role of DNA Methylation in Renal Cell Carcinoma. *Journal of Hematology and Oncology*, 2015. URL `https://doi.org/10.1186/s13045-015-0180-y`.

Moya A Domingo E Novella, Duarte E. Exponential Increases of RNA Virus Fitness During Large Population Transmissions. *Journal of Virology*, 1995.

Assenov Yassen Hullein Jennifer Zucnick Manuela Oakes, Claus Rainer. Heteropgeneity and Evolution of DNA Methylation in Chronic Lympocytic Leukemia. *Blood*, 122, 2013.

Florence Odekunle. Genomics in personalized medicine. *International journal of health science and research*, 2017. URL 10.1007/s00439-011-1028-3.

Kenneth Offit. New genomic, old lessons. *Human genetics*, 2011. URL 10.1007/s00439-011-1028-3.

Bandit S Orawan, Pandawee S. Application of Artificial Neural Networks on Growth Prediction of Staphylococcus in Milk. *International Food Research Journal*, 2016.

Vinvent Vu Oustimov. Artificial neural networks in the cancer genomics frontier. *Translational cancer research*, 2014. URL http://tcr.amegroups.com/article/view/2647/html.

Paul Schmidt Paaby. Dissecting the genetics of longevity in drosophila melanogaster. *Landes Bioscience*, 2009. URL http://www.landesbioscience.com/journals/fly/article/7771.

Rebecca Fry Paige-Bommarito. The role of DNA Methylation in Gene Regulation. *Toxicoepigenics*, 2019. URL https://doi.org/10.1016/B978-0-12-812433-8.00005-8.

Minatu Behrouz Parvin-Hamid, Alizadeh Hoseinali. A Modification on k-nearest Neighbor Classifier. *Global Journal of Computer Science and Technology*, 10, 2010.

Qu L Clark S Patterson, Molloy k. DNA Methylation: Bisulphite Modification and Analysis. *Journal of Visualized Experiments*, 2011a. URL 10.3791/3170.

Qu Wenjia Clark Susan Patterson, Molloy Laura. DNA Methylation: Bishulphite Modification and Analysis. *Journal of Visualized Experiments*, 2011b.

Kemp Kernstine Pfeifer. DNA Methylation Biomarkers in Lung Cancer Diagnosis: Closer to Practical Use? *Translational Cancer Research*, 2017. URL http://dx.doi.org/10.21037/tcr.2017.01.17.

Phillips. The Role of Methylation in Gene Expression. *Nature Education*, 1, 2008.

Guochen Yuang Pinello, Giouse Lo Bosco. Applications of alignment-free methods in epigenomics. *Briefing in bioinformatics*, 2014.

Guochen Yuang Pinello, Giouse Lo Bosco. National health service (uk). 2018. URL `www.nhs.uk`.

Diorio C Pouliot, Labrie Y. The Role of Methylation in Breast Cancer Susceptibility and Treatments. *International Journal of Cancer Research and Treatment*, 2015.

ProteinTech. Epigenic modifications – dna methylation. 2017. URL `https://www.ptglab.com/news/blog/epigenetics-modifications-dna-methylation/`.

Joshua E. Stern Linda Wiens Qinghua-Feng, Stephen E. Hawes. DNA Methylation in Tumor and Matched Normal Tissues from Non-small Cell Lung Cancer Patients. *Cancer Epidemiology, Biomarkers Prevention*, 2008. URL `10.1158/1055-9965`.

Down TA Hawa MI Rakyan, Beyan H. Identification of Type 1 Diabetes-associated DNA Methylation Variable Positions that Precede Disease Diagnosis. *PLos Genet*, 2011.

Razin. DNA Methylation and Gene Expression. *Microbiology Reviews*, 55, 1991. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC372829/`.

Vucic Emily Rowbotham, Marshal Erin. Epigenetic Changes in Aging and Age-related Disease. *Journal of Aging Science*, 3, 2014.

Ficz Gabriella Oxley David Tomaszewski Bartlomiej Sabine, Cameron Kerry. The Aging Brain: Effects on DNA Repair and DNA Methylation on Mice. *Genes*, 2017.

Schuebeler. Function and Information Content of DNA Methylation. *Nature*, 2012. URL `10.3791/3170`.

Langevin Scott. Leukocyte-adjusted Epigenome-wide Association Studies of Blood from Solid Tumor Patients. *University of Cincinnati College of Medicine. Department of environmental health*, 2015.

Mercedes Serrano. The Cerebellum: Disorders and Treatment. *Handbook of Clinical Neurology*, 155, 2018. URL https://doi.org/10.1016/B978-0-444-64189-2.00015-9.

Hira Mubeen Shahid-Raza. Genetic markets: importance, uses and applications. *International journal of scientific research publications*, 2016. URL http://www.ijsrp.org/research-paper-0316/ijsrp-p5137.pdf.

