Phenotype Prediction with Neural Additive Models

Jessica Houghton
Department of Statistics
Stanford University
Stanford, CA 94305
jlhought@stanford.edu

Sasha Yousefi
Department of Statistics
Stanford University
Stanford, CA 94305
syousefi@stanford.edu

1 Abstract

Machine Learning must balance two different objectives: attain optimal performance while maintaining interpretability. While Deep Neural Networks have shown to outperform linear models in genomic tasks, these models often do not allow us to determine which parts of the genome lead to a predicted outcome. In this paper, we propose the use of a Neural Additive Model (NAM) for phenotypic prediction. In a NAM, each feature feeds into a separate deep neural net subnet. Each subnet learns a separate additive model for each feature and those subnets are added together at an output layer. In our experiments, we trained our NAM on single nucleotide polymorphism (SNP) data from all consenting patients in the UK biobank. We then predicted on two binary labels and two continuous label: red hair, bilirubin levels, celiac disease, and age of diabetes diagnosis. The NAM proved to have superior predictive power than previously used linear and tree based methods for all tasks. Additionally, NAMs allow us to capture features (SNP) importances as well as minor allele contributions to a particular phenotype. From our ranking of SNP importances, we were able to identify which variations in the genome the model deemed important for prediction. For example, for celiac disease, our model recognizes variant rs3891175 as having a significant contribute to model output. The NAM ranks individuals with the risk allele (Thymine) as having a higher chance of developing celiac. This variant aligns with the literature, as it is located in the HLA-DBQ1 gene, which regulates immune pathways and inflammation in response to gluten. Our NAM is able to identify many risk variants and alleles that align with current literature, as well as identify a few elusive variants with little known functions. All in all, our NAM provides promising results in identifying potentially novel risk variants while corroborating the current literature.

2 Introduction

Disease intervention has always been the central aim of medicine. Scientists and engineers alike have spent decades decoding ways to predict and prevent disease. Finally, scientists have settled on genetic data as a way to make meaningful predictions. In the past, scientists have used machine learning methods such as gradient boosted trees in order to make phenotypic predictions.[3] Additionally, deep learning methods such as CNNs and Transformers have been used to predict phenotypes from genetic sequences.[24] However, both fall short as boosted trees lack the complexity that deep learning methods can express and deep learning methods lack the interpretability that simpler models provide. Thus, In our project, we have chosen to utilize Neural Additive Models to make phenotypic predictions from genomic data. Neural Additive Models combine the expressivity of DNNs while still maintaining the inherent intelligibility of general additive models.[1] This structure is vital for the medical landscape as interpretable models are key for physician decisions.

We will be looking at genomic data from individuals in the UK biobank in order to predict disease phenotypes in individuals. As genome sequencing technology has improved in the past few years, the ability to sequence one's genome has gotten cheaper and faster. Thus, there is a large influx of genetic data that has yet to be analyzed and has the ability to make scientific discoveries in human

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

health. This problem is important as there are many diseases that are genetically related through a complex relationship between many different genes. Furthermore, there are still many regions of the genome that scientists have very little to no understanding of their biological function. In our project, we hope to figure out some of the important genetic features that can predict four phenotypes: celiac disease, total bilirubin levels, age at diabetes diagnosis, and red hair.

3 Related Work

As the UK Biobank is a unique prospective cohort study of genetic and phenotypic data with unprecedented size, researchers all over the world are working to tap into the amazing depth of scientific knowledge and potential captured in the dataset. Given the size and scope of the data, it goes without being said that this dataset lends itself to unique machine learning and deep learning opportunities to understand the role genetics has in determine health outcomes. Many machine-learning approaches have already been applied to various aspects of the UK Biobank data, some but not all of which will be described in this section. However, it is first imperative to describe previous statistical methods for calculating polygenic risk score (PRS), or individual disease risk based on genetic variants.

The first approach to determining polygenic risk score (PRS) derives from linkage disequilibrium (LD), or the nonrandom associations between alleles of different genetic variants. This is called the "clumping/pruning and thresholding" approach, where a smaller subset of genetic variants are incorporated through LD pruning while accounting for clumping, the evidence of association with the studied trait. Then, the PRS is calculated by summing over all SNPs that have crossed a p-value threshold.[15] However, this approach has limited predictive accuracy since it does not incorporate the information from the discarded SNPs not included in the smaller subset. LDpred is another commonly used statistics-based method for calculating polygenic risk scores which uses a Bayesian approach to model the correlation structure between all variants instead of a subset of SNPs. This approach has shown to improve accuracy compared to the pruning method especially with large sample sizes.[23]

Although these genome-wide association studies (GWAS) are the state-of-the-art method for determining the association between a single nucleotide polymorphism (SNP) and a phenotype, these methods have many limitations. First, the typical effects of a risk allele are typically small and there is still a large amount of genetic heritability that has yet to be explained. Secondly, many of these models fail to account for epistasis, which is the interaction between different genes which can impact a phenotype. Thus, there is a push for machine learning methods to handle more complex effects and interactions of genes on given phenotypes. However, most machine learning methods lack the interpretability that statistical methods provide.

