Review

Nikoletta Katsaouni, Araek Tashkandi, Lena Wiese and Marcel H. Schulz*

Machine learning based disease prediction from genotype data

https://doi.org/10.1515/hsz-2021-0109
Received January 10, 2021; accepted June 15, 2021;
published online July 5, 2021

Abstract: Using results from genome-wide association studies for understanding complex traits is a current challenge. Here we review how genotype data can be used with different machine learning (ML) methods to predict phenotype occurrence and severity from genotype data. We discuss common feature encoding schemes and how studies handle the often small number of samples compared to the huge number of variants. We compare which ML methods are being applied, including recent results using deep neural networks. Further, we review the application of methods for feature explanation and interpretation.

Keywords: deep neural networks; disease prediction; machine learning.

Introduction

Linking genotypic changes to understand phenotypic traits is an important challenge. Some disorders, such as Gaucher's disease (Cox 2001), Rett syndrome (Grillo et al. 2013) and familial hypercholesterolemia (Kastelein et al. 2020) are monogenic and caused by variants in a single gene. These diseases follow the Mendelian mode of inheritance. Linkage analysis is a group of statistical methods that use inheritance patterns in a family, in order to estimate the associations between specific regions and a trait of interest, and it is very powerful for monogenic diseases (Schote et al. 2020).

However, for complex diseases such as diabetes or cardiovascular diseases more than one gene is contributing to the disease (polygenic), and in these cases genome-wide analyses of variants are necessary. Trying to interpret the results of such genome-wide association studies (GWAS) for understanding complex traits is a current challenge and many results are available (Buniello et al. 2018). There is a plethora of large-scale efforts underway to create large resources of human genotypic variation (e.g. 1000 Genomes, UK Biobank (Sudlow et al. 2015)). Although in these GWAS, common variants (minor allele frequency >1% in a study population) are assessed, over the last years, additional attention was attributed to the analysis and interpretation of rare variants (Bellenguez et al. 2017; Hopfner et al. 2020) (minor allele frequency <1%). Despite their low frequency in a population, they can have a causal effect on phenotypes (Christophersen et al. 2017)

There is a number of problems why the genetic markup of a human individual cannot be easily linked to the likelihood of developing a certain phenotype. First of all, the predominant finding in many GWAS is that individual variants related to the phenotype have only small effect sizes. Even the combined set of all individual marker single nucleotide polymorphisms (SNPs), that are statistically enriched using genome-wide significance estimates, only explains a small percentage of the heritability of a phenotype. Heritability refers to the proportion of phenotype variability that can be accounted by genetic factors. Further, there is a high correlation between neighboring SNPs, due to the inheritance of large genomic blocks in human chromosomes. Thus, the probability that a limited set of genomic markers are sufficient to predict any complex phenotype is low by current estimates (Boyle et al. 2017).

One important aspect of using genotype data, besides finding links to the phenotype of interest, is to use them for

***Corresponding author: Marcel H. Schulz**, Institute for Cardiovascular Regeneration, Goethe University, 60590 Frankfurt am Main, Germany; German Center for Cardiovascular Research (DZHK), Partner Site RheinMain, 60590 Frankfurt am Main, Germany; and Cardio-Pulmonary Institute, Goethe University Hospital, Frankfurt am Main, Germany, E-mail: marcel.schulz@em.uni-frankfurt.de
Nikoletta Katsaouni, Institute for Cardiovascular Regeneration, Goethe University, 60590 Frankfurt am Main, Germany. https://orcid.org/0000-0001-8323-2824
Araek Tashkandi, Institute of Computer Sciences and Engineering, University of Jeddah, 21959 Jeddah, Saudi Arabia
Lena Wiese, Institute of Computer Science, Goethe University, 60629 Frankfurt am Main, Germany

the design of polygenic risk scores (PRS). The idea of a polygenic risk score is to estimate the tendency of a human to develop a certain phenotype, a disease or response to a therapy using only genotypic summary information. PRSs have gained reputation for the discovery of complex traits (Dudbridge 2013).

A PRS computes the sum of risk alleles for an individual, weighted by the risk allele effect sizes obtained from GWAS summary statistics, that is the information content of one SNP with respect to the phenotype of interest (Choi et al. 2020). A PRS assumes a linear combination of markers and often assumes independence of those. However, there are also approaches that take into account the correlation structure between neighboring SNP markers, and standardized software for their creation has been developed, e.g., LDpred (Privé et al. 2020; Vilhjálmsson et al. 2015). The authors of Choi et al. (2020) present a practical approach to PRS analysis by showing and discussing the end-to-end analysis pipeline. Choi et al. (2020) also discuss the special cases that no (or not sufficient) GWAS data for the target phenotype are available – in which base traits similar to the target trait can be analyzed – as well as multi-trait analyses. The authors stress the importance of verifying the validity of statistical assumptions underlying PRS analysis – for example, the assumption that SNPs are uncorrelated. As a PRS assumes a linear combination of markers, it makes it inadequate to capture non-linearities between the variants (Levine et al. 2017). While in some cases the linear additive effects are enough to generate well-performing risk prediction models (Gola et al. 2020), comparative studies have shown that the application of machine learning techniques can improve the performance of PRS by weighting the contribution of individual variants in a data-driven way (Paré et al. 2017).

While Bracher-Smith et al. (2020) provides a survey with a focus on psychiatric disorders, other papers review GWAS and their effects on disease development from a health management point of view. In Torkamani et al. (2018) the authors address PRSs for adults, focusing on four diseases that can be caused by genetic inheritance. But, the complexity of GWAS and SNP analysis has also spurred criticism regarding the applicability of PRSs in clinical practice (Wray et al. 2021). Accordingly, it is debated whether large-scale clinical use of conventional PRSs with appropriate reliability of the outcomes will be possible in the near future.

Other methods are tailored to predict the existence of a disease or disease subtype directly from the genotype data, not the GWAS summary statistics only. These methods often rely on machine learning methods for classification and regression and include established approaches such as logistic regression (LR), commonly used by GWAS for the estimation of the impact of individual SNPs, support vector machines (SVMs) and methods based on decision trees (DTs) such as random forests (RF) or tree-bagging approaches (López et al. 2018; Oriol et al. 2019; Wei et al. 2013). In order to cope with learning problems, where the number of SNPs is much larger than the number of samples, feature regularization, such as LASSO or forward selection, is being applied (Oriol et al. 2019; Wang et al. 2018).

The recent advances in deep learning methods have spurred applications in biomedicine, and although promising, their use has been controversial due to problems such as limited interpretability (Ching et al. 2018). However, deep neural networks (DNNs), which consist of many layers of neurons with non-linear activation functions, have been proven to outperform state-of-the-art performances in many domains due to their flexibility and generalization power. This includes approaches such as convolutional neural networks (CNNs), recurrent neural networks (RNNs) or autoencoders. Increasingly, different types of DNNs are applied for the problem of disease prediction from genotype data (Montanez et al. 2018; Yin et al. 2019). The dimensionality issue in most of the studies is addressed by filtering the genomic information using prior knowledge and GWAS summary statistics. This can be achieved by statistical methods and ML models. Thus, different approaches from statistical genetics and machine learning have been proposed to summarize the effect of large sets of genomic variants to explain a phenotype. In order to achieve a wide acceptance, several questions remain: What are the best mathematical and algorithmic approaches to exploit genotypic information for disease prediction? How to address problems, such as the huge amount of interacting genetic variants resulting in highly non-linear associations with the phenotype?

