Deep Learning for Phenotypic Prediction

Sapienza

Buzatu Ionelia

October 2019

Abstract

Artificial neural networks are well suited for problems with a large signal-to-noise ratio when the primary objective is prediction rather than a selection of relevant predictor variables. SNPs are single base-pair changes in the DNA sequence that occur with high frequency in the human genome. SNPs are typically used as markers of a genomic region, with the large majority of them having a minimal impact on biological systems. SNPs can have functional consequences, however, causing amino acid changes, changes to mRNA transcript stability, and changes to transcription factor binding affinity. Prior studies have used SNPs to identify statistically significant genetic variants in case-control studies. This study demonstrates the application of deep learning for phenotypic prediction. In this study, one of the models has a feed-forward architecture, because this kind of architecture is suitable for genetic prediction problems when there are no special relations among the input data features. The first model consists of four fully connected layers which predicts the systolic and diastolic blood pressure. The second model is a classifier for obesity. The starting point of the neural network is an artificial neuron, which takes as input a vector. The weights are the parameters of the model that are learned during training. In genomics, the input might be a DNA sequence, in which the nucleotides A, C, T and G are encoded as [1,0,0,0], [0,1,0,0], [0,0,1,0] and [0,0,0,1]. In this study, the model takes as input a vector of encoded single-nucleotide polymorphism. Each SNP was encoded as 0 for the minor allele (bb), 1 for the heterozygous allele (Ab), and 2 for the major allele (AA). Though other forms of SNPs encodings have been exploited, those have yielded the best results for phenotypic prediction. The output of the neural network is the prediction of interest. Two different models have been trained for blood pressure prediction. The feed-forward model has a mean square error of 15 and the convolutional model has a mean square error of 13. Though no pattern in the genomic sequence has been found, the convolutional network led to a better accuracy and its training was much faster than the feed-forward, due to lesser amount of parameters compared to the feed-forward. The limit of this pipeline is that does not account for low trait heritability, such as environmental effects and chance effects that contribute to the phenotypic differences of interest. The pipeline demonstrates that using SNPs as single input is not enough to make predictions for common phenotypes such as blood pressure or obesity.

Acknowledgements

I want to thank my thesis advisor, Professor Aris Anagnostopoulos. Also, I want to thank Professor Ioannis Chatzigiannakis for giving me the opportunity for this thesis and in particular, I thank Manuel Del Verme.

Contents

List of Figures

Chapter 1

Introduction

The data in this study comes from the UK Biobank international health resource. UK Biobank aims to improve the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses including cancer, heart diseases, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and forms of dementia. It is following the health and well-being of 500,000 volunteer participants and provides health information, which does not identify them, to approved researchers in the UK and overseas, from academia and industry. [1]. The data has been generated by Affymetrix. The Affymetrix platform prints short DNA sequences as a spot on the chip that recognizes a specific SNP allele. Alleles (i.e. nucleotides) are detected by differential hybridization of the sample DNA. Genotyping has been undertaken on all 500,000 participants. The following chapters demonstrates the application of deep learning on phenotypic predictions exploting all the single nucleotide polymorphisms (SNPs) of each participant.

1.1 SNPs

The modern unit of genetic variation is the single nucleotide polymorphism or SNP. SNPs are single base-pair changes in the DNA sequence that occur with high frequency in the human genome [21]. For the purposes of genetic studies, SNPs are typically used as markers of a genomic region, with the large majority of them having a minimal impact on biological systems. SNPs can have functional consequences, however, causing amino acid changes, changes to mRNA transcript stability, and changes to transcription factor binding affinity [8]. SNPs are by far the most abundant form of genetic variation in the human genome. SNPs are notably a type of common genetic variation; many SNPs are present in a large proportion of human populations [2].

While mutation is the change in DNA sequence by addition of a nucleotide or a whole sequence, replacement like in SNP (point mutation, also known as genetic variants), deletion of a nucleotide or sequence, SNP is just a type of mutation. Polymorphism means exhibitng different forms. when a DNA strand undergoes SNP, it can have four types of sequences due to four types of nucleotide bases. Also, these different forms may result in different expressions. Commonly occurring SNPs lie in stark contrast to genetic variants that are implicated in more rare genetic disorders, such as cystic fibrosis [12]. These conditions are largely caused by extremely rare genetic variants that ultimately induce a detrimental change to protein function, which leads to the disease state. Variants with such low frequency in the population are sometimes referred to as mutations, though they can be structurally equivalent to SNPs - single base-pair changes in the DNA sequence. In the genetics literature, the term SNP is generally applied to common single base-pair changes, and the term mutation is applied to rare genetic variants. SNPs typically have two alleles, meaning within a population there are two commonly occurring base-pair possibilities for a SNP location. The frequency of a SNP is given in terms of the minor allele frequency or the frequency of the less common allele. For example, a SNP with a minor allele (G) frequency of 0.40 implies that 40% of a population has the G allele versus the more common allele (the major allele), which is found in 60% of the population.

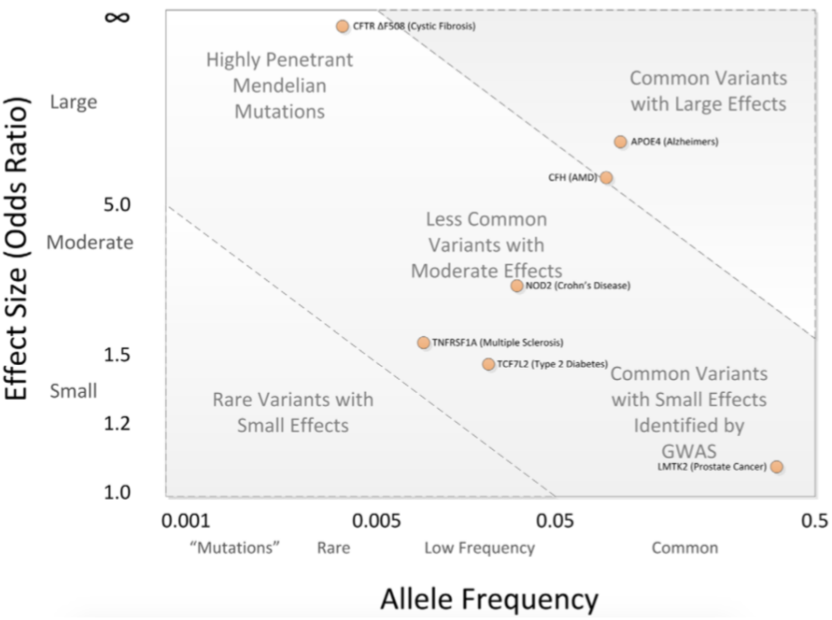


Figure 1.1: Spectrum of Disease Allele Effects. Disease associations are often conceptualized in two dimensions: allele frequency and effect size. Highly penetrant alleles for Mendelian disorders are extremely rare with large effect sizes (upper left), while most GWAS findings are associations of common SNPs with small effect sizes (lower right). The bulk of discovered genetic associations lie on the diagonal denoted by the dashed lines.