The GraBLD method was shown to enhance predictive performance in polygenic score heuristics and LDpred for height, BMI, and diabetes status within 130K UK Biobank participants. The method uses gradient boosted regression trees to optimize the weights of each SNP in the score, followed by a regional adjustment for linkage disequilibrium. This approach takes advantage of machine learning techniques combined with statistical summary-level data from genome-wide association studies (GWAS) to make predictions. [18]

One of the most successful machine learning methods for genotype-phenotype data are an ensemble method of gradient tree boosting models.[5] In particular, XGBoost is a popular option that takes advantage of many regression trees to handle data of mixed type and are computationally efficient.[3] Although these methods have decent predictive power, there is still a lack of the ability to evaluate the importance of each SNP on the prediction. However, the SHapley Additive exPlanation (SHAP) value method has been shown to efficiently and theoretically interpret the importance of each feature in tree ensemble models with local feature interaction effects.[16]. This method has been evaluated with the UK Biobank with obesity as a phenotype to reproduce previous research on the top known SNPs associated with obesity.[10]

Although statistical and machine learning methods have been used in genotype-phenotype prediction, the use of deep learning in phenotype prediction has yet to be explored given the difficulty of the explainability of black box models. A promising new publication proposes an interpretable Transformer for Gene Expression Modeling (T-GEM) for cancer phenotype prediction from RNA-seq and scRNA-seq gene expression data. This model utilizes transformers which have been widely applied in NLP and imaging prediction to capture important biological functions associated with phenotypes. In particular, the beginning layers of their model has broader attention and higher layers focus more on phenotype-related genes. Then, the importance of each feature can be interpreted by using entropy and attribution scores of attention weights of a head. [24]

Another deep learning approach to phenotype prediction uses interpretable neural network architectures that are constructed with the embedding of previous biological knowledge. GenNet uses only biologically plausible connections (exon, pathway, chromosome annotations, etc) to make a more memory efficient instead of incorporating all possible SNPs into fully connected layers. This model was applied to SNPs, exons, and genes in the UK Biobank to predict 15 different phenotypes including hair color and dementia. This model proved to replicate well known genes such as OCA2 for hair color, as well as novel genes such as ZNF773 for schizophrenia.[8] Although both T-GEM and GenNet relate to our project in theory for finding the most important genetic sequences for phenotype prediction, the input data for our project will be of a different type. We aim to employ another one of these promising interpretable deep learning methods called Neural Additive Models (described below) to uncover the genetic associations with phenotypes outlined in SNP data from the UK Biobank. To our knowledge, there has yet to be any deep learning models applied to SNP-only data given the complex nature of sequence genetic data without translation invariance. This is important as it can inform scientists what position specific genetic modifications may be the most responsible for health-related outcomes.

4 Data

The dataset we will be using is the UK biobank. The UK Biobank is a large-scale biomedical database containing in-depth health and genetic information from half a million UK participants. The UK biobank study follows half a million patients from age 40 to 69 who initially volunteered in a cohort in 2006; these patients will be followed for 30 year thereafter. They were interviewed about lifestyle and medical metrics, attained a physical examination, and most importantly for our case, had their genomes sequenced. The database is frequently updated and is globally accessible to approved researchers taking part in meaningful healthcare research. More than 800,000 genetic variants were directly genotyped and greater than 90 million variants have been imputed.

A sample of the data consists of a set of anonymous patients who may have a particular Single Nucleotide Polymorphism (SNP) at a locus in their genome. These patients would have a record of their genome, health/lifestyle history, as well as routine lab measurements. We will be using the sequenced genetic data along with their health/lifestyle history to make phenotypic predictions. For example, we will be using the red hair phenotype as well as bilirubin levels as our preliminary phenotypes. Our dataset labels would be binary (0, 1) for the red hair trait and a positive value on a continuous scale for bilirubin levels. The training data for each patient would contain gender as well as the value at each SNPs associated with either red hair or bilirubin (obtained from a snpnet polygenic risk scores result). The SNP genotype is expressed as a 0, 1, or 2 to express how many copies of the allele there are. To avoid class imbalance in the red hair phenotype, we downsampled our negative examples in order to have perfectly balanced classes. Additionally, the average bilirubin score for the dataset was 9.12 and the standard deviation was 4.42. We handled missing data by removing any approved UK biobank patient that did not have data for the red hair phenotype or bilirubin levels.

There are a few obstacles with the data at present. First, the database is restricted to UK citizens, which could mean demographic biases and lower representation of certain groups. Additionally, the age range of the UK biobank would exclude data from certain age brackets, which could limit use cases for those ages.

5 Approach

While implementing deep neural networks in biomedicine we often run into the issue of interpretability. While DNNs are powerful black-box predictors, they often are unclear on how they make their decisions, which hinder their applicability in healthcare. Thus, our first approach is to use Neural Additive Models (NAMs) which according to Agarwal et al., 2021 “combine some of the expressivity of DNNs with the inherent intelligibility of generalized additive models”.[1] At a high level, NAMs