In this review, we have collected recent papers that solve the problem of predicting disease risk or phenotypic prevalence from SNP data. The focus is on studies that apply ML methods, including recent approaches with deep neural networks. We review how feature selection of informative SNPs is addressed and how SNP information is encoded. We detail which ML methods are used and how DNNs compare to more established ML methods. Finally, we investigate whether studies use methods for feature explanation and interpretation to highlight genomic variants of interest.

Overview of disease prediction models using genotype data

In the upcoming sections, we survey many of the used approaches for SNP analysis. We start with an overview of feature selection and encoding methods applied on the data, followed by a review of specific ML methods for the actual analysis. While some papers favor one ML model, other papers aim for a more unbiased comparison of different approaches; that is why comparison papers are surveyed in their own subsection. Last but not least – going beyond the mere analysis – we cover (the only few) approaches that consider explainability and interpretability issues after the analysis.

The main steps for the construction of disease prediction models are illustrated in Figure 1. Firstly, the SNPs are preprocessed in a way that less important ones are filtered out, missing values are handled properly and data of uncorrelated individuals are preserved. Hence, the initially large amount of SNPs is limited by retaining only a sufficiently relevant subset. This can be accomplished with classical statistical methods as well as machine learning models. Afterward, in the second step the genomic data has to be encoded in a way that is convenient for the model of interest and facilitates the learning process. Finally, in the third and fourth steps the ML models are constructed, evaluated and probably optimized iteratively. Each of the depicted steps in Figure 1 is described thoroughly in the following section.

SNP preprocessing and selection using statistical methods

Data preprocessing and appropriate feature selection have a decisive role in the accurate performance and predictability of machine learning algorithms (Okser et al. 2014). Denoising, filtering of redundant or irrelevant information and dimensionality reduction are some strategies commonly used (Kruppa et al. 2012; Shi et al. 2016). However, the type and the necessity of feature selection and transformation are highly dependent on the machine learning model. The preprocessing algorithms of the reviewed studies select relevant SNP subsets by direct application on the raw genotype data obtained per individual, or make use of the summary statistics p-values as computed by a GWAS. It is often observed that deep learning approaches (e.g. CNNs) tend to demand less data transformation (López et al. 2018). Although there is a plethora of machine learning models, the choice of the preprocessing strategy is mainly restricted to the techniques described in Table 1. These methods are explained in more details below:

Linkage disequilibrium (LD) is quite often used to quantify the correlations between alleles and to identify whether a 'haplotype' is presented in a disproportionately large percentage in the population (Slatkin 2008). The development of next generation sequencing (NGS) technologies has enhanced the identification of SNPs (Davey et al. 2011), such that the use of methods that consider LD has attracted more attention and software for the facilitation of the analysis has been developed (Machiela and Chanock 2015; Zhang et al. 2019). The significant interactions for the corresponding disease of interest are chosen according to the GWAS studies such as in Bellot et al. (2018) and the p-values are set accordingly. In Oriol et al. (2019) LD-based clumping (p-value ≤ 0.01, $r^2 \leq 0.05$) is used to find the statistically relevant SNPs which are in LD, while in Montanez et al. (2018) multiple p-value thresholds were used to assign four different groups of SNPs (5 SNPs (1×10^{-5}), 32 SNPs (1×10^{-4}), 248 SNPs (1×10^{-5}) and 2465 SNPs (1×10^{-2})) and used a Bonferroni correction for the statistical tests.

In order to estimate the distributions of genotype frequencies of two alleles of one gene locus the **Hardy-Weinberg equilibrium** (Mayo 2008) is calculated. This preprocessing step can be used to further clean the data and take into consideration the genotype information that is 'stable' across generations (Gaudillo et al. 2019; Oriol et al. 2019).

Although in some cases the missing SNPs are treated as an additional feature (López et al. 2018) to be handled by the deep learning model, most of the studies are applying **missing call filtering** to exclude the non-informative SNPs (Gaudillo et al. 2019; Ghafouri-Fard et al. 2019; Pirmoradi et al. 2020; Romagnoni et al. 2019). A recent study showed that even though Random Forests are able to deal with missing information properly, there are weaknesses when

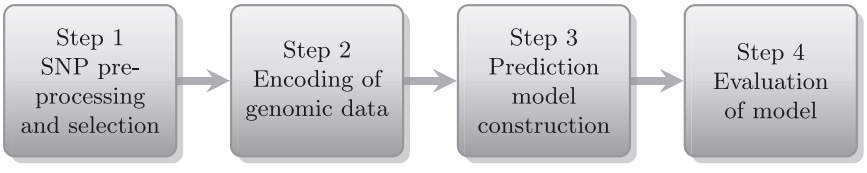

Figure 1: Workflow for the construction of disease prediction models using genotype data.

Table 1: Use of different SNP selection and association analysis techniques for genotype data.

	López et al. (2018)	Wei et al. (2013)	Montanez et al. (2018)	Gaudillo et al. (2019)	Mieth et al. (2020)	Wang et al. (2018)	Xu et al. (2020)	Romagnoni et al. (2019)	Bellot et al. (2018)	Pirmoradi et al. (2020)	Oriol et al. (2019)	Yin et al. (2019)	Sun et al. (2020)	Ghafouri-Fard et al. (2019)	Badré et al. (2020)
Missing call filtering	✓			✓			✓	✓	✓					✓	
Logistic regression		✓	✓		✓			✓							✓
Minor allele frequencies		✓					✓	✓	✓		✓		✓		
Hardy-Weinberg equil.				✓							✓				
Chi-square test				✓	✓										
Random forest (OOBE, mean)				✓											
Decision trees						✓									
Genetic algorithms						✓									
Linear regression						✓					✓				
Redundancy filtering										✓					
Linkage disequilibrium*											✓				
Identity-by-descent											✓				
Convolutional neural networks												✓			
Principal component analysis			✓												

*The inputs to the algorithm are the significance *p*-values of GWAS summary statistics.

their ratio is more than 25% (López Ibáñez et al. 2016). The arbitrary selection of samples to generate the decision trees can not only increase the generalization ability of the Random Forest, but also reduces in this way the amount of missing information to be handled at each level.

The distinction between common and rare variants can be accomplished using **minor allele frequency (MAF) and marker call rate** as in Oriol et al. (2019), Romagnoni et al. (2019), Wei et al. (2013), and Xu et al. (2020). Additionally, in Gaudillo et al. (2019) they added a hypothesis testing step by using the **Chi-square test**, a non-parametric test to measure statistical significance for categorical variables. Similarly, Mieth et al. (2020) applied a **permutation based threshold calibration** to retain the important SNPs.

Finally, removing correlated individuals can favor the performance of the model. In Oriol et al. (2019) the **Identity-By-Descent (IBD)** estimation is performed with a correlation threshold of 0.25 to find the individuals which are related to each other. Similar samples are removed afterwards.

Even though not described in the studies that use ML approaches for disease prediction, there are additional preprocessing techniques that are well established in GWAS, such as Cochran-Armitage trend test, failure rate, and heterozygosity rate estimation (Anderson et al. 2010; Ani et al. 2021).

SNP selection using ML methods for association analysis

Before we review common methods for SNP selection, we would like to point out a possible source of confusion. Some of the methods mentioned below are also used for the computation of GWAS *p*-values, such as logistic regression. In fact, in a GWAS context each variant is assessed regarding its enrichment in the population with the phenotype. The set of these individual *p*-values are denoted as GWAS summary statistics. Obviously, there is a direct relation to the task of feature selection for an ML model, in which strong features are the ones that separate the two groups of interest. Thus, we found that people that use ML models sometimes use related approaches for the SNP selection.