[18]

1.2 Common Disease Common Variant Hypothesis

The genetic mechanisms that influence common disorders are different from those that cause rare disorders [10]. This studies provides an opportunity to investigate the impact of common variants on complex disease; The idea that common diseases have a different underlying genetic architecture than rare disorders, coupled with the discovery of several susceptibility variants for common disease with high minor allele frequency (including alleles in the apolipo-protein E or APOE gene for Alzheimers disease [11] and PPARg gene in type II diabetes [12]), led to the development of the common disease/common variant (CD/CV) hypothesis [13]. This hypothesis states simply that common disorders are likely influenced by genetic variation that is also common in the population. There are several key ramifications of this for the study of complex diseases. First, if common genetic variants influence disease, the effect size (or penetrance) for any one variant must be small relative to that found for rare disorders. For example, if a SNP with 40% frequency in the population causes a highly deleterious amino acid substitution that directly leads to a disease phenotype, nearly 40% of the population would have that phenotype. Thus, the allele frequency and the population prevalence are completely correlated. If, however, that same SNP caused a small change in gene expression that alters risk for a disease by some small amount, the prevalence of the disease and the influential allele would be only slightly correlated. As such, common variants almost by definition cannot have high penetrance. Secondly, if common alleles have small genetic effects (low penetrance), but common disorders show heritability (inheritance in families), then multiple common alleles must influence disease susceptibility. For example, twin studies might estimate the heritability of a common disease to be 40%, that is, 40% of the total variance in disease risk is due to genetic factors. If the allele of a single SNP incurs only a small degree of disease risk, that SNP only explains a small proportion of the total variance due to genetic factors. As such, the total genetic risk due to common genetic variation must be spread across multiple genetic factors. These two points suggest that traditional family-based genetic studies are not likely to be successful for complex diseases, prompting a shift toward population-based studies.

The frequency with which an allele occurs in the population and the risk incurred by that allele for complex diseases are key components to consider when planning a genetic study, impacting the technology needed to gather genetic information and the sample size needed to discover statistically significant genetic effects. The spectrum of potential genetic effects is sometimes visualised and partitioned by effect size and allele frequency (figure above). Genetic effects in the upper right are more amenable to smaller family-based studies and linkage analysis, and may require genotyping relatively few genetic markers. Effects in the lower right are typical of findings from GWAS, requiring large sample sizes and a large panel of genetic markers. Effects in the upper right, most notably CFH, have been identified using both linkage analysis and GWAS. Effects in the lower left are perhaps the most difficult challenge, requiring genomic sequencing of large [18] samples disease.

In the last years the common disease/common variant hypothesis has been tested for a variety of common diseases, and while much of the heritability for these conditions is not yet explained, common alleles certainly play a role in susceptibility. The National Human Genome Institute GWAS catalog (http:// www.genome.gov/gwastudies) lists over 3,600 SNPs identified for common diseases or traits, and in general, common diseases have multiple susceptibility alleles, each with small effect sizes (typically increasing disease risk between 1.22 times the population risk) [14]. These results support that for most common diseases, the CD/CV hypothesis is true, though it should not be assumed that the entire genetic component of any common disease is due to common alleles only.

1.3 Linkage Disequilibrium

Linkage disequilibrium (LD) is a property of SNPs on a contiguous stretch of genomic sequence that describes the degree to which an allele of one SNP is inherited or correlated with an allele of another SNP within a population and it is one of the critical factors affecting genome wide association studies (GWAS).

African-descent populations are the most ancestral and have smaller regions of LD due to the accumulation of more recombination events in that group. European-descent and Asian/descent populations were created by founder events (a sampling of chromosomes from the African population), which altered the number of founding chromosomes, the population size, and the generational age of the population. These populations on average have larger regions of LD than African-descent groups.

For the purposes of genetic analysis, LD is generally reported in terms of r2, a statistical measure of correlation. High r2 values indicate that two SNPs convey similar information, as one allele of the first SNP is often observed with one allele of the second SNP, so only one of the two SNPs needs to be genotyped to capture the allelic variation. There are dependencies between these two statistics; r2 is sensitive to the allele frequencies of the two markers, and can only be high in regions of high D.

1.4 What is GWAS

Genome-wide association studies (GWAS) have evolved over the last ten years into a powerful tool for investigating the genetic architecture of human disease. GWAS studies use SNPs to find loci affecting a phenotype and SNP allele frequencies in two populations with different phenotypes. GWAS also find chromosome locations where sequences differences affect the phenotype, thus not necessarily finding the causal sequence differences but just finding the place were we would look for the causal sequence differences.

A central goal of human genetics is to identify genetic risk factors for common, complex diseases such as schizophrenia and type II diabetes, hypertention and for rare Mendelian diseases such as cystic fibrosis and sickle cell anemia. Genome-wide association study or GWAS measures and analyzes DNA sequence variations from across the human genome in an effort to identify genetic risk factors for diseases that are common in the population. The ultimate goal of GWAS is to use genetic risk factors to make predictions about who is at risk and to identify the biological underpinnings of disease susceptibility for developing new prevention and treatment strategies. GWAS when applied to more common disorders, like heart disease or various forms of cancer, linkage analysis has not fared as well. This implies the genetic mechanisms that influence common disorders are different from those that cause rare disorders [22]. The end result of GWAS under the common disease/common variant hypothesis is that a panel of 500,000 to one million markers will identify common SNPs that are associated to common phenotypes.

The de facto analysis of genome-wide association data is a series of single-locus statistic tests, examining each SNP independently for association to the phenotype. The statistical test conducted depends on a variety of factors, but first and foremost, statistical tests are different for quantitative traits versus case/control studies. [19] GWAS performs single-locus and multi-locus analysis. Most GWAS focus on the

detection of main effects by using an allele- or genotype-based test for each single-nucleotide polymorphism (SNP) separately. However, the identified genetic effects tend to be moderate and explain only a small fraction of the overall heritability [7]. One of the early successes of GWAS was the identification of the Complement Factor H gene as a major risk factor for age related macular degeneration or AMD [9].
A recent GWAS revealed DNA sequence variations in several genes that have a large influence on warfarin dosing [5]. Cystic fibrosis (and most rare genetic disorders) can be caused by multiple different genetic variants within a single gene. Because the effect of the genetic variants is so strong, cystic fibrosis follows an autosomal dominant inheritance pattern in families with the disorder. One of the major successes of human genetics was the identification of multiple mutations in the CFTR gene as the cause of cystic fibrosis [12]. This was achieved by genotyping families affected by cystic fibrosis using a collection of genetic markers across the genome, and examining how those genetic markers segregate with the disease across multiple families. This technique, called linkage analysis, was subsequently applied successfully to identify genetic variants that contribute to rare disorders like Huntington disease [13].