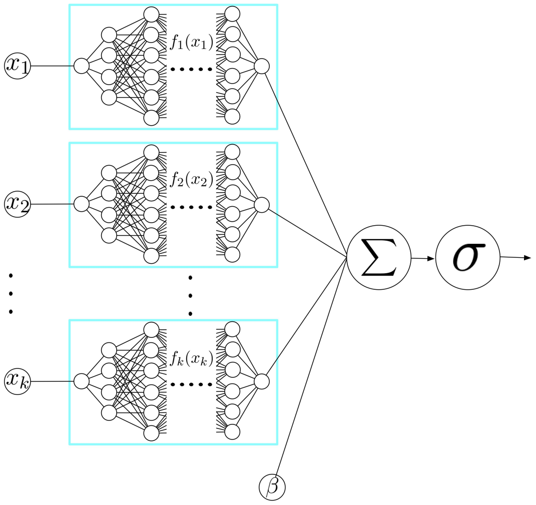

Figure 1: Neural Additive Model for Binary Classification

are quite similar to generalized additive models, but are based on neural networks instead of boosted trees, which offers more flexibility while maintaining accuracy. In a NAM, each feature feeds into a separate deep neural net subnet. Each subnet learns a separate additive model for each feature and those subnets are added together at an output layer. Subnets are learned in parallel in order to optimize runtime. Lastly, in the classification setting, a sigmoid is applied to achieve probabilities for each class. After training, subnets could be replaced with feature importance graphs, which show how the output varies with the possible feature values. Our NAM will follow the backbone described in Agarwal et al., 2021.[1]

First, we must apply this pipeline to a genetic framework, which will involve manipulation to transform genome sequences into features. We have attached the code to parse data in the UK Biobank in the supplementary materials. As described in the data portion of this paper, after attaining UK biobank access, we combined demographic information (gender) along with SNP data related to the relevant phenotype as our training data. We attained SNP data for each consenting individual, and transformed them into a dataframe with the patient ID as the index and the SNP id or demographic feature as the column. As our labels, we attained either binary values or values on a continuous scale to represent the presence or value of the associated phenotypic trait (i.e. red hair phenotype or bilirubin levels). For binary traits, we downsampled the negative cases to achieve a balanced dataset between labels - this ensures that the AUC is a reliable metric to measure predictive power. Some patients in the database had missing SNP values in their genome. This is likely the result of sequencing errors. For these values (represented as -9 in our database instead of the typical 0/1/2), we imputed them by using the median value for that SNP across all relevant patients.

Hyperparameters

Num epochs	10, 20, 30, 40, 50
Num learners	1, 3, 5, 10
Data augmentation	No PCs, PCs added
Optimizers	Adam

Table 1: Hyperparameters tuned to optimize our NAM models. 10 PCs were included to adjust for population stratification.

For genetic datasets, one must take genomic difference on the basis of demographics into account. Many phenotypes vary significantly based on ethnic background - for example, European populations are much more susceptible to Celiac Disease. Thus, SNPs that may turn up "significant" could simply be capturing demographic differences between one population and the rest. The SNP could have little/no affect on the phenotype, but if the phenotypically dominant demographic group displays that

SNP, then it would turn up as "significant". To mitigate this problem, we look to principal component analysis (PCA). As suggested by Byun et al. [2], using top PCs as covariates corrects for stratification in genomic studies. PCA will appropriately apply a correction to markers with large difference in allele frequency across ancestral populations. Thus, we incorperated the top ten PCs as covariates for each dataset in order to control for ancestral genomic variations.

A major advantage of NAMs is the ability to identify the contribution of each feature (in our case SNP) to the final prediction. Since our model is additive, we can get the contribution of each feature by doing the following procedure: 1) zeroing out the input matrix, 2) for the feature (SNP) of interest, adding back a row for each SNP value (0 = no minor allele, 1 = 1 copy of minor allele, 2 = two copies of minor allele), 3) running a forward pass on the model to get the contribution of each allele of the SNP to prediction. After running this procedure for all SNPs for our phenotype, we used measured feature importance for each SNP. We did so by taking the mean of the three genotypes and calculating the sum of absolute distance for each genotype from the mean. This metric allows lets us see if the model identifies whether having 0, 1, or 2 copies of an allele significantly affects your likelihood for expressing a the phenotype of interest.

Lastly, as a baseline model we plan to run a Gradient Boosting Tree model, XGboost, on our dataset. XGboost is a decision tree ensemble algorithm which attempts to accurately predict target variables by combining the estimates of a set of simpler, weaker models [3]. We will run XGboost with an interaction level of 100 on the training data described above to get a benchmark comparison for our chosen phenotypes.

6 Experiments

NAM Results for Prediction of Classification Tasks

Model Performance						
Model	Phenotype	Val AUC	Test AUC	Sensitivity	Specificity	Operating Point
XGBoost	Red Hair	0.908	0.918	0.865	0.867	0.466
NAM	Red Hair	0.947	0.961	0.897	0.907	0.546
XGBoost	Celiac Disease	0.733	0.764	0.671	0.781	0.226
NAM	Celiac Disease	0.853	0.852	0.737	0.841	0.1523
NAM	Celiac + Red Hair	0.773	0.677	0.682	0.614	0.451

Table 2: AUC results for the validation and test sets with our Neural Additive Model and Baseline XGBoost. Two phenotypes were explored, red hair and celiac disease. Both tasks had a binary label. Sensitivity, specificity, and operating point were also recorded. Additionally, AUC metrics for multitask learning of the celiac and red hair (Celiac + Red Hair) are given.