In Badré et al. (2020), Montanez et al. (2018), and Wei et al. (2013), the authors used **Logistic Regression (LR)** as preprocessing step in order to select the most informative SNPs by estimating the relevance scores. In other words, the regression coefficients for each SNP attribute are determined according to their impact on phenotype and the more significant SNPs are obtained. In Mieth et al. (2020) the LR preprocessing step is also followed by an average filter with a well-defined window size, to calculate the global prediction relevance scores. Montanez et al. (2018) applied LR under an additive model, as they claim that the additive models are able to determine the additive and dominant effects accurately.

The **stepwise Multiple Linear Regression (MLR)** was chosen in Wang et al. (2018) where the SNPs with larger coefficients are selected. Forward stepwise selection is based on significance, while simultaneously the insignificant ones were excluded.

Tree-based machine learning algorithms are a widely used ML category. According to Saeys et al. (2008), the tree-based methods outperform other selection techniques when high-dimensional data should be handled. Based on this claim (Gaudillo et al. 2019), used RFs to evaluate the SNPs according to the disease phenotype. The importance of each input was estimated by comparing the mean and the out-of-bag error (OOBE) over the forest before and after permutation. Another tree-based approach was used in Wang et al. (2018), where the correlations were estimated using DTs. This was accomplished by calculating the difference of each attribute before and after the DT. The SNPs with the highest information gain were selected for the prediction model.

The authors of Wang et al. (2018), in the context of Ensemble Models for feature selection, used **genetic algorithms (GA)** as well. The parameters which were optimized by the GA are the attributes (SNPs) of the chromosomes. The reproduction (in terms of GA training) was repeated until the fitness function converged for a given number of chromosomes and the performance was estimated using fivefold cross-validation.

A recent study Yin et al. (2019) has used a **CNN** in order to select the appropriate promoter regions, which are used afterwards by their predictive model. In total 64 SNP positions upstream and downstream (56 and 8 accordingly) of each promoter region were used for the Promoter-CNN to determine the genome regions, which are highly related to the disease (in this case the ALS) phenotype. The decoupling between the feature selection and the disease specific domain knowledge can propose a model which is efficiently generalised to genome-sized data. The single promoter classifier of Yin et al. (2019) identified some of the already known ALS disease genes as well as some novel risk factors. Although the selected promoter regions had a positive impact on the predictive model's performance, no biological justifications for the direct association with the disease are provided.

Finally, in Montanez et al. (2018) the dimensionality reduction and the identification of outliers is done with the **principal component analysis (PCA)**.

Encoding of genotype information

It is essential that the genotype information is mapped into an interpretable form for the machine learning models. The genotypes can be one of AA, Aa and aa to denote homozygous major alleles, heterozygous alleles or homozygous minor alleles, respectively. These are mostly encoded according to one of the following strategies:

- In Montanez et al. (2018), Badré et al. (2020), and Xu et al. (2020) the genotypes at the typed locus can take one of the integer values in the range [0, 2]. The same genotypic encoding was used in Gaudillo et al. (2019), with the difference that the values are now chosen to be in the range [1, 3].
- A different approach was applied for the DeepCOMBI model in Mieth et al. (2016). In this application the genotypes are encoded in a binary system with three bits. Therefore, the three possible inputs can be 100, 010 or 001.
- Pirmoradi et al.(2020) decided to use the information of the target label in the encoding strategy, which is not considered by the above mentioned techniques. Therefore, they applied the mean encoding technique. In mean encoding each encoded SNP (or feature) i is defined as $\frac{\text{Number of true targets under feature } i}{\text{Total number of targets under feature } i} = \frac{\text{Number of cases for SNP i}}{\text{Number of instances for SNP i}}$. This approach can be beneficial for multi-dimensional data (in contrast to other methods) takes into consideration the phenotype during the encoding. The encoded values are assigned to the SNPs considering the separability of each class (case or control) in each specific dataset.

Handling the missing values of SNPs is another aspect to be considered. While many studies, as it is described in Section SNP selection using ML methods for association analysis, use missing call filtering to exclude the missing, non-informative SNPs, Lopez et al. (2018) used the missing (or unknown) values as another attribute and assumed that the RFs can handle this information. Finally, Sun et al. (2020) proposed to encode SNP data in the range [0, 2] and divide the encoding by 2, such that their final representation lies in the range [0, 1]; the authors claim that this transformation can boost the performance of their neural network model.

Although missing SNPs can be handled by the before mentioned techniques, in many cases the inference of the missing ones can be valuable. A statistical method to infer the not genotyped SNPs, often named as "hidden", is imputation (Halperin and Stephan 2009; Marchini and Howie 2010). Imputing genotypes to overcome the restrictions of the current technologies and to increase the resolution of the study dataset is mainly performed in two steps: prephasing and imputation. During the pre-phasing process the haplotypes for each individual are estimated. During the imputation step the alleles of the "hidden" SNPs are inferred by the available SNPs with the use of a reference population, such as HapMap Gibbs et al. (2003). There are several algorithms for imputation such as IMPUTE2 (Howie et al. 2012), TUNA (Nicolae 2006), Beagle (Browning and Browning 2009) and neural network approaches (Sun and Kardia 2008).

Machine learning models for disease prediction

Different Machine Learning (ML) models can be leveraged for disease prediction from genotype data. In the following, we present a literature review of the commonly used ML models for predicting different diseases. A summary of the reviewed studies can be found in Table 2.

Neural networks

For predicting amyotrophic lateral sclerosis (ALS), Yin et al. (2019) used Deep Learning. A deep CNN is employed to identify eight genomic promoter regions associated with ALS (Promoter-CNN) per chromosome. Moreover, it is used to classify the patients based on a combination of promoter regions (ALS-Net). The Promoter-CNN has two convolution layers, followed by two dense layers. ALS-Net's design is based on the structure of genome data. Its architecture and hyperparameters are optimized using nine-fold cross-validation and the AdaGrad algorithm separately for each chromosome. The performance of their approach was assessed by testing ALS-Net. It was compared against logistic regression (as a basic PRS), SVM, RF, and Adaptive Boosting (AdaBoost). All the models use the same promoter regions used in their ALS-Net model.

Badré et al. (2020) used a DNN for estimating polygenic risk scores for breast cancer risk. They compared the performance of a DNN with other models: best linear unbiased prediction (BLUP), BayesA, and LDpred. DNN outperformed the other models in estimating the breast cancer PRS with 67.4% of Area Under the receiver operating characteristic Curve (AUC). They found that the DNN has a high accuracy in assigning genetic risk degree, either high or normal.

Table 2: Summary of ML models for disease prediction using genotypic data. In the Table the most used metrics (at least in two studies) are displayed for the evaluation of the performance. The studies are displayed in chronological and alphabetical order.