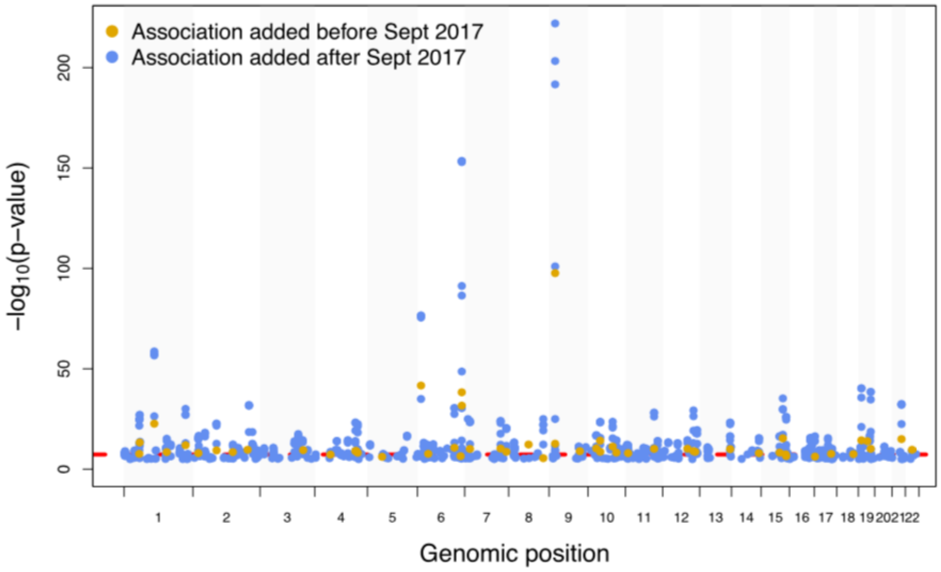


Figure 1.2: GWAS Catalog associations for coronary artery disease plotted across all chromosomes. Associations added after 2017 represented in blue, previous associations in orange. The dashed red line indicates genome-wide significance at P-value = 5 108 [4].

1.5 Beyond GWAS - Deep Learning

The benefits of multi-locus analysis are that it performs an unbiased analysis for interactions within the selected set of SNPs. It is also far more computationally and statistically tractable than analyzing all possible combinations of markers [19].
Machine learning algorithms provide several alternatives for performing multi-SNP analyses and artificial neural networks are well suited for problems with a large signal- to-noise ratio when the primary objective is prediction rather than selection of relevant predictor variables. [15]. Deep learning methods are a class of machine learning techniques capable of identifying highly complex patterns in large datasets.
Machine learning tasks fall within two major categories: supervised and unsupervised. In supervised learning, the goal is predicting the label (classification) or response (regression) of each data point by using a provided set of labelled training examples. In unsupervised learning, such as clustering and principal component analysis, the goal is learning inherent patterns within the data themselves.
Deep learning can build predictive models of gene expression from genotype data and can be used for studying the splicing-code model as well as for the identification of long noncoding RNAs. Deep learning has also been used for the interpretation of regulatory control in single cells; for example, the detection of DNA methylation in single cells [3]. In cancer genomics, deep learning can extract the high-level features between combinatorial somatic mutations and cancer types and learn prognostic information from multicancer datasets.
The starting point of a neural network is an artificial neuron, which takes as input a vector of real values and computes the weighted average of these values followed by a nonlinear transformation, which can be a simple threshold7. The weights are the parameters of the model that are learned during training. The power of neural networks stems from individual neurons being highly modular and composable, despite their simplicity. The output of one neuron can be directly fed as input into other neurons. By composing neurons together, a neural network is created.
The input into a neural network is typically a matrix of real values. In genomics, the input might be a DNA sequence, in which the nucleotides A, C, T and G are encoded as [1,0,0,0], [0,1,0,0], [0,0,1,0] and [0,0,0,1]. Neurons that directly read in the data input are called the first, or input, layer. Layer two consists of neurons that read in the outputs of layer one, and so on for deeper layers, which are also referred to as hidden layers. The output of the neural network is the prediction of interest, e.g., whether the input DNA is an enhancer.
Researchers have observed that performance can improve with very deep networks (more than 100 layers). In most genomics applications, fewer than five layers are sufficient. Even relatively shallow networks can still have millions of parameters, and the most important factor that determines the success of a model is the availability of a large corpus of labeled training data. In this study the models do not have more than four layers.

Chapter 2

Related Work

2.0.1 On GWAS

Genome-Wide Association Studies (GWAS) are used to identify statistically significant genetic variants in case-control studies. The main objective is to find single nucleotide polymorphisms (SNPs) that influence a particular phenotype (i.e. disease trait). GWAS typically use a p-value threshold of 5 108 to identify highly ranked SNPs. While this approach has proven useful for detecting disease-susceptible SNPs, evidence has shown that many of these are, in fact, false positives. Consequently, there is some ambiguity about the most suitable threshold for claiming genome-wide significance. Many believe that using lower p-values will allow us to investigate the joint epistatic interactions between SNPs and provide better insights into phenotype expression. One example that uses this approach is multifactor dimensionality reduction (MDR), which identifies combinations of SNPs that interact to influence a particular outcome. [6]. In this paper [16], the authors use the principal component analysis method applied to GWAS. This paper [16] reports genetic association of blood pressure (systolic, diastolic, pulse pressure) among UK Biobank participants of European ancestry with independent replication in other cohorts, and robust validation of 107 independent loci. They identify new independent variants at 11 previously reported blood pressure loci. In combination with results from a range of in silico functional analyses and wet bench experiments, their findings highlight new biological pathways for blood pressure regulation enriched for genes expressed in vascular tissues and identify potential therapeutic targets for hypertension. The authors of this paper makes use of exome content, along with the GWAS data. Elevated blood pressure is a strong, heritable [24]. and modifiable driver of risk for stroke and coronary artery disease and is a leading cause of global mortality and morbidity [14]. Genome-wide association study (GWAS) meta-analyses, and analyses of custom or exome content, had identified and replicated genetic variants of mostly modest or weak effect on BP at over 120 loci711. They also report association analyses between BP traits and genetic variants among 140,000 participants in UK Biobank, a prospective cohort study of 500,000 men and women aged 4069 years with extensive baseline phenotypic measurements, stored biological samples and follow-up by electronic health record linkage. Their study provides new biological insights into blood pressure regulation. Their GWAS analyses consists of systolic (SBP) and diastolic (DBP) blood pressure and of pulse pressure (PP) using single-variant linear regression under an additive model, based on 9.8 million single-nucleotide variants (SNVs) with minor allele frequency (MAF) 1% and imputation quality score. Their results consists of thirty-two validated novel loci findings. In total they validate 107 loci. Of the 107 validated loci, 24 were reported for association with SBP as the primary trait (most significant from combined meta-analysis), 41 were reported for DBP and 42 were reported for PP, although many loci were associated with more than one BP trait. They perform quality control first. The model classifies a loci of being a valid candidate for blood pressure association or not. The binary case study model is: BP SNV + sex + age + age2 + BMI + UKBB vs. UKBL + top 10 PCs.

Their findings are mostly common variants, with modest effect sizes. They report that the lack of rare variant discovery could also be due to the challenge of detecting rare variants from imputed data in microarray data [16].

2.0.2 On Deep learning

Predicting phenotypes from genetic data is a major area of interest of deep learning. A first step in performing these types of predictions is to specify what genetic variants are present in an individual genome. This problem has been addressed by DeepVariant, which applies a CNN to make variant calls from short-read sequencing. The method treats DNA alignments as an image with a performance that appears to exceed that of standard variant callers. Another study has been recently reported the extension of DeepSEA to the study of regulatory variants in autism spectrum disorder [30]. The same team has published ExPecto, the ab initio prediction of gene expression levels and variant effects from sequences from more than 200 tissues and cell types. Also recently, Sundaram et al. have trained a DNN by using hundreds of common variants from population sequencing of nonhuman primate species to identify pathogenic variants in rare human diseases [25].