NAM Results for the Prediction of Regression Tasks

Model Performance			
Model	Phenotype	Val MSE	Test MSE
XGBoost	Bilirubin Levels	11.497	11.75
NAM	Bilirubin Levels	10.487	10.366
XGBoost	Age at Diabetes Diagnosis	204.983	205.721
NAM	Age at Diabetes Diagnosis	160.590	165.647
NAM	Bilirubin + Diabetes Age	115.032	93.623

Table 3: MSE results for the validation and test sets with our Neural Additive Model and XGBoost baseline model. Two phenotypes were explored, bilirubin levels and age at diabetes diagnosis. Both tasks had a continuous label. Additionally, MSE metrics for multitask learning of the bilirubin and age at diabetes diagnosis (Bilirubin + Diabetes Age) are given.

In our results, we showcase analysis and evaluation of four different phenotypes: Red Hair, Celiac Disease, Bilirubin Levels, and Age at Diabetes Diagnosis. For our classification tasks, Red Hair and Celiac, we report AUC (area under the ROC curve), Specificity, Sensitivity, and operating point. We chose the best operating point threshold so that the classifier gives the best trade off between the costs of failing to detect positives against the costs of raising false alarms. For our regression tasks, Bilirubin Levels and Age at Diabetes Diagnosis, we use MSE (mean squared error) to evaluate our results. We report all metrics for the test and validation sets as shown in Table 1 and Table 2. Additionally, using our feature importance metric we describe in the Approach section, we list two of our top 10 SNPs (Table 3) in this section as well as their accompanying graphic (Fig 2). The graphic portrays the influence of each genotype on the predictive outcome. All genomic analysis of SNPs and phenotypes were done using the Global Biobank Engine [**engine**] as well as GWAS Catalog [6].

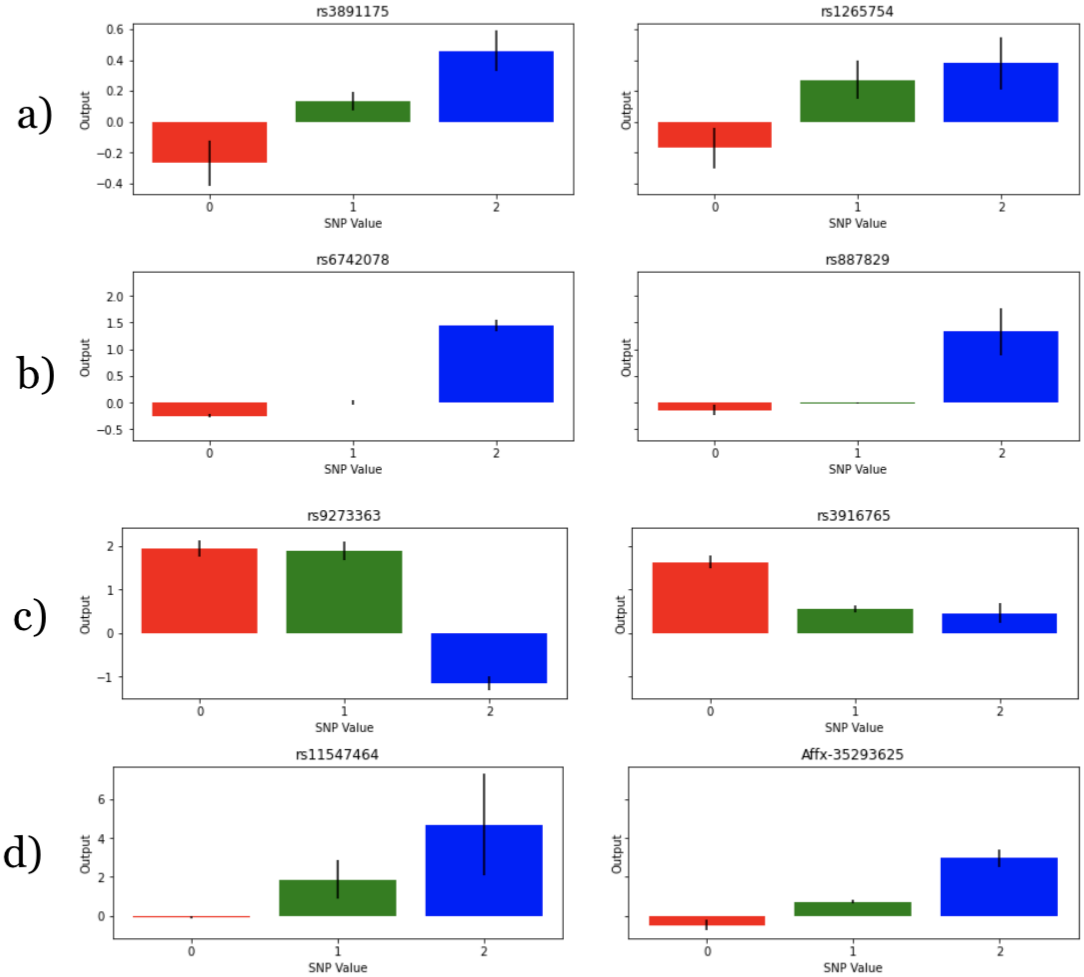

Figure 2: Allelic Contributions for the top 2 most important SNPs for each phenotype. Alleles (0 copies of alternate, 1 copy of alternate (heterozygote), 2 copies of alternate) are listed on the x axis, and contribution to model output is listed on the y axis. A) corresponds to celiac disease, b) corresponds to total bilirubin levels, c) corresponds to age diabetes diagnosis, d) corresponds to red hair

