Review

Machine Learning to Advance Human Genome-Wide Association Studies

Rafaella Sigala , Vasiliki Lagou , Aleksey Shmeliov , Sara Atito , Samaneh Kouchaki , Muhammad Awais , Inga Prokopenko , Adam Mahdi , Ayse Demirkan *

Posted Date: 28 November 2023

doi: 10.20944/preprints202311.1819.v1

Keywords: genome-wide association; human genetics; machine learning

Review

Machine Learning to Advance Human Genome-Wide Association Studies

Rafaella E. Sigala [1], Vasiliki Lagou [1], Aleksey Shmeliov [1], Sara Atito [2,3], Samaneh Kouchaki [2,3], Muhammad Awais [2,3], Inga Prokopenko [1,2] Adam Mahdi [4] and Ayse Demirkan [1,2,*]

[1] Section of Statistical Multi-omics, Department of Clinical and Experimental Medicine; rafaella.sigala1@gmail.com, v.lagou@surrey.ac.uk, a.shmeliov@surrey.ac.uk, i.prokopenko@surrey.ac.uk
[2] Surrey Institute for People-Centred Artificial Intelligence, University of Surrey, UK; sara.atito@surrey.ac.uk, samaneh.kouchaki@surrey.ac.uk, muhammad.awais@surrey.ac.uk, i.prokopenko@surrey.ac.uk
[3] Centre for Vision, Speech and Signal Processing. University of Surrey, UK; sara.atito@surrey.ac.uk, samaneh.kouchaki@surrey.ac.uk, muhammad.awais@surrey.ac.uk
[4] Oxford Internet Institute, University of Oxford, UK; adam.mahdi@oii.ox.ac.uk
* Correspondence: a.demirkan@surrey.ac.uk

Abstract: Machine learning, including deep learning, reinforcement learning, and generative artificial intelligence are revolutionising every area of our lives when data is made available. With the help of these methods, we can decipher information from larger datasets while addressing the complex nature of biological systems in a more efficient way. Although machine learning methods have been introduced to human genetic epidemiological research as early as 2004, those were never used to their full capacity. In this review, we outline some of the main applications of machine learning to assigning human genetic loci to health outcomes. We summarise widely used methods and discuss their advantages and challenges. We also identify several tools, such as Combi, GenNet and GMSTool, specifically designed to integrate these methods for hypothesis-free analysis of genetic variation data. We elaborate on the additional value and limitations of these tools from a geneticist's perspective. Finally, we discuss the fast-moving field of foundation models and large multi-modal omics biobank initiatives.

Keywords: genome-wide association; human genetics; machine learning

1. Introduction

Genome-wide association study (GWAS) is a hypothesis-free statistical approach for assessing associations between genetic variants and phenotypes in a sample population [1]. Up to date, more than 60,000 genetic associations have been reported in more than 6,000 GWASs [2] with their summary statistics being publicly available in repositories, such as the GWAS Catalog [3], GWAS Atlas [2], and Roslin gene atlas [4], among others. Although the most popular area of GWAS application has been human genetics, this approach has also been successfully applied in genetic research for fungi [5], bacteria [6], plants [7], and animals, including wild and companion animals [8], as well as livestock populations [9], proving its widespread use across agricultural, veterinary and medical sciences. GWAS not only have been instrumental in discovering genetic variants as potential causal factors for human diseases but also encouraged the development of genotyping platforms and new statistical methods, as well as investment into huge DNA biobanks with petabytes of phenotype and genotype data from various ethnic groups [10].

The main ambition and outcomes anticipated when designing GWASs has been two-fold. First, is to understand the biological pathogenesis of human diseases or variation in quantitative traits, such as height or endophenotypes. Such knowledge can be used for development of diseases therapeutic options by blocking the downstream path of a culprit molecule, or by replacing the

doi:10.20944/preprints202311.1819.v1

missing molecules. Second, is to identify the individuals at risk of a certain disease, often by calculating polygenic risk scores (PRS). In the case of quantitative traits, the quantitative prediction is translated into a liability threshold, (e.g. > 30) to predict obesity.

1.1. The Road from GWAS Findings to Drug Discovery

GWAS for several diseases have led to the identification of a large number of associated variants in functionally plausible genes as in the case of *FTO* for obesity [11], *SLC30A8* for type 2 diabetes [12] and *APOE* for Alzheimer's disease [13]. In a more recent GWAS meta-analysis, missense variants in *GLP1R* locus with significant effects on random glucose were functionally followed up [14]. It was shown that some of these variants responded differently to GLP-1R agonist drugs, commonly used in managing diabetes, indicating the importance of tailored treatments based on genetic variability.