Shawe. Support Vector Machine. *Cambridge University Press*, 2000.

AlYakoob Smaoui. Analyzing the Dynamics of Cellular Flames using Karhunen-loeve Decomposition and Autoassociative Neural Network. *Journal of Sientific Computing*, 2003.

Kaleigh Smith. Genetic polyporphism and snps. *Genotyping, haplotype assembly problem, Haplotype map, functional genomics and proteomics*, 2002.

Andrew Read Strachan. Human molecular genetics. *Taylor and Francis group*, 2018. URL https://www.routledge.com/Human-Molecular-Genetics-5th-Edition/Strachan-Read/p/book/9780815345893.

Kang Kyung-In Kang Moon-Young Cho Sung-Hoon, U-Yeol Park. Application of Support Vector Machines in Assessing Conceptual Cost Estimates. *Journal of computing in civil engineering*, 2007.

Suman Lakhanpaul Tabassum-Jehan. Single nucleotide polyphorism (snp) – methods and applications in plant genetics. *Indian journal of biotechnology*, 2006.

Izumiyama N Furugori E-Sawabe M Arai T Esaki Y Mafune K Kammori M Fujiwara M Takubo, Nakamura K. Telomere Shortening with Aging in Human

Liver. *The Journals of Gereontology Series A; Biological Science and Medical Sciences*, 2014.

Nadeem Tariq. Breast cancer detection using artificial neural networks. *Journal of molecular biomarkers*, 2018. URL 0.4172/2155-9929.1000371.

Xiwei Wu Kemp H Kernstine Arthur D Riggs Gerd P Pfeifer Tibor-Rauch, Zunde Wang. DNA Methylation Biomarkers for Lung Cancer. *Tumor Biology*, 33, 2012. URL 10.1007/s13277-011-0282-2.

Takasaki T Ikeda Noriaki S Tsuji, Ishiko A. Estimating Age of Humans Based on Telomere Shortening. *Forensic Science International*, 2002.

Jose Florez Udler, Marc McCarthy. Genetic risk scores for diabetes diagnosis and precision medicine. *Endocrine*, 2019. URL https://doi.org/10.1210/er.2019-00088.

Juana Fernandez-Tajes Ulloa. Bivalve omics: State of the art and potential applications for the biomonitoring of harmful marine compounds. 2013. URL 10.3390/md11114370.

Martinez-Chantar Maria Luz Mato Jose Lu Shelly Varela-Rey, Woodhoo Ashwin. Alcohol, DNA Methylation and Cancer. *Alcohol Research Current Reviews*, 35, 2013.

Waddington. The epigenotype. *Endeavour*, 1942.

Zhen Wang and John Moult. Snps, protein structure, and disease. *Human mutation*, 2000. URL http://moult.ibbr.umd.edu/pdfs/SNPsproteinstructureanddisease.pdf.

Samini Goli Warton, Mahom Kate. Methylated Circulating Tumor DNA in Blood: Power in Cancer Prognosis and Response. *Endocrine Related Cancer Journal*, 23, 2016.

WHO. Genomics and world health: Report of the advisory committee on health research. *World Health Organization*, 2002. URL https://www.who.int/genomics/geneticsVSgenomics/en/.

Petra A. Link Jihnhee Yu Woloszynska-Read, Smitha R James. DNA Methylation-dependent Regulation of boris/ctcfl Expression in Ovarian Cancer. *Cancer Immunology Research*, 2007. URL https://cancerimmunolres.aacrjournals.org/content/canimmarch/7/1/21.

Wei Zhang Bixi Zhong Yanda Li Xianglin-Zhang, Huan Fang. Ribosomal dna methylation as stable biomarkers for detection of cancer in plasma. 2019. URL https://doi.org/10.1101/651497.

Dazhong Zhuang Xylinas, Melanie Hassler. An Epigenomic Approach to Improving Response to Neoadjuvant Cisplatin Chemotherapy in Bladder Cancer. *Biomolecules*, 6, 2016. URL https://doi.org/10.3390/biom6030037.

Jinhua Liu Yonghong-Zhang, Spphie Petropoulous. The Signature of Liver Cancer in Immune Cells DNA Methylation. *Clinical Epigenetics*, 8, 2018. URL https://clinicalepigeneticsjournal.biomedcentral.com/articles/10.1186/s13148-017-0436-1.

Howard C Yuval. Principles of DNA Methylation and Their Implicaitons for Biology and Medicine. *The Lancet*, 2018.

Thuraisingham Bhavani Zhouyan, Murat Kantarcioglu. Adversarial Support Vector Machine Learning. 2012.

Lee SH Ng E Zhuang, Jones A. The Dynamics and Prognostic Potential of DNA Methylation Changes at Stem Cell Gene Loci in Women's Cancer. *PLos Genet*, 8, 2012.