	Dataset					Metrics				
	Disease	Method	Cases	Controls	Dataset size	Accuracy	Precision	AUC	Sensitivity	Specificity
Wei et al. (2013)	Inflammatory bowel disease	LR	13,458	22,442	35,900	–	–	86.4	–	–
Shaik Mohammad et al. (2016)	Autism spectrum disorders (ASD)	ANN	138	138	276	63.8	–	72	–	–
López et al. (2018)	Type 2 diabetes (T2D)	RF	248	429	453	–	–	89.0 ± 0.04	–	–
Montanez et al. (2018)	Obesity	DNN	962	1192	2154	–	–	99.08	96.04	97.12
Wang et al. (2018)	Obesity	SVM/kNN	74	65	139	67	–	70	72	62
Gaudillo et al. (2019)	Asthma	RF-SVM	–	–	143	62.5	65.3	64	69	–
Ghafouri-Fard et al. (2019)	Autism spectrum disorders (ASD)	DNN	487	455	942	73.67	–	80.59	82.75	63.95
Yin et al. (2019)	Amyotrophic lateral sclerosis (ALS)	CNN	4511	7397	11,908	76.9	71.1	–	–	–
Badré et al. (2020)	Breast cancer	DNN	26,029	23,082	49,111	60.1	–	65.9	–	–
Sun et al. (2020)	Age-related eye diseases (AMD)	DNN	2341	5462	7803	–	–	81.8	–	–
Xu et al. (2020)	Smoking behaviour	SVM	213	224	437	–	–	89.7	–	–

Montanez et al. (2018) use a DNN with the association results of GWAS to predict obesity. First, LR is implemented for feature selection between each SNP and obesity phenotype. Second, they select four sets of SNPs with different p-value association thresholds (as already mentioned above) as features to initialize a multi-layer feedforward neural network. To classify the obesity cases, the DNN classifier is trained with stochastic gradient descent using back-propagation. Moreover, an adaptive learning rate is used for tuning the network architecture and the regularization parameters for each SNP set. Instead of cross-validation, the dataset is randomly split into 60, 20, 20% for training, validation, and testing, respectively. Deterioration was observed in the model performance by decreasing the SNP number (i.e., increasing p-value threshold) and progressively performance improvement with more SNPs. This case explained that with only highly ranked SNPs, the ML model has a limited predictive capability to classify between case and control. The best performance was achieved with p-value smaller than 1×10^{-2} (2465 SNPs). In contrast, the five SNPs with p-value smaller than 1×10^{-5} produce the worst performance. They recommended further improvement by using an enhanced way of identifying the reduced sets of genetic variants and stated stacked autoencoders as an approach.

With the aim of evaluating the negative effects of different supplements in the development and amelioration of age-related eye diseases (AMD) Sun et al. (2020) used a DNN for a total of 7803 samples. A grid search was conducted to optimize the network parameters and the training was done using 10-fold cross-validation. The DNN was compared to LASSO, Ridge, random survival forest (RSF), and the benchmark genetic PRS and it was shown to perform best.

A DNN was also used by Ghafouri et al. (2019) for the diagnosis of Autism Spectrum Disorders (ASD). Concerning the training, SNP information from 487 patients was used, using 10-fold cross-validation and the L2 regularizer. A total number of 15 SNPs within four genes highly associated with ASD were included in the model. The performance of the proposed model outperformed the previous similar study of Shaik Mohammad et al. (2016) where a smaller accuracy was achieved.

Pirmoradi et al. (2020) proposed a self-organizing auto-encoder (SOAE) which is able to construct its architecture automatically in an optimal manner, avoiding in

this way over-fitting and reducing the computational cost of a large hyperparameter grid search. The SOAE was applied for five different disease prediction scenarios: Thyroid Cancer, mental retardation, breast cancer, colorectal cancer and autism. The combination of the mean encoding (see above) as feature selection step with the SOAE outperformed previously published works for some of the cases. However, it should be mentioned that the number of samples was limited (between 111 and 567) and the generalization performance to independent disease cohorts was not investigated.

Random forest

López et al. (2018) used RF to identify relevant SNPs related to Type 2 diabetes (T2D). Even though the used SNP dataset has a significant percentage of missing data, some SNP information for many samples was missing, RF managed to produce high prediction performance. Specifically, an extra attribute value was added at every SNP, performing a kind of a unique-value imputation method. There was no optimization for RF hyperparameters (the number of decision trees is fixed to 1000). Three experimental scenarios are used: the raw data, SNP data with clinical information (sex, Body Mass Index (BMI), and age), and data considering the SNP interactions by defining SNP relevance values. The SNP-value relevance strategy suggests replacing each SNP with three boolean variables that resemble the allele combinations (AA, Aa and aa), in addition to the before-mentioned attribute that indicates the existence or not of the SNP for the current sample. Using the combined SNP data with clinical information produces the highest performance. They compared the RF performance with SVM and LR on the three experimental scenarios by 10-fold cross-validation and AUC. RF prediction performance outperforms them. The problem of imbalance in the data was handled by balancing the classes for every fold. Relevance values obtained for the SVM and LR were similar but different from the RF. Furthermore, SVM and LR have a high variability of the relevance values obtained, whereas RF has the lowest.

Support vector machines

SVMs have been widely used for disease prediction based on genotype information. Gaudillo et al. (2019) constructed a SVM model to predict an individual's susceptibility to asthma. Before the prediction by the SVM, the high dimensionality of the data was reduced using RF. The hyperparameters were optimized by testing multiple combinations in a given range of values and the final RF-SVM model was compared with the RF-kNN by the leave-one-out cross-validation (LOOCV). The performances of the models were estimated by the accuracy, precision, sensitivity and the ROC curve as metrics. The combination of RF as feature selection with SVM as classifier outperformed the baseline SVM, where the SVM was applied on the whole dataset and not on the one selected by the RF SNPs. In this way, the combination of RF and the SVM was promoted. Gaudillo et al. (2019) also mentioned that the SVM is more powerful for high-dimensional data compared to a kNN classifier.

A similar approach was followed in Wang et al. (2018), where again SVM and kNN models were compared but now for the prediction of obesity risk. The RBF kernel was used for the SVM model and the estimation of the hyper-parameters was accomplished using a 5-fold cross-validation. The performance of the models was a result of the sensitivity, specificity, accuracy and Matthews correlations coefficient. The SVM (calculated over 5-fold cross-validation) outperforms again the kNN approach as well as the DT.

A recent study Xu et al. (2020) has applied SVMs with linear kernels on SNPs in order to predict smoking behaviour. Smoking is a behavior that not only has genetic factors but is also strongly influenced by environmental conditions. The results were again based on a 10-fold cross-validation. Additionally, the SVM had better performance in comparison to the RF model for multiple number of SNPs.

Logistic regression

Wei et al. (2013) implemented a risk assessment for the two common forms of inflammatory bowel disease (IBD): Crohn's disease (CD) and ulcerative colitis (UC). Penalized LR with the L1 penalty was applied for this prediction task. The dataset is randomly divided into three folds. The first fold was used for SNPs pre-selection, the second fold for model training, and the third fold for model testing. Ten-fold cross-validation is used for selecting the best value for the penalty parameter. The resulting AUCs from training and testing are similar. They claim that their achieved prediction performance is the best ever reported for CD and UC. Moreover, they compared the AUC of LR with SVM with RBF kernels and gradient boosted trees (GBT). They found that SVM gave a comparable AUC and GBT gave a low-performance AUC.

Table 3: Overview of results in comparative studies.

	Disease	ML methods	Optimal model	Hyperparameter optimisation	Best model
Oriol et al. (2019)	Alzheimer's disease	BSWiMS, NB, RF, KNN BSWiMS features, LASSO, RPART, SVM mRMR features, ensemble of all	Ensemble model	None	AUC = 0.719
Aguiar-Pulido et al. (2010)	Schizophrenia	LNN, MLP, RBF, EC, MDR, Bayesian networks, SVM, decision tables, DTNB, BFTree, AdaBoost	LNN and MLP	None	AUC = 0.9439
Bellot et al. (2018)	Human traits	CNNs, Bayesian linear models, and MLPs	BayesB for height trait; CNN and Bayesian models comparable for other traits	DeepEvolve for CNNs and MLPs	Correlation = 0.47
Romagnoni et al. (2019)	Crohn's disease	LR, GBT and NN	Models had similar accuracy, GBM most robust	10-fold cross-validation	AUC close to 0.80

BSWiMS, Bootstrap Stage-Wise Model Selection; RPART, Recursive Partitioning and Regression Trees; mRMR, minimum-Redundancy-Maximum-Relevance; RBF, Radial Base Functions; EC, Evolutionary Computation; MDR, Multifactor Dimensionality Reduction; DTNB, Decision Table Naïve Bayes Hybrid Classifier; BFTree, Best-First Decision Tree classifier.