Several examples of therapeutically actionable GWAS variants, which map to genes modulated by currently used drugs for diseases, have been described [15]. A higher success rate in clinical trial progression is associated with causal evidence from human genetic studies prioritising genes encoding for approved drug targets or their interacting proteins [16]. The translation of GWAS signals into therapeutic targets requires the integration of multiple omics layers, as well as clinical knowledge of the pathophysiology of the disease. Open-source informatic solutions can assist in the identification and prioritisation of targets. For example, Open Targets Platform aggregates evidence scores from 22 different data sources capturing information from genetic associations, somatic mutations, known drugs, affected pathways, literature mining, differential expression and animal models [17]. Among these data sources, Open Targets Genetics aims to overcome the challenges of identifying the most likely causal variant and the actual causal genes at each GWAS locus for common, complex traits/diseases by integrating genetic and functional genomics features [18]. The application of complex statistical models on larger studies with broader phenotyping and better knowledge of disease pathophysiology offer opportunities not only for de novo drug development but also drug repurposing. However, most GWAS signals do not present large enough effect sizes to be translated into drug targets, apart from some notable exceptions, such as *APOE* for Alzheimer's disease [19].

1.2. GWAS Applications beyond Gene Discovery: Cumulative Genetic Profiles and Causal Relationships

GWAS findings demonstrated that most common non-communicable diseases show high polygenicity with each individual associated variant, accounting for a small proportion of phenotypic variance. It set the floor for a more efficient identification of individuals at high/low disease risk by calculating PRS, summing the weighted effect size of each associated variant **[20]**. PRS have been first introduced for highly polygenic mental disorders for which initial GWAS underachieved [21,22]. Later they were constructed for various diseases, such as coronary artery disease, hypercholesterolemia and T2D [23]. Although it improved disease prognosis compared to conventional risk factors, their value in clinical practice is still questionable highlighting the existing challenges [24]. For example, PRS studies can face ancestry biases with limited transferability across populations due to differences in risk allele frequencies, heritability, linkage disequilibrium and clinical heterogeneity [25]. The majority of existing PRS have been constructed based on variants identified in European populations. These Eurocentric PRS might be less predictive for other ethnic groups with substantially lower allele frequencies for these variants. Furthermore, certain modifiable factors, such as diet, alcohol consumption, smoking and physical activity, correlate with genetic ancestry influencing phenotypic variance and PRS accuracy [26]. Even within populations of the same ethnicity genetic differences are present leading to bias when PRS are trained and tested on different subpopulations [27]. Overall, more advanced methods are necessary to improve risk prediction models, making PRS implementation into clinical practice a reality.

Additionally, GWAS results fuelled development of novel approaches enabling discoveries of the complex relationships between human traits, exposomal and intrinsic factors. Among the most popular approaches is Mendelian Randomisation, a method powered by a plethora of GWAS data to estimate the causal effect of exposure on an outcome dissecting the causal relationships between phenotypes [28]. Mendelian Randomisation relies on effect estimates and standard errors obtained from individual SNPs in either single GWAS or through meta-analysis of GWAS. Finally, genetic correlation between two different phenotypes is not necessarily measured on the same individuals and can be calculated by using GWAS outputs [29].

GWAS approach relies on statistical modelling testing each SNP separately and the identified individual variants only account for a small proportion of the heritability of complex diseases and traits. This is partially due to lack of robust methodology for studying SNP-SNP interactions. Typically, GWAS analysis requires a large sample size for statistical power, which is achieved by meta-analysis of hundreds of GWAS conducted on distinct populations [30]. Unfortunately, GWAS findings often lack direct biological interpretation and post-GWAS methods are necessary for drug development.

2. Machine Learning Solutions for GWAS

Machine learning, a subfield of artificial intelligence, deals with the development of algorithms capable of learning from the data. Recently, the application and development of machine learning methods for genomics have undergone a rapid growth. It proved valuable for analysing complex, high-dimensional genomics data and extracting previously unknown information. Examples of machine learning applications in the wider omics field range from the identification of DNA sequences (splice sites [31], promoters [32], enhancers [33]), nucleosome positioning [34], taxonomic annotation [35], microbial enterotyping [36], sequence errors learning [37], microbial host body site and subject classification [38], viral escape prediction [39], protein 3D structure estimation [40], evolutionary population genetics inference [41] and genomic selection [42].