6.1 Celiac Disease

Celiac Disease is an autoimmune disorder that causes a reaction in the body to gluten. Our NAM not only showed that Celiac Disease is highly predictive, but identified important SNP variants that are known to affect gluten malabsoprtion in the literature. Our NAM achieved an AUC of 0.852, which general success in differentiating between disease and control patients (Table 1). This model significantly outperformed our XGboost model, which had an AUC of 0.733 (Table 1). As a reminder, Genetics of Significant Celiac SNPs

SNP	Location	Allele	Gene	Most severe consequence	Affected Traits
rs3891175	Chr6:32666690	C/T	HLA-DQB1	5 prime utr variant	celiac disease, aspartate aminotransferase measurement, autoimmune hepatitis
rs1265754	Chr6:32335915	T/A	TSBP1	missense variant	inguinal hernia, celiac disease

Table 4: Significant Celiac SNPs and their associated location, allele (major/ minor), Gene, most severe mutation caused by variant, and affected traits as determined in the literature

we down-sampled our dataset in order to get accurate measures of predictive power. Using our feature-importance metric, we we list the top ten most impactful SNPs our model used for celiac disease prediction (Supplementary Table 1). Additionally, we provide an accompanying graphic for this phenotype indicating how the genotype of each SNP affected model output (Supplementary Figure 2).

The most important SNP that the model identified for predicting Celiac Disease is rs3891175, which is a variant in the 5 prime UTR of the HLA-DQB1 gene. The HLA complex form a network of genes that regulate the immune response of the human body. These genes are located on chromasome six, and HLA-DQB1 has been associated with traits such as celiac disease, aspartate aminotransferase measurement, and autoimmune hepatisis. Aspartate Aminotransferase levels are commonly found to be elevated in patients with celiac disease in association with liver dysfunction [13]. Since HLA genes are associated with such a broad mechanism, we would expect a mutation in rs3891175 to affect multiple pathways. Through our visual representation of feature importance (Figure 2.a) we are able to see that two copies of the minor allele (T/T) increase our model output, making our model more likely to predict celiac disorder. On the other hand, two copies of the major allele at this SNP decreased the model's output, meaning the model is less likely to predict celiac.

The second most important SNP that the model identified for predicting Celiac Disease is rs1265754, which is a SNP in the TSBP1 gene on chromosome 6. This SNP has been linked to inguinal hernia and celiac disease in the literature. Much of this gene function is unknown, but it is hypothesized to play a role in fetal development. Gene expression is highest in the testis but is also seen in other tissue types such as the brain, lens of the eye, and the medulla. The TSBP1 gene is found in close proximity to the HLA cluster, which could mean TSBP1's affect on immune function and celiac disorder is due to Linkage Disequilibrium (when an allele of one genetic variant is inherited or correlated with an allele of a nearby genetic variant). It could also mean that variations in the TSBP1's transcription could have downstream affects on the HLA gene cluster. In figure 2.a, we see a similar trend to the rs3891175 SNP, where two copies of the minor allele increase model output and two copies of the major allele decrease model output. Additionally, while less so than a homozygous recessive genotype, the heterozygote also contributes notably to model output for celiac prediction, adding a positive contribution for celiac prediction from the T/A phenotype.

The other SNPs our model deemed significant are all on chromosome 6, and are either in the HLA gene cluster or in neighboring segments. Most SNP variants shown in Supplementary Table 8 have been linked to immune responses for diseases such as celiac, Rheumatoid arthritis, Hypothyroidism, and diabetes. Their associated allelic contributions to model outputs are shown in Supplementary Figure 2.

6.2 Total Bilirubin

The NAM model trained on total bilirubin levels was quite predictive compared to previous models. As seen in Table 2, the MSE for this model was 10.487 on the validation set and 10.366 on the test set, which both outperformed the XGBoost model. In addition, the most important SNPs for our NAM model in predicting total bilirubin levels are many variants that are already known to be associated with bilirubin in literature. The most important SNP, rs6742078 is an intron variant

Genetics of Significant Total Bilirubin SNPs

SNP	Location	Allele	Gene	Most severe consequence	Affected Traits
rs6742078	Chr2:233763993	G/T	UGT1A1, UGT1A3, UGT1A4, UGT1A5, UGT1A6, UGT1A7, UGT1A8, UGT1A9, UGT1A10	intron variant	serum metabolite measurement, bilirubin measurement, response to tenofovir, blood protein measurement, bilirubin measurement x insomnia, bilirubin measurement x response to tenofovir, aldosterone measurement, circulating cell free DNA measurement
rs887829	Chr2:233759924	C/T	UGT1A3, UGT1A4, UGT1A5, UGT1A6, UGT1A7, UGT1A8, UGT1A9, UGT1A10	intron variant	metabolite measurement, serum metabolite measurement, bilirubin measurement, blood metabolite measurement, biliverdin measurement, bilirubin measurement x insomnia, X-11530 measurement, bilirubin measurement x response to tenofovir, total cholesterol measurement, blood protein measurement, cholelithiasis x bilirubin measurement

Table 5: Significant Total Bilirubin SNPs and their associated location, allele (major/ minor), Gene, most severe mutation caused by variant, and affected traits as determined in the literature. Variants with N/A values had attributes that were not found in the literature. Additionally, gender has N/A values for their genomic descriptions. We did not include PCs as variants since we did not want to capture demographic differences in phenotypes.