Comparative studies

Some studies provide a comparison between different ML models to find an optimal model for a specific disease prediction task. We provide a review of these studies. Moreover, in Table 3, we compare the surveyed works on the used ML models, the optimal ML model, and the implemented hyperparameter optimization along with the best-achieved performance.

Oriol et al. (2019) conducted comparisons of Machine Learning models for predicting Late-Onset Alzheimer's Disease (LOAD) from genetic data and for identifying the genetic markers associated with the risk of LOAD. They used the benchmark tool FRESA.CAD (Feature Selection Algorithms for Computer-Aided Diagnosis) to compare: Bootstrap Stage-Wise Model Selection (BSWiMS), Least Absolute Shrinkage and Selection Operator (LASSO), Naive Bayes, RF, K Nearest Neighbors (KNN) with BSWiMS features, Recursive Partitioning and Regression Trees (RPART), SVM with minimum-Redundancy-Maximum-Relevance (mRMR) feature selection filter, and the ensemble of all these methods. They were evaluated by cross-validation, the balanced error, the AUC, the accuracy, specificity, and sensitivity. They found that SVM with mRMR filter had the lowest performance, and the best ML model was LASSO (AUC = 0.703). However, the methods' ensemble produces the best performance (AUC of 0.719, sensitivity of 70%, and specificity of 65%).

Aguiar-Pulido et al. (2010) compared 12 ML techniques for classifying schizophrenia disease. They focus on using SNPs at the two most studied genes in relation to schizophrenia DRD3 and HTR2A (each of them or both). The 12 ML techniques are Linear Neural Network (LNN), Multilayer Perceptron (MLP), Radial Base Functions (RBF), Evolutionary Computation (EC), Multifactor Dimensionality Reduction (MDR), Bayesian Networks, SVM, Decision Tables, Decision Table Naïve Bayes Hybrid Classifier (DTNB), Best-First Decision Tree classifier (BFTree) and AdaBoost. The used dataset contains 260 positive subjects and 354 negative subjects. They obtained six further datasets from it by adding different percentages of simulated negative subjects. Simulated subjects were made to address missing data, but it is not explained why only negative subjects were simulated. In total, six simulated and the real datasets were used. They implemented 252 Quantitative Genotype – Disease Relationship (QGDR) classification models out of these 12 machine learning techniques, seven datasets, and the three states of SNPs at two schizophrenia-related genes. The models were tested by the 10-fold cross-validation. The best classification accuracy was achieved by LNN using SNPs of both genes. They stated two examples of the best LNN models, the first model with 40 inputs and 152 neurons (AUC= 0.8405), and the second LNN with two inputs and eight neurons (AUC = 0.9439). However, details on the used models' parameters were missing, and no hyperparameter optimization is mentioned. MLP produces almost the same classification accuracy as LNN. We noticed that the largest percentage of simulated negative subjects produces the highest classification accuracy which might be a sign of overfitting.

Bellot et al. (2018) compared CNNs, MLPs, and Bayesian linear models (BayesB, and Bayesian Ridge Regression (BRR)) for genomic prediction of complex human traits. Five phenotypes or SNP sets of the considered traits are height, bone heel mineral density (BHMD), body mass index (BMI),

systolic blood pressure (SBP), and waist-hip ratio (WHR). A genetic algorithm implemented by the DeepEvolve framework is used for hyperparameter optimization for CNNs and MLPs independently for each trait. From 10 to 50 k set of SNPs, preselected by single-marker regression analyses, were used for comparing the models. The models' prediction accuracy is evaluated by correlation of the predicted phenotype value with the observed phenotype measurement in the test set, e.g. the height of a person. BayesB achieved the best prediction performance in the height trait and was followed by BRR and MLPs. CNN models were the worst-performing ones in the height trait. For the other traits, the performance of the CNNs is either comparable or they insignificantly outperformed BayesB and BRR models.

For predicting Crohn's Disease Romagnoni et al. (2019) compare linear and non-linear ML models. Different penalized LR, GBT, and DNN models were applied. The used LR regularizations were Lasso (L1), Ridge (L2), and ElasticNet (combined L1 and L2). Three GBT implementations were compared: XGBoost, LightGBM and CatBoost. A feedforward fully connected dense DNN was employed with residual connections, dropout, batch normalization and with varying number of hidden layers and neurons. All the models' hyperparameters were optimized by 10-fold cross-validation on the training dataset, where the selected optimal values are the ones that maximize the AUC values. They tested the impact of different pre-processing steps such as Quality Control (QC) methods and coding strategies with Lasso LR. They discovered that imputation of missing genotypes and QC methods increased the classification accuracy of LR. They observed that non-linear ML models produced higher AUC for CD prediction than linear models. They hence claim that non-linear models can capture the non-linear epistatic interactions between genotypes. However, both the linear and non-linear models produce AUC values close to 0.80; they justify this by the limited epistatic effects in the genotyping dataset. The classification AUC of the neural network with a single hidden layer did not improve by increasing the number of hidden neurons. Similarly, the AUC of the DNN with multiple hidden layers did not improve by increasing the number of hidden layers. Both types of neural network approaches obtained almost the same AUC. The different algorithms of GBTs achieved the same range of AUC values as the other DNN implementations (AUC = 0.80).

Explainability and interpretability

Explainable Artificial Intelligence (X-AI) covers approaches to make machine-based reasoning accessible and reviewable to human judgment. One demand in this direction is to prefer interpretable models over black-box models (Rudin 2019). For the scope of this review paper, the explainability focuses on the ML-based identification of SNPs relevant for the disease under investigation. With an importance score for each SNP the researchers can quantify the influence that the SNP has for the particular disease. We observed that some authors indeed considered explainability of ML models and hence interpretability of ML decisions in their articles.

Another study López et al. (2018) has investigated the influence of SNPs for risk prediction by assigning a feature importance score. In particular, they compute relevance values for each SNP by the Gini importance. After applying one-hot encoding the authors conclude from their study that information on the presence of SNPs is more relevant than their impact on the gene expression. A comparison of relevance values obtained for LR and SVM shows the better stability of relevance values obtained by RF.

As a main feature importance criterion, in Romagnoni et al. (2019) the authors consider permutation feature importance (PFI): by distorting the original data set by permuting values of one feature and observing the outcome on the prediction accuracy. Additionally the authors also consider as importance criteria the weights for the individual features obtained by logistic regression and the split gain of gradient boosting trees. Notably, the authors observed a certain variability of the features identified as important depending on the subset of the data used for training; in a similar setting, this effect was also observed in Oriol et al. (2019). The authors of Romagnoni et al. (2019) counter this with a repeated random permutation approach and considering feature importance individually for each permutation. Another analysis of the authors considers robustness with regard to the identified loci (containing important features). They define a model's robustness in terms of the detected loci that contain SNPs, which appear to be correlated with the phenotype, as resulted by training the model on different data folds. Even though linear and non-linear models reached similar classification accuracy, they found the non-linear models robust in identifying and classifying genetic markers. LGBM (with gain and PFI) produced very similar robustness among the different data folds: Several runs of the algorithm on different folds identified a high ratio of common loci.