2.1. Machine Learning Methods Frequently Adapted for GWAS

PubMed and Google Scholar were searched for journal articles that included the keywords “machine learning” and “genome wide association study”. We focused on papers written in English and published from 1 January 2004 to 6 November 2023. An initial set of 147 articles was selected and then reviewed based on their title, keywords and abstract for inclusion. Papers that did not match inclusion criteria were eliminated, resulting in 109 articles. We then assessed full text of those papers, which were further categorised based on their context and relevance including research articles that applied machine learning algorithms to GWAS, PRS and review papers. We also included benchmarking research which used real data excluding the ones that used only synthetic data. From this set of articles duplicate papers were also deleted. This resulted in 79 relevant papers, of which 60 were research articles and 19 review articles. The methodology in each research article was analysed to identify the specific machine learning tools and their unique features. The most common methods included Support Vector Machines (SVMs), random forest and neural networks. We provide short background for these methods below.

Random forest [43] is an ensemble learning method commonly used in GWAS. In a random forest, several weak classifiers (e.g. tress) are constructed, each using a random subset of the training data and a random subset of the features. This randomness in data and feature selection is a key element of the method, which mitigates the risk of overfitting and helps ensure the model's generalisation to new, unseen data. Each tree in the forest independently makes predictions based on its specific subset of the data. When a new data point is presented to the model, it passes through each decision tree and their individual predictions are aggregated. In classification tasks, the final prediction is often determined by a majority vote among the trees, while in regression tasks, it is the average of the predictions. Random forests are particularly strong at handling high-dimensional genomic data commonly encountered in GWAS, providing insights into the importance of individual genetic features and interactions among them [44]. Random forests can also be used to perform feature importance rankings, helping researchers to identify key genetic variables contributing to complex traits, as discussed below.

SVMs [45] are a class of machine learning algorithms designed to classify data by identifying the optimal hyperplane that best separates different classes in a high-dimensional feature space. In the context of GWAS, SVMs map genetic data that is often represented as high-dimensional feature vectors in multi-dimensional space. The goal is to identify the hyperplane (decision boundary) that maximises the margin between different genetic variations associated with particular trait or disease. SVMs work by selecting support vectors, which are the data points closest to the decision boundary. These vectors play a key role in determining the orientation and position of the hyperplane. The choice of the optimal hyperplane is critical because it minimises the risk of overfitting and aims to generalise well to unseen data. SVMs can also handle non-linear relationships through kernel functions, transforming the input data into a higher-dimensional space, where a linear separator becomes feasible.

Neural networks [46] rapidly gained significance in GWAS, mainly due to their ability to uncover complex genetic patterns within high-dimensional genomic datasets. The basic building block of a neural network is the artificial neuron (also referred to as a node). Each neuron transforms input data through a weighted sum, which is followed by the application of an activation function. By connecting neurons in layers, neural networks can model increasingly abstract and complex relationships. In the context of GWAS, these networks are often designed as deep neural networks [47,48] with multiple hidden layers, to extract hierarchical features from genetic data. Neural networks are especially suited at capturing non-linear relationships among genetic variants [48]. During the training process they adjust their internal parameters to minimise prediction errors. This training process involves feeding the network with genetic data and adjusting its parameters until it can make more accurate predictions. Once the model has been trained, neural networks can be used for a variety of tasks, including classification, regression and feature selection.

2.2. Machine Learning Application Areas in GWAS

In this section we present the methods, benchmarking efforts, and specifically designed tools which integrate machine learning approaches working with high dimensional genetic data, the results of which are promising in identifying novel disease-associated susceptibility loci. These studies suggest that machine learning could be used instead of traditional statistical GWAS methods, potentially aiding in the better understanding of complex multifactorial genetic diseases and prediction of individuals at risk. Benchmarking efforts of using machine learning in field of GWAS are mainly focused on four methods: gradient boosting, random forest, SVM and neural networks. Here, we simplify the classification of applications by prioritising top GWAS results**, detecting epistasis among selected loci, prioritising variants for GWAS, predicting traits, identifying variant/loci and supporting PRS.**

Prioritization of top GWAS results Machine learning applications developed for post-GWAS prioritisation (up until 2020) were summarised by Nicholls et al. [49] who pointed out that 7 out of 19 post -GWAS prioritisation methods were ensemble methods, namely gradient boosting and random forest. One remarkable benchmarking effort in this field was done by Vitsios and Petrovski (2020) [50] and compared seven different machine learning methods to prioritise genes for amyotrophic lateral sclerosis, chronic kidney disease and epilepsy. They implemented a diverse pool of gene-annotation sources: generic resources (disease and/or tissue agnostic), resources filtered by tissue and disease-specific features. They also developed “mantis-ml” as an automated machine learning framework to enable learning from sets of gene-associated features. Random forest was reported as the top-performing classifier. Another benchmarking effort earlier was by Roshan, et al. (2011) who introduced random forest as a ranking method of causal variants for GWAS [51], once a GWAS is already performed. Their method helped to loosen the Bonferroni threshold, by 2 times the number of SNPs passing the threshold and showed that both methods improve the ranks of causal variants and associated regions.