located on chromosome 2 and is associated with the UGT1A1, UGT1A3, UGT1A4, UGT1A5, UGT1A6, UGT1A7, UGT1A8, UGT1A9, and UGT1A10 genes [21]. These genes are a part of the UDP glucuronosyltransferase family which is responsible for glucorinidation of bilirubin and other substances. The UGT1A1 is the major gene responsible for bilirubin conjugation, and is essential for detoxification within the liver [12, 14]. In particular, metabolic diseases of the liver such as Crigler-Najjar Syndrome type 1 and Gilbert syndromes are known to be associated with mutations in the UGT1A1 gene [19]. The feature graph in Figure 2b shows that a homozygous dominate genotype in this position is predicted to have slightly lower predicted bilirubin levels, whereas one with a homozygous recessive genotype is predicted to have much higher levels. This indicates that patients with a homozygous recessive are likely to be at risk for unconjugated hyperbilirubinemia according to our model.

A majority of the other top ten most important SNPs are also located on chromosome 2 and are associated with the UGT1A gene family (see Appendix). SNPs rs887829, rs34622615, rs4148325, rs2070959, and rs4124874 predict lower levels of bilirubin for homozygous dominants, slightly increased levels for heterozygotes, and highest predicted levels for homozygous recessives. However, SNPs rs1661052, rs34691116, and rs11045819 are located on chromosome 12 and are related to the human solute carrier (SLC) superfamily. These genes encode proteins that transport compounds such as bilirubin from the blood into the liver. Although not quite as strong of a known association as UGT1A1, the SLC22A18, SLC01B3, and SLC01B1 genes has been found to be associated with bilirubin levels and unconjugated hyperbilirubinemia [20, 11] . Finally, gender was the sixth most important feature in the bilirubin level prediction for our model. Our model predicts that if one identifies as female (0) they are likely to have a slightly higher predicted level of bilirubin compared to males (1). However, this does not match up with literature, as the National Health and Nutrition Examination Survey show males having significantly higher bilirubin levels (0.72 +/- 0.004) [25]. In total, our top reported SNPs for bilirubin prediction are in line with current literature documenting the association of the UGTA and SLC gene families.

6.3 Age Diabetes Diagnosed

Genetics of Significant Age Diabetes Diagnosed SNPs

SNP	Location	Allele	Gene	Most severe consequence	Affected Traits
rs9273363	Chr6:32658495	C/A	HLA-DQA1, HLA-DQB1	regulatory region variant	type 1 diabetes mellitus, chromic lymphocytic leukemia
rs3916765	Chr6:32717773	G/A	HLA-DQB3, MTC03P1	intergenic variant	acute myeloid leukemia, type 2 diabetes mellitus

Table 6: Significant Age Diabetes Diagnosed SNPs and their associated location, allele (major/minor), Gene, most severe mutation caused by variant, and affected traits as determined in the literature. Variants with N/A values had attributes that were not found in the literature. Additionally, gender has N/A values for their genomic descriptions. We did not include PCs as variants since we did not want to capture demographic differences in phenotypes.

The age of diabetes diagnosis was also quite predictive compared to baseline methods. As seen in Table 2, our NAM model had a 160.590 validation MSE and 165.647 test MSE, both of which outperform the XGBoost's performance. Although this is a prediction problem of what age a patient was diagnosed with diabetes rather than a binary disease outcome, the model was still able to use some significant SNPs related to diabetes for prediction. The most important SNP for age at diabetes diagnosis were variable in terms of the chromosomes spanned (4, 6, 11, 20) and the genes associated. The top SNP was rs9273363 which is located on chromosome 6 and associated with the HLA-DQA1 and HLA-DQB1 genes [22]. The human leukocyte antigen (HLA) system is crucial in the body's response to bacterial and viral infections. HLA-DQA1 and HLA-DQB1 are within the HLA class II molecules which are used to display peptides from extracellular pathogens [17]. These two genes make up a protein complex and certain haplotypes (HLA-DQA1*05 and DQB1*02 and HLA-DQA1*03 and DQB1*03) are known to affect the risk of type 1 diabetes [4, 9]. The hypothesis is that these haplotypes can lead to autoimmune damage of insulin-producing cells, leading to high blood sugar levels. In Figure 2c, homozygous recessives are predicted to be diagnosed with diabetes earlier than homozygous dominants and heterozygotes.

The second most important SNP in our model is rs3916765 which is located on chromosome 6 and associated with the HLA-DQB3 and MTC031. MTC031 is a pseudogene for MTC03 which encodes the mitochondria cytochrome c oxidase subunit III (COX3) enzyme. It is known that mitochondrial dysfunction is related to the development of diabetic cardiomyopathy [7]. The other most important SNPs have little to none known association with diabetes, and further investigation is needed to determine the relationship between the genes. It is possible that these SNPs affect genes that are new discoveries yet crucial in diabetes diagnosis and development.

6.4 Red Hair

In predicting the red hair phenotype from genomic inputs, our model identified crucial SNP variants that are commonly known to affect pigmentation in the literature. As red hair is known to have a rather high predictive power, our AUC was around 0.961 for the test set, which indicates high accuracy (Table 1). This outperforms our baseline XGboost model (AUC = 0.918), showing that the NAM is superior for red hair phenotypic prediction. Through our feature importance metric, we identity the top ten most important SNPs (Supplementary Table 11) along with their genomic contributions to the model output (Suplementary Figure 6).