Another study [6] that applies deep learning to understand epistatic interactions proposed a novel framework, based on nonlinear transformations of combinatorically large SNP data, using stacked autoencoders, to identify higher-order SNP interactions. Autoencoders are a type of neural network architecture. They focus on the challenging problem of classifying preterm births. Evidence suggests that this complex condition has a strong genetic component with unexplained heritability reportedly between 20%-40%. This claim is substantiated using a GWAS data set. Latent representations from original SNP sequences are used to initialize a deep learning classifier before it is fine-tuned for classification tasks (term and preterm births). Their findings show that important information pertaining to SNPs and epistasis can be extracted from 4666 raw SNPs generated using logistic regression (p-value=5 103) and used to fit a deep learning model.

The real breakthrough here came from a team at Google and Verily. They invented a tool called DeepVariant for identifying DNA mutations. They did this by converting DNA sequences to images and feeding them through a convolutional neural network. That was a real eye-opener. It suggested that all kinds of problems in population genetics could also be treated in this way.

Chapter 3

Data Description

3.1 Files Description

3. Dataset description (Describe the various files, geno, phenotype and how to process them)

The ukb8627.tab[21GB] consists of the phenotype data of all the participants that have been taking part since the UK biobank started to collect their data. Over the years 502638 participants data has been collected but Y withdraw consent to participate to the project. Currently the total genomes are 486383. The 'f.eid' is the id of the patient and 10951 columns each rapresant a Field. data showcase, also described in the readme.

The ukb_category.txt[7.1K] describes categories ids with hierarchy.

The ukb_field.txt[234K] Description of fields ids with parent category. Field with biological similarities are grouped under a specific category. The below table shows all the fieald under the category CANCER

The genetic data consists of:
- log2ratios
- b-allele-frequencies

File names	
Calls BIM	ukb_snp_chrN_v2.bim
Calls FAM	ukbA_cal_v2_sP.fam
Relatedness	ukbA_rel_sP.txt
Imputation BGEN	ukb_imp_chrN_v3.bgen
Imputation BGI	ukb_imp_chrN_v3.bgen.bgi
Imputation MAF+info	ukb_mfi_chrN_v3.txt
Imputation sample chr1-22	ukbA_imp_chrN_v3_sP.sample
Imputation sample chrX	ukbA_imp_chrX_v3_sP.sample
Imputation sample chrXY	ukbA_imp_chrXY_v3_sP.sample
Haplotypes BGEN	ukb_hap_chrN_v2.bgen
Haplotypes BGI	ukb_hbg_chrN_v2.bgi
Intensity	ukb_int_chrN_v2.bin
Confidences	ukb_con_chrN_v2.txt

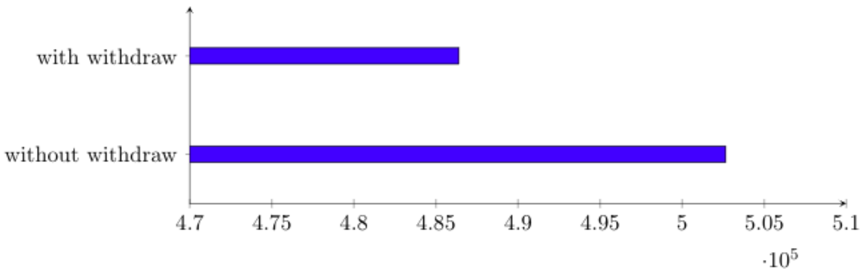


Figure 3.1: Participants

Field ID	Description
40021	Cancer record origin
40009	Reported occurrences of cancer
40005	Date of cancer diagnosis
40008	Age at cancer diagnosis
40006	Type of cancer: ICD10
40016	Type of cancer: ICD10 addendum
40013	Type of cancer: ICD9
40011	Histology of cancer tumour
40017	Type of cancer: ICD9 addendum
40012	Behaviour of cancer tumour
40019	Cancer report format

These files contain the B Allele Frequency baf and Log2Ratio log2r transformed intensitiy values for performing CNV calling. There is a separate file for baf and log2r per chromosome. These are plaintext files with space separated columns. The rows corresond to markers ordered as the calls BIM file and the columns correspond to samples ordered as the calls FAM file.

a) `Calls`
The genotype data calls are in binary PLINK format (.bed, .bim, .fam) [20] The BIM file determines the order of markers in the calls and all of the other genotype data sets. The SNP id is the rsid where it is available or the Affymetrix SNP id otherwise. The positions are in GRCh37 coordinates. The FAM file contains the id of the participants and determines the order of samples in the calls and all of the other genotype data sets. NOTE: that the fam is the same for all beds, so take as reference `ukb_cal_chr1_v2.fam`

b) indIDS[3.7M] : Ids of the individuals as found in the FAM file.

c) `ukb_snp_bim`
Bim files for the 'calls' data above

d) `Confidencies`
These files contain the Affymetrix 'confidence' that a genotype belongs to the call cluster. This is a plaintext file with space spearated columns. Values are in the range 0-1 with 0 being most confident. Missing values are represented by -1. The

order of markers and Samples are given by the BIM and FAM files. e) `Haplotypes`

Phased haplotypes in BGEN format. The sample file lists the order of the samples in the .bgen files.

f) `Imputations`
contain imputed genotype of the individuals. The imputed genotype calls are in BGEN v1.2 format (.bgen, .sample, .bgi). The sample file lists the order of the samples in the .bgen files. The sample file includes the 'Sex' field for every sample (corresponding to 'Inferred.Gender' in the SampleQC file). The list of variants in the files can be found with bgenix [17]. The 1st column of the marker list file is alternate ids which is a unique identifier for each marker. For markers in the genotype data set alternate ids is the genotype marker id (rs id in SNPQC file). The second column is rsid or the reference panel marker id, it is not guaranteed to be unique. The alleles in the imputation are aligned with REF/ALT, first allele is the ref allele on the fwd strand.

g) `Intensities`
contains A/B Intensity values measured by Affymetrix. Two intensity values A/B for each genotype (marker, indivudal) pair each represented as a 4-byte float. The set of A,B values for each marker are ordered consecutively by sample (analagous to a matrix with rows=SNPs and columns=Samples) e.g. SNP1SAMPLE1A SNP1SAMPLE1B SNP1SAMPLE2A SNP1SAMPLE2B ... Missing pairs of intensities are represented by -1 -1. The order of the markers and Samples are given by the BIM and FAM files with the calls. Affymetrix transform the A,B values into 'contrast' and 'strength' for their calling algorithm. The values are: contrast (X) = log2(A/B) strength (Y) = log2(AB)/2

h) `The relatedness`
file lists the pairs of individuals related up to the third degree in the data set. It is a plaintext file with space separated columns.