The COMBI approach Mieth et al. (2016) uses an SVM to obtain a relevance value for each SNP and then applies a hypothesis testing on the ones with a high score. Extending this approach (Mieth et al. 2020), use neural networks and then apply layerwise relevance propagation (LRP); by averaging over all sample-wise explanations, a global

score for each SNP is obtained. Finally association of highly scored SNPs with the investigated SNPs is statistically tested.

As a general tool, the LIME method (Ribeiro et al. 2016) provides explanations for a complex model by approximating it locally (for a single prediction) with a surrogate model. The interpretable surrogate model is trained on a perturbed dataset (in order to vary the input data samples) where closeness to the data sample for which an explanation is seeked is taken into account. Several articles (Badré et al. 2020; Ghafouri-Fard et al. 2019; Sun et al. 2020) apply LIME to obtain feature importance at sample level for neural networks. In this way they identify SNPs that either have a protective or a harmful effect on the investigated disease. In Sun et al. (2020), the authors furthermore split the data set into two subgroups (high risk vs. low risk) and observe differences in the features importance between the two subgroups.

In addition to LIME, Badré et al. (2020) also apply DeepLift (Shrikumar et al. 2017). Notably only one third of the SNPs were determined to be relevant by both interpretation tools (LIME and DeepLift) at the same time. The authors attribute this to the fact that LIME is model-agnostic while DeepLift uses internals of the Neural Network. Interestingly, only roughly one 10th of SNPs with association to breast cancer (as reported in previous studies) were indeed identified by the NN model.

In a recent study (Orlenko and Moore 2021), focus on Random Forest on GWAS addressing Glaucoma and Alzheimer's disease and compare SNP importance obtained by three metrics to importance scores obtained by a HIBACHI sensitivity analysis. The three tested metrics were (1) PFI scores, (2) built-in feature importance coefficients (BIC) by exploiting the RF-internal performance optimization (entropy or Gini importance), as well as (3) mean SHAP (Lundberg and Lee 2017) values. PFI values were superior in determining the correct ranking of features according to their importance for the prediction. Yet, the authors also state that PFI is sensitive to correlations in the data.

Discussion and conclusion

As kind of a baseline approach polygenic risk scores (PRSs) are gaining increasing attention and it is often highlighted that application of PRSs in clinical practice can improve individualized disease prediction. Moving away from conventional PRSs, which only use GWAS summary statistics, Ho et al. (2019) compare PRSs to ML models and argue that ML methods can handle the data analysis with less statistical assumptions; for example, in terms of correlations between SNPs. On this background, with the survey presented in this paper we hence aim to give an overview of recent trends in the application of ML methods for SNP analysis. Based on our comparative survey we want to underline the potential of ML methods to improve disease prediction and investigate the complex interaction of SNPs and phenotype. However, we can also observe certain shortcomings in the current status of how ML methods are applied for this purpose:

Data set sizes

Many authors discuss in-depth the fact that importance of individual SNPs is difficult to assess because correlated or interacting SNPs have a combined effect on disease risk. This problem could indeed be addressed by modern (and explainable) machine learning methods applied to much larger GWAS data sets than are used in the reviewed papers such that optimally novel insights regarding the influence of subsets of SNPs on certain phenotypes could be revealed.

In particular, the above mentioned preprocessing (feature selection and encoding) strategies should be carefully analyzed with respect to whether they unintentionally reduce the information contained in the GWAS data sets.

Bias in data sets

Utility of ML for disease risk prediction is in fact influenced by the validity of the underlying GWASs. Many authors criticize the bias on existing GWAS that mostly focus on individuals of European ancestry. This problem is already identified in applications of conventional PRSs (Lewis and Vassos 2020) – but evidently also has an effect on the generalizability of ML-based predictions (Chen et al. 2020). As is the case in numerous other ML-based decision support systems, the application of ML in healthcare and medicine needs a careful choice regarding the training data sets.

Explainability and feature importance

The variability of feature importance (depending on subsets chosen for training) reported in Romagnoni et al. (2019) and Oriol et al. (2019) is a fact that has to be considered in ML-based SNP analysis approaches and has to be investigated more thoroughly. Similarly, the fact that different explainability tools in Badré et al.

(2020) and Orlenko and Moore (2021) come to divergent feature importance values for the same data set shows the need for a more in-depth analysis. In particular, more flexible methods like SHAP (Lundberg and Lee 2017) (that offer the option to aggregate local interpretations into global ones) will have to be analyzed in more detail when applied to SNP analysis. Because the X-AI field is evolving rapidly, the development of novel methods also has the potential to improve interpretability of SNP-based disease prediction.

However, we would like to note that the ability to rank or prioritize a SNP in the context of an ML model, must not imply a direct causality to the disease. In fact understanding the functional relevance of individual variants, prioritized as part of a GWAS or with ML models, is a difficult process. A recent model using genotypic variation and gene expression patterns proposed that every complex phenotype is connected to a majority of genes expressed in the cell types relevant to the phenotype. Those act as part of cell-specific regulatory networks that affect the causal disease genes and act in highly non-linear ways (Liu et al. 2019). Thus, a variant may act in one of several cell types. The accessibility of DNA regions to regulatory proteins that interact with the DNA, such as transcription factors, is regulated by epigenetic mechanisms. In order to understand cell-type specific DNA accessibility different databases such as RegulomeDB and EpiRegio hold catalogues of DNA elements and their activity in different cell types e.g. (Baumgarten et al. 2020; Boyle et al. 2012). Also, variants can have indirect effects, for example by perturbing regulatory proteins such as transcription factors. Therefore, subsequent detailed biological experiments are necessary to reveal the role of individual variants with the phenotype.

Reproducibility

Being able to reproduce the results obtained by ML highly depends on the availability of appropriate source code and documentation. In the reviewed papers we often found that information on exact model settings that were tested by the authors was missing. For explainability, we can generally observe a similar reproducibility problem, too. Mostly when some explainability tool (like LIME) is applied, the exact settings of the tool are not reported; for example, Wang et al. (2018) just mention that they applied LIME. A notable exception is Badré et al. (2020) who report on algorithmic details and parameters and provide an in-depth comparison to previously identified relevant SNPs.

Generalizability

There is inherently a difficulty to compare the effectiveness of ML methods across studies due to the differences in the diseases under investigation. This already starts with the variety of preprocessing steps applied to the GWAS data sets: most papers follow an ad-hoc decision for filtering of certain samples or imputation of missing values. Generally no systematic approach is available that would justify the application of a certain preprocessing in a comparative manner; the only exception in our review is Oriol et al. (2019) who report on different filtering methods provided by the FRESA.CAD tool. In terms of the range of the discussed ML models, there is no clear favorite method that would propose itself as the optimal solution in all surveyed studies. Hence, the choice of the optimal ML model remains to be study- and disease-specific. Yet, the application of ensemble methods might improve predictions by combining the strengths of different ML models.

(Hyper-)parameter optimization

Finding the optimal settings for an ML model is one of the crucial factors when looking for the right choice of an ML method. Identifying the right (hyper-)parameters for such a model would, for example, enable faster convergence during model training, less overfitting and better accuracy of the outcomes. An extensive hyperparameter optimization for example by grid search – on the downside – incurs a heavy performance and time penalty; this is for example observed in Sun et al. (2020). Moreover, even when a hyperparameter search is used often a detailed description of the optimization strategy is missing in the papers. Still, a hyperparameter optimization (in particular, an automated topology optimization for neural networks) will likely improve the predictive accuracy and should be applied more systematically.