From Table 3, we provide the top two most influential SNPs that our model decided to use for prediction, rs11547464 and Affx-35293625. From the literature, we see that both are significant for

Genetics of Significant Red Hair SNPs

SNP	Location	Allele	Gene	Most severe consequence	Affected Traits
rs11547464	Chr16:89919683	G/A	MC1R	Missense variant	hair color, hair color measurement
Affx-35293625	Chr16:89986117	T/C	DEF8	Missense variant	skin of upper limb, skin of scalp and neck, malignant melanoma of upper limb, malignant melanoma of lower limb, Actinic keratosis

Table 7: Significant Red Hair SNPs and their associated location, allele (major/ minor), Gene, most severe mutation caused by variant, and affected traits as determined in the literature

red hair development. They are located in the MC1R and DEF8 genes, which are both related to skin pigmentation and hair color. While the MC1R gene is primarily known for melanin production, the neighboring DEF8 gene has a host of functions, including responding to DNA damage. DEF8's affects on pigmentation may be due to Linkage Disequilibrium.

Additionally, in Figure 2 (d), we see the direct effects of each SNP genotype on model prediction of Red Hair. We see that having two copies of the minor allele (GG) for variant rs12931267 significantly increases our additive model output, meaning that the model is more likely to predict having red hair. Having one copy slightly increases the output and having no copies does not increase output. We see a similar pattern with the Affx-35293625, where having two copies of the the minor allele significantly increases the model's prediction of red hair. The other SNPs our model deemed significant are all on chromosome 16, and are either on the MC1R gene or in a neighboring segment (DEF8, FANCA, etc.). All SNP variants shown in Supplementary Table 1 have been linked to melanin production and pigmentation in the literature.

6.5 Multitask Models

One of the benefits of NAMs compared to XGBoost is the ability to do multitask prediction. Theoretically, multitask prediction can lead to much better performance as the model is able to apply knowledge about the function from one prediction to another prediction. However, this was not the case with our multitask models. Our NAM for multitask classification of celiac disease and red hair had a test AUC of 0.677, which was much worse than both of the indivudal NAM models. Similarly, the NAM for multitask regression of total bilirubin and age at diabetes diagnosis had a test MSE of 93.623. This result fell in between the MSE for bilirubin and age at diabetes diagnosis. Although these models did not perform as well as planned, we hypothesize that the paired phenotypes do not share much information between tasks. This could cause a negative transfer between tasks given that the paired phenotypes are unrelated to each other. Further investigation between other pairs of phenotypes is needed to test this theory. For example, it may be interesting to run a multitask classification model on black hair and red hair to see if the model does in fact perform better than the individual models given that the two phenotypes are quite similar.

7 Conclusions

The Neural Additive Model offers a balance of the interpretability of linear models with the expressivity of neural networks, making it a perfect choice for medical prediction problems. Within the medical domain where human lives are at stake, it is crucial to have interpretable models that can prove that the decisions make clinical sense in order to gain the trust of medical professionals and patients. Using SNP data from the UK Biobank, we applied the NAM to four phenotypes: celiac disease, total bilirubin level, age at diabetes diagnosis, and red hair. Each of these NAM models outperformed the XGBoost baseline model and were quite accurate in MSE and AUC performace. Even more importantly, the NAM allowed us to quantitatively determine the most important SNPs for prediction. The celiac model identified SNPs such as rs3891175 and rs1265754, both of which are related to celiac in the literature. SNPs such as rs6742078 and rs887829 were found to be of high importance in total bilirubin levels. These variants are both associated with the UGT1A gene family, which is important in the glucorinidation of bilirubin. For the diabetes NAM mode, SNPs rs32658495 and rs32717773 were important for prediction, both of which are associated with the HLA class II molecules which are hypothesized to lead to damaging of insulin producing cells. Finally, the red hair NAM model most important SNPs were rs11547464 and Affx-35293625 which are both located on chromosome 16. These SNPs are associated with the MC1R and DEF8 genes which are related to hair pigmentation. It is reassuring that each of these models learned SNPs that are known to be clinically relevant. It is also important to note that not all of the important SNPs are associated with the given phenotype. However, it is possible that these SNPs could be worth investigating as potential associations with the phenotypes.

However, the multitask models for classification and regression did not perform as well as hoped. This may be due to the fact that the paired phenotypes are unrelated to each other. Future work should investigate whether having two related phenotypes can show improved performance in the multitask model compared to individual models. In addition, a combined regression and classification multitask could also increase accuracy. As it was not feasible in this project, it is also worth performing hyperparameter tuning on the number of hidden layers within each neural network. It is possible that an increase in the number of hidden layers within each network could lead to more expressivity, at the expensive of computation time.

8 Contributions

Jessica Houghton and Sasha Yousefi both worked equally on data cleaning, which includes handling missing data, genotypic data transformations, and UK biobank access. Additionally, both authors worked equally on developing principal components and other data augmentation techniques. Both authors worked equally in implementing the Neural Additive Models and the Baseline Models. Additionally, both authors worked equally on creating feature importance graphics, feature importance metrics, attaining the top SNPs contributing to each model. Each author worked on research for two phenotypes, and each contributed to this paper equally. We would like to thank Alexander Ioannidis, Daniel Mas Montserrat, and the Rivas lab for getting us access to the UK biobank dataset. Additionally, we'd like to thank Kari Hanson for introducing our research advisors to us through the course, CME 291 at Stanford. Lastly, we would like to thank the Rishabh Agarwal and the google brain team for providing the skeleton code for this Neural Additive Model.