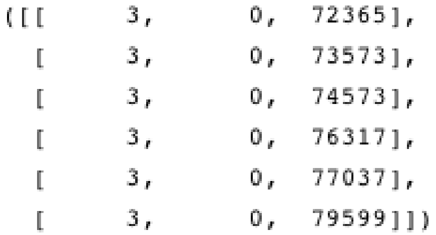
ID1	string	Sample id for individual 1 in related pair.
ID2	string	Sample id for individual 2 in related pair.
HetHet	numeric	Fraction in common of heterozygous genotype.
IBS0	numeric	Fraction sharing zero alleles (output from KING software).
Kinship	numeric	Estimate of the kinship coefficient output from KING software.

3.2 Data Extraction and Normalization

The date was extracted from the binary files. The fam file is used to get the index of each patient, which then is used to extract the SNPs of the patient from the bed file. Each SNP has an index in the bim file. The bim file fields are the following: chromosome, genetic distance, basepair distance.
The bed file consists of the snp value: 0 for minor allele, 1 for the hetherozygous allele and 2 major allele. To have values between 0 and 1, th data has been normilized accoding to the Z-score Normalization/standardization. In this technique, the values

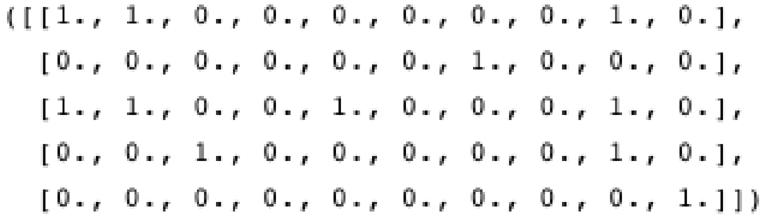
are normalized based on the mean and standard deviation. The formula is below:

$$z = \frac{x - \bar{x}}{\sigma}$$



```
([[      3,       0,  72365],
  [      3,       0,  73573],
  [      3,       0,  74573],
  [      3,       0,  76317],
  [      3,       0,  77037],
  [      3,       0,  79599]])
```

Figure 3.2: An example of the bim file fields.



```
([[1., 1., 0., 0., 0., 0., 0., 0., 1., 0.],
  [0., 0., 0., 0., 0., 0., 1., 0., 0., 0.],
  [1., 1., 0., 0., 1., 0., 0., 0., 1., 0.],
  [0., 0., 1., 0., 0., 0., 0., 0., 1., 0.],
  [0., 0., 0., 0., 0., 0., 0., 0., 0., 1.]])
```

Figure 3.3: An example of the bed file array with SNPs encoding.

Chapter 4

Methodology

4.1 Deep Learning Pipeline

There are three common families of architectures for connecting neurons into a network: feed-forward, convolutional and recurrent. Feed-forward is the simplest architecture. Every neuron of layer i is connected only to neurons of layer i + 1, and all the connection edges can have different weights. Feed-forward architecture is suit- able for generic prediction problems when there are no special relations among the input data features.

In a convolutional neural network (CNN), a neuron is scanned across the input matrix, and at each position of the input, the CNN computes the local weighted sum and produces an output value. This procedure is very similar to taking the position weight matrix of a motif and scanning it across the DNA sequence. CNNs are useful in settings in which some spatially invariant patterns in the input are expected.

Recurrent neural networks (RNN) are designed for sequential or time-series data. At each point in the sequence, a neural network, which could be feed-forward or convolutional, is applied to generate an internal signal, which is also fed to the next step of the RNN. Hidden layers of the RNN can be viewed as memory states that retain information from the sequence previously observed and are updated at each time step.

In addition, there are neural network architectures used for unsupervised learning. The most common are autoencoders that perform nonlinear dimentionality reduction, in contrast to principal component analysis, which is linear. In an autoencoder, the output is set to be the input, and the network is encouraged to find a low-dimensional space that compresses the original information and reconstructs the inputs.

Training a neural network starts with a labeled dataset of (Xi, Yi), where each Xi is the ith input, and Yi is its output label. Each training point Xi is fed into the network, and the networks output is evaluated against the true label Yi to produce a loss L(Y i, Yi). Loss is the sum of the errors made for each example. The lower the loss, the better the model. Commonly used loss functions include squared error and cross-entropy. Squared error measures the difference between predicted and actual output values, which is especially relevant when the output is a continuous value. Cross-entropy measures the difference between two probability distributions over the same set of underlying events or classes, as is appropriate when the output is categorical.

In this study the mean square error is used to calculate the error of the blood pressure model and the cross/entropy is to calculate the error of the obesity classifier.

4.1.1 Feed-Forward Model

A multi-layer feed-forward neural network is implemented in this study based on the formal definitions in [15]. Computational units (neurons) take as input (x1 , x2 , . . . xn), and a +1 intercept term, and generate outputs hW,b(x) = f(W x)=f(i=1Wixi+b),where f : $R \mapsto R$ is the activation function. Each x corresponds to a snp encoded as 0,1 or 2, for minor, heterozygous and major allele respectively. Under supervised learning conditions, uniform adaptive optimization governs weight initialization. A rectifier nonlinear activation function f is implemented to control weight summing and node activation according to:

$$f(x) = max(0, x)$$

where x is the input to a computational unit. The network structure contains input, hidden and output layers where nl denote the number of layers and Ll a particular layer. Parameters (W, b) = (W (1) , b(1) , W (n)), b(n)) are described in the network where W (l) denotes the weight of the synaptic connection between unit j in layer l, and unit i in layer l + 1. An intercept node b(l), associated with unit i in layer l + 1 is introduced as a bias to overcome the problem associated with input patterns that are zero. The number of nodes in a layer is denoted by sl for l (this does not include the bias unit). Thus, a(l)i refers to the activation of node i in layer l. Given the parameters W, b, the neural network hypothesis is defined as hW,b(x) which outputs a real number. The network is trained using a sample set of observations (x(i),y(i)) where y(i) R2. With a fixed training set (x(1), y(1)), . . . , (x(m), y(m)) of m examples, the neural network is trained using gradient descent and the cost function is calculated using:

$$J(W, b) = \Big[\frac{1}{m}\sum_{i=1}^{m} J(W, b, x^{(i)}, y^{(i)}\Big] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}(W_{ji}^{(l)})^2$$

$$= \Big[\frac{1}{m}\sum_{i=1}^{m}(\frac{1}{2}\|h_{y,b}(x^{(i)} - y^{(i)})\|^2)] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}(W_{ji}^{(l)})^2$$

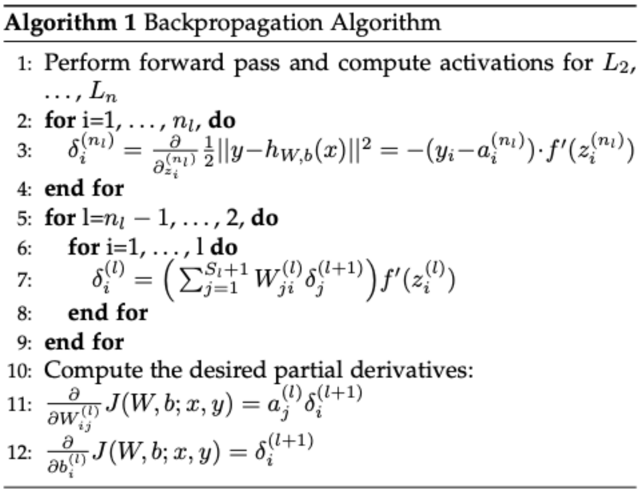
where the first expression is the average sum of squared errors and the second, a weight decay term for decreasing the strength of weights, and preventing overfitting. The relative importance of the two expressions is controlled with the weight decay parameter .