Inclusion of additional data

Going beyond the application of ML methods just on the genomic information, there might be a benefit when augmenting the GWAS data with additional inputs – for example, gene annotations, or clinical or environmental information. Some of the surveyed articles (López et al. 2018; Sun et al. 2020; Wang et al. 2018) consider extra information (like age, sex, smoking status, etc.). Moreover, one article (Xu et al. 2020) considers gene ontology (GO) enrichment to enable pathway analyses. Nevertheless, in

other use cases pathway analysis is more widely used; see for example (Pers et al. 2015) in the area of gene function analysis or the more general overview in White et al. (2019). Hence, data integration and multimodal/multi-omics data analysis are further topics for a more in-depth analysis when it comes to evaluating the benefits of ML for SNP-based disease prediction.

Conclusion

To sum up, as in many other areas, we see a potential of modern ML methods to widen the knowledge in the realm of GWAS-based disease prediction. In particular, the breadth of applied ML methods and the range of investigated diseases show an increased interest of the scientific community to make use of these new tools. Yet, the issues raised in our discussion also clearly show that reliability of these tools in practice is limited by several factors and automated predictions have to be questioned with due skill, care and diligence by experienced researchers and medical experts. In general, an appropriate user-friendly representation of the properties of each applied ML model should also be considered (Seifert et al. 2019) that summarizes the obtained analysis results in terms of accuracy, generalizability or bias and that can be intuitively understood by ML unexperienced researchers.

Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.
Research funding: This project is part of the "Center for Data Science and AI" funded by the Alfons und Gertrud Kassel-Stiftung. This work was supported by the DFG Cluster of Excellence Cardio Pulmonary Institute (CPI) [EXC 2026].
Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

Aguiar-Pulido, V., Seoane, J.A., Rabuñal, J.R., Dorado, J., Pazos, A., and Munteanu, C.R. (2010). Machine learning techniques for single nucleotide polymorphism–disease classification models in schizophrenia. Molecules 15: 4875–4889.

Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P., and Zondervan, K.T. (2010). Data quality control in genetic case-control association studies. Nat. Protoc. 5: 1564–1573.

Ani, A., van der Most, P.J., Snieder, H., Vaez, A., and Nolte, I.M. (2021). Gwasinspector: comprehensive quality control of genome-wide association study results. Bioinformatics 37: 129–130.

Badré, A., Zhang, L., Muchero, W., Reynolds, J.C., and Pan, C. (2020). Deep neural network improves the estimation of polygenic risk scores for breast cancer. J. Hum. Genet. 66: 1–11.

Baumgarten, N., Hecker, D., Karunanithi, S., Schmidt, F., List, M., and Schulz, M.H. (2020). EpiRegio: analysis and retrieval of regulatory elements linked to genes. Nucleic Acids Res. 48: W193–W199.

Bellenguez, C., Charbonnier, C., Grenier-Boley, B., Quenez, O., Le Guennec, K., Nicolas, G., Chauhan, G., Wallon, D., Rousseau, S., Richard, A.C., et al. (2017). Contribution to Alzheimer's disease risk of rare variants in trem2, sorl1, and abca7 in 1779 cases and 1273 controls. Neurobiol. Aging 59: 220–e1.

Bellot, P., de Los Campos, G., and Pérez-Enciso, M. (2018). Can deep learning improve genomic prediction of complex human traits? *Genetics* 210: 809–819.

Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., et al. (2012). Annotation of functional variation in personal genomes using regulomedb. Genome Res. 22: 1790–1797.

Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An expanded view of complex traits: from polygenic to omnigenic. Cell 169: 1177–1186.

Bracher-Smith, M., Crawford, K., and Escott-Price, V. (2020). Machine learning for genetic prediction of psychiatric disorders: a systematic review. Mol. Psychiatr. 26: 1–10.

Browning, B.L. and Browning, S.R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. 84: 210–223.

Buniello, A., MacArthur, J.A., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2018). The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 47: D1005–D1012.

Chen, I.Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., and Ghassemi, M. (2020). Ethical machine learning in healthcare. Annu. Rev. Biomed. Data Sci. 4, https://doi.org/10.1146/annurev-biodatasci-092820-114757.

Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M.M., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. J. R. Soc. Interface 15: 20170387.

Choi, S.W., Mak, T.S.-H., and O'Reilly, P.F. (2020). Tutorial: a guide to performing polygenic risk score analyses. Nat. Protoc. 15: 2759–2772.

Christophersen, I.E., Rienstra, M., Roselli, C., Yin, X., Geelhoed, B., Barnard, J., Lin, H., Arking, D.E., Smith, A.V., Albert, C.M., et al. (2017). Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. Nat. Genet. 49: 946–952.

Cox, T. (2001). Gaucher's disease—an exemplary monogenic disorder. QJM Int. J. Med. 94: 399–402.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., and Blaxter, M.L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12: 499–510.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. PLoS Genet. 9: e1003348.

Gaudillo, J., Rodriguez, J.J.R., Nazareno, A., Baltazar, L.R., Vilela, J., Bulalacao, R., Domingo, M., and Albia, J. (2019). Machine

learning approach to single nucleotide polymorphism-based asthma prediction. PloS One 14: e0225574.

Ghafouri-Fard, S., Taheri, M., Omrani, M.D., Daaee, A., Mohammad-Rahimi, H., and Kazazi, H. (2019). Application of single-nucleotide polymorphisms in the diagnosis of autism spectrum disorders: a preliminary study with artificial neural networks. J. Mol. Neurosci. 68: 515–521.

Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., et al. (2003). The international hapmap project. Nature 426: 789–796.

Gola, D., Erdmann, J., Müller-Myhsok, B., Schunkert, H., and König, I.R. (2020). Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status. Genet. Epidemiol. 44: 125–138.

Grillo, E., Rizzo, C.L., Bianciardi, L., Bizzarri, V., Baldassarri, M., Spiga, O., Furini, S., De Felice, C., Signorini, C., Leoncini, S., et al. (2013). Revealing the complexity of a monogenic disease: Rett syndrome exome sequencing. PloS One 8: e56599.

Halperin, E. and Stephan, D.A. (2009). Snp imputation in association studies. Nat. Biotechnol. 27: 349–351.

Ho, D.S.W., Schierding, W., Wake, M., Saffery, R., and O'Sullivan, J. (2019). Machine learning snp based prediction for precision medicine. Front. Genet. 10: 267.

Hopfner, F., Mueller, S.H., Szymczak, S., Junge, O., Tittmann, L., May, S., Lohmann, K., Grallert, H., Lieb, W., Strauch, K., et al. (2020). Rare variants in specific lysosomal genes are associated with Parkinson's disease. Mov. Disord. 35: 1245–1248.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44: 955–959.

Kastelein, J.J., Reeskamp, L.F., and Hovingh, G.K. (2020). Familial hypercholesterolemia: The most common monogenic disorder in humans. J. Am. Coll. Cardiol. 75: 2567–2569.

Kruppa, J., Ziegler, A., and König, I.R. (2012). Risk estimation and risk prediction using machine-learning methods. Hum. Genet. 131: 1639–1654.

Levine, M.E., Langfelder, P., and Horvath, S. (2017). A weighted snp correlation network method for estimating polygenic risk scores. In: *Biological networks and pathway analysis*. Springer, New York, U.S., pp. 277–290.

Lewis, C.M. and Vassos, E. (2020). Polygenic risk scores: from research tools to clinical instruments. Genome Med. 12: 1–11.

Liu, X., Li, Y.I., and Pritchard, J.K. (2019). Trans effects on gene expression can drive omnigenic inheritance. Cell 177: 1022–1034.e6.