References

[1] Rishabh Agarwal et al. *Neural Additive Models: Interpretable Machine Learning with Neural Nets*. 2020. DOI: 10.48550/ARXIV.2004.13912. URL: https://arxiv.org/abs/2004.13912.

[2] Jinyoung Byun et al. "Ancestry inference using principal component analysis and spatial analysis: a distance-based analysis to account for population substructure". In: *BMC Genomics* 18.789 (2017), p. 1. DOI: https://doi.org/10.1186/s12864-017-4166-8.

[3] Tianqi Chen and Carlos Guestrin. "XGBoost". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Aug. 2016. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145%2F2939672.2939785.

[4] Federica Farina et al. "HLA-DQA1 and HLA-DQB1 alleles, conferring susceptibility to celiac disease and type 1 diabetes, are more expressed than non-predisposing alleles and are coordinately regulated". In: *Cells* 8.7 (2019), p. 751.

[5] Jerome H Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232.

[6] *GWAS catalogue*. https://www.ebi.ac.uk/gwas/search?query=rs6025. Accessed: 2022-12-15.

[7] Steven Hicks et al. "Type II diabetes increases mitochondrial DNA mutations in the left ventricle of the Goto-Kakizaki diabetic rat". In: *American Journal of Physiology-Heart and Circulatory Physiology* 304.7 (2013), H903–H915.

[8] Arno van Hilten et al. "GenNet framework: interpretable deep learning for predicting phenotypes from genetic data". In: *Communications biology* 4.1 (2021), pp. 1–9.

[9] *HLA-DQA1 gene*. https://medlineplus.gov/genetics/gene/hla-dqa1/#conditions. Accessed: 2022-12-35.

[10] Pål V Johnsen et al. "A new method for exploring gene–gene and gene–environment interactions in GWAS with tree ensemble methods and SHAP values". In: *BMC bioinformatics* 22.1 (2021), pp. 1–29.

[11] Andrew D Johnson et al. "Genome-wide association meta-analysis for total serum bilirubin levels". In: *Human molecular genetics* 18.14 (2009), pp. 2700–2710.

[12] Michael Kaplan et al. "Neonatal Jaundice and Liver Diseases". In: *Fanaroff and Martin's Neonatal-Perinatal Medicine*. 2020. Chap. 91.

[13] Jaimy Kim and George Wu. "Celiac Disease and Elevated Liver Enzymes: A Review". In: *Journal of Clinical and Translational Hepatology* 9.1 (2021), pp. 116–124.

[14] Robert Kliegman et al. "Pediatric Pharmacogenetics, Pharmacogenomics, and Pharmacoproteomics". In: *Nelson Textbook of Pediatrics*. 2020. Chap. 72.

[15] Cathryn M. Lewis and Evangelos Vassos. "Polygenic risk scores: from research tools to clinical instruments". In: *Genome Medicine* 12.1 (May 2020), p. 44. ISSN: 1756-994X. DOI: 10.1186/s13073-020-00742-5. URL: https://doi.org/10.1186/s13073-020-00742-5.

[16] Scott M Lundberg et al. "From local explanations to global understanding with explainable AI for trees". In: *Nature machine intelligence* 2.1 (2020), pp. 56–67.

[17] Sadeep Medhasi and Narisara Chantratita. "Human Leukocyte Antigen (HLA) System: Genetics and Association with Bacterial and Viral Infections". In: *Journal of Immunology Research* 2022 (2022).

[18] Guillaume Paré, Shihong Mao, and Wei Q. Deng. "A machine-learning heuristic to improve gene score prediction of polygenic traits". In: *Scientific Reports* 7.1 (Oct. 2017), p. 12665. ISSN: 2045-2322. DOI: 10.1038/s41598-017-13056-1. URL: https://doi.org/10.1038/s41598-017-13056-1.

[19] Anna Peters and William Balistreri. "Metabolic Diseases of the Liver". In: *Nelson Textbook of Pediatrics*. 2020. Chap. 384.

[20] Serena Sanna et al. "Common variants in the SLCO1B3 locus are associated with bilirubin levels and unconjugated hyperbilirubinemia". In: *Human molecular genetics* 18.14 (2009), pp. 2711–2718.

[21] *Variant: rs6742078*. https://www.ebi.ac.uk/gwas/variants/rs6742078). Accessed: 2022-12-35.

[22] *Variant: rs9273363*. https://www.ebi.ac.uk/gwas/variants/rs9273363. Accessed: 2022-12-35.

[23] B. J. Vilhjálmsson et al. “Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores”. In: *Am J Hum Genet* 97.4 (Oct. 2015), pp. 576–592.

[24] Ting-He Zhang et al. “Transformer for Gene Expression Modeling (T-GEM): An Interpretable Deep Learning Model for Gene Expression-Based Phenotype Predictions”. In: *Cancers* 14.19 (2022), p. 4763.

[25] Stephen D Zucker, Paul S Horn, and Kenneth E Sherman. “Serum bilirubin levels in the US population: gender effect and inverse correlation with colorectal cancer”. In: *Hepatology* 40.4 (2004), pp. 827–835.