A gradient is a partial derivative and it is computed with respect to a single parameter. The two partial derivatives are computed becuase there are two parameters, a and b in the model. A derivative raprasents how much a given quantity changes when some other quantity varies, so for instance how much does our MSE loss change when we vary each one of our two parameters. The parameters W(l) and each b(l) are initialized to a ij i random value close to zero before training commences. This is a necessary step that prevents hidden layer units learning the same function of the input. The cost function is used to minimize J (W, b) and parameters W, b are updated with:

$$w_{ij}^{(l)} := w_{ij}^{(l)} - \alpha \frac{\delta}{\delta W_{ij}^{(l)}} J(W, b)$$

$$b_i^{(l)} := b_i^{(l)} - \alpha \frac{\delta}{\delta b_i^{(l)}} J(W, b)$$

where α is the learning rate.



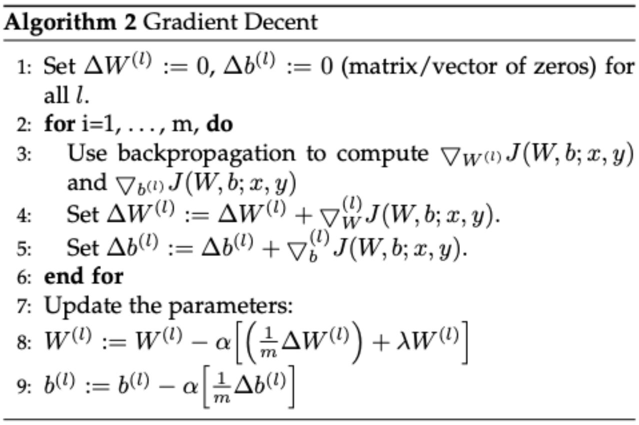
```
Algorithm 1 Backpropagation Algorithm
Perform forward pass and compute activations for $L_2$, $\ldots$, $L_n$
for i=1, $\ldots$, $n_l$, do
    $\delta_i^{(n_l)} = \frac{\partial}{\partial_{z_i}^{(n_l)}} \frac{1}{2} ||y - h_{W,b}(x)||^2 = -(y_i - a_i^{(n_l)}) \cdot f'(z_i^{(n_l)})$
end for
for l=$n_l - 1, \ldots, 2$, do
    for i=1, $\ldots$, l do
        $\delta_i^{(l)} = \left( \sum_{j=1}^{S_l+1} W_{ji}^{(l)} \delta_j^{(l+1)} \right) f'(z_i^{(l)})$
    end for
end for
Compute the desired partial derivatives:
$\frac{\partial}{\partial W_{ij}^{(l)}} J(W, b; x, y) = a_j^{(l)} \delta_i^{(l+1)}$
$\frac{\partial}{\partial b_i^{(l)}} J(W, b; x, y) = \delta_i^{(l+1)}$
```

Figure 4.1: Backpropagation Algorithm

The backpropagation algorithm (see figure above) computes the partial derivatives W(l)ij J(W, b; x, y) and b(l)i J(W, b; x, y) of the cost function for a single sample J(W, b; x, y).
Each node `i` in layer l is used to compute an error element $\delta_i^{(l)}$ that measure the nodes contribution to errors in the output. With respect to output nodes, the error term δ_i^{nl} (where layer n_l is the output layer), represents the difference between the networks activation and the true target value. While hidden units compute a δ_i^l using a weighted average of the error terms of the nodes that use a_i^{nl} as input. When the derivatives have been computed, the derivatives for the overall cost function with the gradient discent are obtained:

4.1.2 Convolutional Model

A convolution is the simple application of a filter to an input that results in an activation. Repeated application of the same filter to an input results in a map of activations called a feature map, indicating the locations and strength of a detected feature in an input, such as an image or like in this study a one dimentinal array of encoded SNPs.
The innovation of convolutional neural networks is the ability to automatically learn a large number of filters in parallel specific to a training dataset under the constraints of a specific predictive modeling problem, such as image classification. The result is highly specific features that can be detected anywhere on input images.
The convolutional neural network, or CNN for short, is a specialized type of neural network model designed for working with two-dimensional image data, although



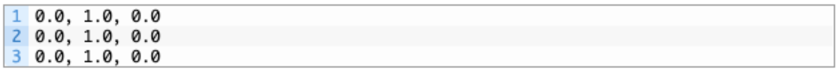
```
Algorithm 2 Gradient Decent
Set $\Delta W^{(l)} := 0$, $\Delta b^{(l)} := 0$ (matrix/vector of zeros) for all $l$.
for i=1, . . . , m, do
    Use backpropagation to compute $\nabla_{W^{(l)}} J(W, b; x, y)$ and $\nabla_{b^{(l)}} J(W, b; x, y)$
    Set $\Delta W^{(l)} := \Delta W^{(l)} + \nabla_W^{(l)} J(W, b; x, y)$.
    Set $\Delta b^{(l)} := \Delta b^{(l)} + \nabla_b^{(l)} J(W, b; x, y)$.
end for
Update the parameters:
$W^{(l)} := W^{(l)} - \alpha\Big[\Big(\frac{1}{m}\Delta W^{(l)}\Big) + \lambda W^{(l)}\Big]$
$b^{(l)} := b^{(l)} - \alpha\Big[\frac{1}{m}\Delta b^{(l)}\Big]$
```

Figure 4.2: Gradient Descent

they can be used with one-dimensional and three-dimensional data. In the related work chapter we have seen that many state of the art models have used convolutions with one dimentional inpout that consisted of DNA sequence, binary encodings for methylation states or variants [3]. Central to the convolutional neural network is the convolutional layer that gives the network its name. This layer performs an operation called a convolution.
In the context of a convolutional neural network, a convolution is a linear operation that involves the multiplication of a set of weights with the input, much like a traditional neural network. Given that the technique was designed for two-dimensional input, the multiplication is performed between an array of input data and a two-dimensional array of weights, called a filter or a kernel. But one-dimentional filters are used extensively as well in the literature [23].
The filter is smaller than the input data and the type of multiplication applied between a filter-sized patch of the input and the filter is a dot product. A dot product is the element-wise multiplication between the filter-sized patch of the input and filter, which is then summed, always resulting in a single value. Because it results in a single value, the operation is often referred to as the scalar product. Using a filter smaller than the input is intentional as it allows the same filter (set of weights) to be multiplied by the input array multiple times at different points on the input. Specifically, the filter is applied systematically to each overlapping part or filter-sized patch of the input data, left to right, top to bottom.
This systematic application of the same filter across an image is a powerful idea. If the filter is designed to detect a specific type of feature in the input, then the application of that filter systematically across the entire input image allows the filter an opportunity to discover that feature anywhere in the image. This capability is commonly referred to as translation invariance, e.g. the general interest in whether the feature is present rather than where it was present. The output from multiplying the filter with the input array one time is a single value. As the filter is applied multiple times to the input array, the result is a two-dimensional array of output values that represent a filtering of the input. As such, the two-dimensional output array from this operation is called a feature map. Once a feature map is created, each value in the feature map is passed through a nonlinearity, such as a ReLU. ReLU is applied to both blood pressure models, the feed-forward and the convolutional models.
In summary, the convolutional model consists of the input of one-dimentional array

of single nucleotide polymorphism, and a filter, which is a set of weights, and the filter is systematically applied to the input data to create a feature map.