López Ibáñez, B., Vinas, R., Torrent-Fontbona, F., and Fernández-Real Lemos, J.M. (2016). Handling missing phenotype data with random forests for diabetes risk prognosis. In: *1st ECAIWorkshop on artificial intelligence for diabetes*. European Conference on Artificial Intelligence (ECAI). Zenodo, The Hage, Netherlands, pp. 39–42.

López, B., Torrent-Fontbona, F., Viñas, R., and Fernández-Real, J.M. (2018). Single nucleotide polymorphism relevance learning with random forests for type 2 diabetes risk prediction. Artif. Intell. Med. 85: 43–49.

Lundberg, S.M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In: *Advances in neural information processing systems*. Curran Associates Inc., Red Hook, NY, USA, pp. 4765–4774.

Machiela, M.J. and Chanock, S.J. (2015). LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 31: 3555–3557.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 11: 499–511.

Mayo, O. (2008). A century of Hardy–Weinberg equilibrium. Twin Res. Hum. Genet. 11: 249–256.

Mieth, B., Kloft, M., Rodríguez, J.A., Sonnenburg, S., Vobruba, R., Morcillo-Suárez, C., Farré, X., Marigorta, U.M., Fehr, E., Dickhaus, T., et al. (2016). Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. Sci. Rep. 6: 36671.

Mieth, B., Rozier, A., Rodriguez, J.A., Hohne, M.M.-C., Gornitz, N., and Muller, K.R. (2020). DeepCOMBI: explainable artificial intelligence for the analysis and discovery in genome-wide association studies, *bioRxiv*.

Montanez, C.A.C., Fergus, P., Montaez, A.C., Hussain, A., Al-Jumeily, D., and Chalmers, C. (2018). Deep learning classification of polygenic obesity using genome wide association study snps. 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, New York, U.S. ,pp. 1–8.

Nicolae, D.L. (2006). Testing untyped alleles (tuna)—applications to genome-wide association studies. Genet. Epidemiol. 30: 718–727.

Okser, S., Pahikkala, T., Airola, A., Salakoski, T., Ripatti, S., and Aittokallio, T. (2014). Regularized machine learning in the genetic prediction of complex traits. PLoS Genet. 10: e1004754.

Oriol, J.D.V., Vallejo, E.E., Estrada, K., Peña, J.G.T., and Initiative, A.D.N. (2019). Benchmarking machine learning models for late-onset Alzheimer's disease prediction from genomic data. BMC Bioinf. 20: 1–17.

Orlenko, A. and Moore, J.H. (2021). A comparison of methods for interpreting random forest models of genetic association in the presence of non-additive interactions. BioData Min. 14: 1–17.

Paré, G., Mao, S., and Deng, W.Q. (2017). A machine-learning heuristic to improve gene score prediction of polygenic traits. Sci. Rep. 7: 1–11.

Pers, T.H., Karjalainen, J.M., Chan, Y., Westra, H.-J., Wood, A.R., Yang, J., Lui, J.C., Vedantam, S., Gustafsson, S., Esko, T., et al. (2015). Biological interpretation of genome-wide association studies using predicted gene functions. Nat. Commun. 6: 1–9.

Pirmoradi, S., Teshnehlab, M., Zarghami, N., and Sharifi, A. (2020). A self-organizing deep auto-encoder approach for classification of complex diseases using snp genomics data. Appl. Soft Comput. 97: 106718.

Privé, F., Arbel, J., and Vilhjálmsson, B.J. (2020). LDpred2: better, faster, stronger. Bioinformatics 36: 5424–5431.

Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. Association for Computing Machinery, New York, NY, USA, pp. 1135–1144.

Romagnoni, A., Jégou, S., Van Steen, K., Wainrib, G., and Hugot, J.-P. (2019). Comparative performances of machine learning methods for classifying crohn disease patients using genome-wide genotyping data. Sci. Rep. 9: 1–18.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1: 206–215.

Saeys, Y., Abeel, T., and Van de Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. In: *Joint European conference on machine learning and knowledge discovery in databases*. Springer, Heidelberg, Berlin, pp. 313–325.

Schote, A.B., Schiel, F., Schmitt, B., Winnikes, U., Frank, N., Gross, K., Croyé, M.-A., Tarragon, E., Bekhit, A., Bobbili, D.R., et al. (2020). Genome-wide linkage analysis of families with primary hyperhidrosis. PloS One 15: e0244565.

Seifert, C., Scherzinger, S., and Wiese, L. (2019). Towards generating consumer labels for machine learning models. In: *2019 IEEE first International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, Los Angeles, USA, pp. 173–179.

Shaik Mohammad, N., Sai Shruti, P., Bharathi, V., Krishna Prasad, C., Hussain, T., Alrokayan, S.A., Naik, U., and Radha Rama Devi, A. (2016). Clinical utility of folate pathway genetic polymorphisms in the diagnosis of autism spectrum disorders. Psychiatr. Genet. 26: 281–286.

Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the genetic architecture of 30 complex traits from summary association data. Am. J. Hum. Genet. 99: 139–153.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences, *arXiv preprint arXiv:1704.02685*.

Slatkin, M. (2008). Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. Nat. Rev. Genet. 9: 477–485.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12: 1–10.

Sun, T., Wei, Y., Chen, W., and Ding, Y. (2020). Genome-wide association study-based deep learning for survival prediction. Stat. Med. 39: 4605–4620.

Sun, Y.V. and Kardia, S.L. (2008). Imputing missing genotypic data of single-nucleotide polymorphisms using neural networks. Eur. J. Hum. Genet. 16: 487–495.

Torkamani, A., Wineinger, N.E., and Topol, E.J. (2018). The personal and clinical utility of polygenic risk scores. Nat. Rev. Genet. 19: 581–590.

Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am. J. Hum. Genet. 97: 576–592.

Wang, H.-Y., Chang, S.-C., Lin, W.-Y., Chen, C.-H., Chiang, S.-H., Huang, K.-Y., Chu, B.-Y., Lu, J.-J., and Lee, T.-Y. (2018). Machine learning-based method for obesity risk evaluation using single-nucleotide polymorphisms derived from next-generation sequencing. J. Comput. Biol. 25: 1347–1360.

Wei, Z., Wang, W., Bradfield, J., Li, J., Cardinale, C., Frackelton, E., Kim, C., Mentch, F., Van Steen, K., Visscher, P.M., et al. (2013). Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. Am. J. Hum. Genet. 92: 1008–1012.

White, M.J., Yaspan, B.L., Veatch, O.J., Goddard, P., Risse-Adams, O.S., and Contreras, M.G. (2019). Strategies for pathway analysis using GWAS and WGS data. Curr. Protoc. Hum. Genet. 100: e79.

Wray, N.R., Lin, T., Austin, J., McGrath, J.J., Hickie, I.B., Murray, G.K., and Visscher, P.M. (2021). From basic science to clinical application of polygenic risk scores: a primer. JAMA Psychiatry. 78: 101–109.

Xu, Y., Cao, L., Zhao, X., Yao, Y., Liu, Q., Zhang, B., Wang, Y., Mao, Y., Ma, Y., Ma, J.Z., et al. (2020). Prediction of smoking behavior from single nucleotide polymorphisms with machine learning approaches. Front. Psychiatr. 11: 416.

Yin, B., Balvert, M., van der Spek, R.A., Dutilh, B.E., Bohte, S., Veldink, J., and Schönhuth, A. (2019). Using the structure of genome data in the design of deep neural networks for predicting amyotrophic lateral sclerosis from genotype. Bioinformatics 35: i538–i547.

Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M., and Yang, T.-L. (2019). PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics 35: 1786–1788.