For example, below is a hand crafted 33 element filter for detecting vertical lines:

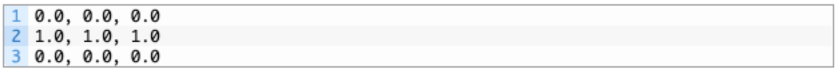


```
0.0, 1.0, 0.0
0.0, 1.0, 0.0
0.0, 1.0, 0.0
```

Figure 4.3: Vertical Lines Filter

Applying this filter to an image will result in a feature map that only contains vertical lines. It is a vertical line detector. Any pixels values in the center vertical line will be positively activated and any on either side will be negatively activated. Dragging this filter systematically across pixel values in an image can only highlight vertical line pixels.

A horizontal line detector could also be created and also applied to the a vector or matrix, for example:



```
0.0, 0.0, 0.0
1.0, 1.0, 1.0
0.0, 0.0, 0.0
```

Figure 4.4: Horizontal Lines Filter

Combining the results from both filters, e.g. combining both feature maps, will result in all of the lines in an input being highlighted. A suite of tens or even hundreds of other small filters can be designed to detect other features in the input. The innovation of using the convolution operation in a neural network is that the values of the filter are weights to be learned during the training of the network.

The network will learn what types of features to extract from the input. Specifically, training under stochastic gradient descent, the network is forced to learn to extract features from the input that minimise the loss for the specific task the network is being trained to solve, e.g. extract features that are the most useful for predicting the systolic blood pressure.

Convolutional neural networks do not learn a single filter; they, in fact, learn multiple features in parallel for a given input by applying multiple filters. For example, it is common for a convolutional layer to learn from 32 to 512 filters in parallel for a given input. This gives the model 32, or even 512, different ways of extracting features from an input, or many different ways of both learning to see and after training, many different ways of seeing the input data. This diversity allows specialization, e.g. not just lines, but the specific lines seen in your specific training data.

Color images have multiple channels, typically one for each color channel, such as red, green, and blue. From a data perspective, that means that a single image provided as input to the model is, in fact, three images. A filter must always have the same number of channels as the input, often referred to as depth. If an input image has 3 channels (e.g. a depth of 3), then a filter applied to that image must also have 3 channels (e.g. a depth of 3). In this case, a 33 filter would in fact be 3x3x3 or [3, 3, 3] for rows, columns, and depth. Regardless of the depth of the input and depth of the filter, the filter is applied to the input using a dot product operation which results in a single value. This means that if a convolutional layer has 32 filters, these 32 filters are not just two-dimensional for the two-dimensional image input, but are also three-dimensional, having specific filter weights for each of the three channels. Yet, each filter results in a single feature map. Which means that the depth of the output of applying the convolutional layer with 32 filters is 32 for the 32 feature maps created.

4.2 Model Training

Both models were trained on a GPU with NVIDIA Tesla M2090 card.
The data was splitted in 50% training set, 25%validation set and 25% for the test set. The prediction for blood pressure used the all data, while the classifier for obesity used one third only.
The predictive model, use slightly more accurare if convolution applied and this gain spedd in training as weel, due to the fact that convolutions have fewer parameters than fully connected networks. Dropout was applied in both cases to reduce overfitting.
To train the network, the derivative of L(Y i, Yi) is computed with respect to the parameters of the network, which are the collection of neuron weights. By updating the weights in a small step in the direction of the derivative, the networks prediction loss can be decreased, and its accuracy can be increased. The derivative can be efficiently computed for each training point via the chain rule from calculus; this process is commonly called back-propagation. In parallel, the networks accuracy is also evaluated on the validation data, which are not used to update the weights. To control the effect of overfitting, dropout was used. Dropout is a training process that randomly ignores nodes to mitigate overfitting.

4.2.1 Training Data

In this study the data has been splitted into 50%, 25%, 25% for the training, validation and test sets, respectivelly.
Training a neural network starts with a labeled dataset of (Xi, Yi), where each Xi is the ith input, and Yi is its output label. Each training point Xi is fed into the network, and the networks output is evaluated against the true label Yi to produce a loss L(Y i, Yi). Loss is the sum of the errors made for each example. The lower the loss, the better the model. Commonly used loss functions include squared error and cross-entropy. Squared error measures the difference between predicted and actual output values, which is especially relevant when the output is a continuous value. Cross-entropy measures the difference between two probability distributions over the same set of underlying events or classes, as is appropriate when the output is categorical. There are two primary classes of phenotypes: categorical (often binary case/control) or quantitative. For some disease traits of interest, quantitative disease risk factors have already been identified [18].

4.3 Encoding X and Y

Distributed representation of word, or neural word embedding, was a recent breakthrough in NLP research based on deep learning. The goal of word embedding is to

derive a linear mapping, i.e., embedding, from the discrete space of individual words to a continuous Euclidean space such that similar words will be mapped to points in close vicinity in the embedding space. The direct benefit of word embedding is that such representation of individual words, vectors in continuous space, becomes differentiable and thus amenable for back-propagation-based neural network modeling. [11]
The challenge of creating a quantitative semantic representation of discrete units of a complex system is not unique to gene systems. For a long time, creating a quantitative representation of words had been challenging for linguistic modeling. Hinton proposed the pioneering idea of learning distributed representations of words [1], i.e., representing the semantics of a word by mapping them to vectors in a high-dimension space. However, Hintons idea did not lead to real implementation in mainstream natural language processing (NLP) research, until recently. The word2vec model achieved success in NLP modeling [2]. This process of distributed representation is often called neural embedding because the embedding function is often expressed by a neural network with a large number of parameters.
In this study the phenotypes or Y of interest are: blood pressure and BMI. For each SNP, or single unit in X, there is a minor allele, major allele and heterozygous allele. Two types of SNP encodings have been used. One embeds each SNP as one-hot encoded vector of size 1X3, where a minor allele is rapresented as [1,0,0], an heterozygous allele is rapresented as [0,1,0] and the major allele is rapresented as [0,0,1]. A second method has been used. The latter had better predictive results than the first one. Based on those results and this paper[11], all the experiments and models training have been conducted using the SNP encoding as below:
minor allel (bb) = 0
heterozygous (Ab) = 1
major allele (AA) = 2

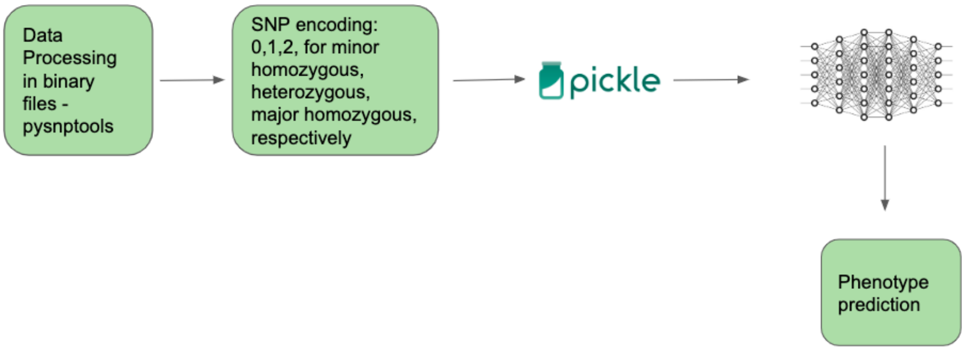

Data Processing in binary files - pysnptools
SNP encoding: 0,1,2, for minor homozygous, heterozygous, major homozygous, respectively
pickle
Phenotype prediction

Figure 4.5: Pipeline: Deep Learning for Phenotypic Prediction.

4.4 Used Softwares and Hardwares

The models are implemented in Pytorch, version (1.0.0). The package pysnptools and the command line application PLINK were used to process the data and extract all single nucleotide polymorphysm. The final trainings were deployed on one GPU with NVIDIA K80 card in amazon cloud.

Chapter 5

Results and Conclusion

The phenotypes of this analysis are: systolic blood pressure, dystolic blood pressure and and BMI or obesity risk. In this study two different model have been implemented for the phenotype blood pressure and one classifier for BMI phenotype.
There are several challenges in studying the joint effects of multiple genetic and environmental variables. First, in typical GWAS, genotypes of up to one million SNPs are determined in several thousand subjects, leading to the small n, large p problem (many more variables (SNPs) than samples). Second, when a large number of SNPs are genotyped on a genome-wide scale, linkage disequilibrium (LD) between SNPs (resulting in correlated variables) needs to be taken into account. For these reasons, standard multi-variable statistical approaches like multiple linear or logistic regression are not well suited for genome-wide data [15], and neural networks are better suited for this kind of problem.
The limit of the pipeline is that does not account for low trait heritability. Low means that environmental effects and chance effects contribute to the phenotypic differences of interest. Here the pipeline demontrates that using SNPs as single input is not enough to make phenotypic predictions. Thus other omics need to be exploited together in order to interpret and obtain a significant association between input and output. One note to be taken in account for future work is that applying quality control, didn't yield better results. Thus removing relatives of second degree didn't have any effects on the final result.
Future sequencing technology that will replace one million SNPs with the entire genomic sequence of three billion nucleotides and replace the microarray sequencing with less noisy ones, will improve the results of deep network applications. Challenges associated with data storage and manipulation, quality control and data analysis will be manifold more complex, thus challenging computer science and bioinformatics infrastructure and expertise.
Merging sequencing data with that from other high-throughput technology for measuring the transcriptome, the proteome, the environment and phenotypes such as the massive amounts of data that come from neuroimaging will better help understand the genotype-phenotype relationship for the purpose of improving healthcare. There may be greater potential for identifying variants disease association and improving phenotypes prediction from the future release of genetic data for all 500,000 UK Biobank participants.

Bibliography

[1] *About UK Biobank*. URL: https://www.ukbiobank.ac.uk/.

[2] Altshuler et al. "Integrating common and rare genetic variation in diverse human populations." In: *Nature* 467 (2010). DOI: 10.1038/nature09298.

[3] Angermueller1 et al. "DNA methylation states using deep learning." In: *Genome Biology* (2017). DOI: 10.1186/s13059-017-1189-z.

[4] Buniello et al. "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics". In: *Nucleic Acids Research* 47 (2019), S51–S57. DOI: 10.1093/nar/gky1120.

[5] Cooper et al. "A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose." In: *Blood* 112 (2008). DOI: 10.1182/blood-2008-01-134247.

[6] Fergus et al. "Utilising Deep Learning and Genome Wide Association Studies for Epistatic-Driven Preterm Birth Classification in African-American Women". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (). DOI: https://ieeexplore.ieee.org/document/8454302.

[7] Frazer et al. "Human genetic variation and its contribution to complex traits." In: *Nat Rev Genet* (2009). DOI: 10.1038/nrg2554.

[8] Griffith et al. "ORegAnno: an open- access community-driven resource for regulatory annotation". In: *Nucleic Acids Research* 36 (2008). DOI: 10.1093/nar/gkm967.

[9] Haines et al. "Complement factor H variant increases the risk of age-related macular degeneration." In: *Science* 308 (2005). DOI: 10.1126/science.1110359.

[10] Hirschhorn et al. "Genome-wide association studies for common diseases and complex traits." In: *Nat Rev Genet* 6 (2005). DOI: 10.1038/nrg1521.

[11] Jingcheng et al. "Gene2vec: distributed representation of genes based on co-expression." In: *BMC Genomics* (2018). DOI: 10.1186/s12864-018-5370-x.

[12] Kerem et al. "Identification of the cystic fibrosis gene: genetic analysis." In: *Science* 245 (1989). DOI: 10.1126/science.2570460.

[13] MacDonald et al. "The Huntingtonfffdfffdfffds disease candidate region exhibits many different haplotypes." In: *Nat Genet* 1 (1992). DOI: 10.1038/ng0592-99.

[14] M.H. et al. "Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013." In: *Lancet* (2015). DOI: 10.1016/S0140-6736(15)00128-28.

[15] Szymczak et al. "Machine learning in genome-wide association studies". In: *Genetic Epidemiology* 33 (2009), S51–S57. DOI: 10.1002/gepi.20473.

[16] Warren et al. "Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk". In: *nature genetics* (). DOI: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5972004.

[17] *Bgenix - Variants.* URL: https://bitbucket.org/gavinband/bgen/wiki/bgenix.

[18] Moore Bush. "Genome-Wide Association Studies." In: *PLOS Computational Biology* (2012). DOI: 10.1371/journal.pcbi.1002822.

[19] Moore Bush. "Genome-Wide Association Studies". In: *PLOS - Computational Biology* (), p. 6. DOI: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531285.

[20] *Calls Formats and PLINK.* URL: https://www.cog-genomics.org/plink/1.9/formats.

[21] The 1000 Genomes Project Consortium. "A map of human genome variation from population-scale sequencing". In: *Nature* (2010), p. 5. DOI: 10.1038/nature09534.

[22] Daly MJ Hirschhorn JN. "Genome-wide association studies for common diseases and complex traits." In: *Nat Rev Genet* (2005). DOI: 10.1371/journal.pcbi.1002822.

[23] Ali Farhadi Joseph Redmon. "YOLOv3: An Incremental Improvement." In: (2017).

[24] M. et al Munfffdfffdoz. "Evaluating the contribution of genetics and familial shared environment to common disease using the UK Biobank". In: *Nature Genetics* (2016). DOI: 10.1038/ng.3618.

[25] L. et al. Sundaram. "Predicting the clinical impact of human mutation with deep neural networks." In: *Nat. Genet.* (2016).