An example of how neural networks can be used to prioritise disease-associated genetic variants, can be found in Liu et al. (2018) [52]. They developed DEOPEN, a model which integrates a deep convolutional neural network and a three-layer feed-forward neural network. This model can predict chromatin accessibility and consider interactions between sequence patterns. The authors also demonstrated how their framework can be used to evaluate genetic variants of interest, including functional variants. Their model outperformed Basset [53] and gkm-SVM [54] for classification of genome susceptibility in 50 random cell lines. Most importantly, DEOPEN can be used to identify known and potentially new transcription factor motifs. The authors applied their framework to a GWAS breast cancer dataset which identified 29 SNPs associated with this condition from 1,057 SNPs that co-occurred with them, through their involvement with a cancer-related transcription factor.

A random forest-based classifier, GCDPipe [55], uses gene-level results derived from GWAS analysis. It expands the list of potential disease gene candidates through the estimation of probability to influence disease risks. GCDPipe identifies gene expression profiles across cell types and tissues with the highest importance for the putative disease genes identification. Additionally, it prioritises drugs based on affinity to the putative disease genes using drug-gene interaction databases. Open Targets recently introduced new techniques for prioritising GWAS results [56]. Their "locus-to-gene" model derives features to prioritise likely causal genes at each GWAS locus, incorporating genetic and functional genomics features such as distance, molecular QTL colocalization, chromatin interaction and variant pathogenicity. The locus-to-gene method uses a machine learning model to determine the weights of each evidence source, referencing on a gold standard of previously identified causal genes and relying on fine-mapping and colocalisation data.

Another method, that uses epigenetic knowledge is DeepPerVar [57], was developed in two versions, based on two datasets: the DeepPerVar-H3K9ac (paired whole genome sequence and HEK9ac CHIP-seq data) and DeepPerVar-methy (paired whole genome sequence and DNA methylation data) to predict quantitative signals and methylation ratios, respectively. Overall, DeepPerVar was able to interpret and prioritise causal variants in a GWAS risk locus linked to Schizophrenia, quantify epigenetic signals and interpret the relationship of non-coding variants with a disease trait.

Epistasis detection among selected loci Random forest was initially suggested as an alternative to model genetic interactions in 2004 [44]. The rationale behind employing random forest is that in situations involving genuine interactions, SNPs exhibit modest individual effects but considerable interaction effects within a population. However, such effects are less likely to be detected at the genome-wide multiple testing thresholds used in GWAS screenings. Moreover, model-based screens that assess the interaction of each SNP with every other SNP in the dataset, aiming to pre-specify interacting SNPs, are impractical for datasets exceeding a thousand SNPs. Given that a typical GWAS dataset usually comprises more than 50,000 SNPs, such an approach becomes unfeasible.

Random forest analysis of interacting genetic models, up to 32 independent SNPs showed that random forest performed better than Fisher's exact test as a screening tool, when genetic heterogeneity as well as random noise is accounted for. In this study, the authors recommended that thousands of trees must be used in order to get stable estimates of the variable importance [44]. An advantage of random forest is that the investigator does not need to propose a model, making it well-suited for hypothesis-free screens such as GWAS or candidate gene studies. It also captures interactions and reflects them in variable importance scores. Drawbacks of the method include lack or concordance between variable importance and predictive index value [58] and high chance of detecting false, spurious associations when the study design is sub-optimal [59]. A recent report described by Leem et al. [60] suggested a three step approach allowing authors to define up to 5-locus interactions in real WTCCC datasets and in synthetic datasets without marginal effects. Also, there have been multiple attempts to find interacting genetic loci by other machine learning methods, such as decision trees (DF-SNPs) [61], Deep Mixed Model [62] and grammatical evolution optimised neural networks (GENN) [63].

Variant prioritisation One important area of machine learning for GWAS has been on prioritising loci to be included in GWAS. To this end, stand-alone but also combinatory tools have been developed for search space reduction. In 2015 Nguyen et al. [64] developed ts-RF which is a two-stage method. In this method, first a p-value assessment is performed to find a cut-off point that separates the genome-wide SNP data into relevant and irrelevant SNPs. The informative SNPs group is further divided into two sub-groups: highly informative and weak informative. Then these two groups are considered when sampling for building trees. So, the feature subspace is encouraged to contain highly informative SNPs when used to split a node at a tree, resulting in better performance. They applied ts-RF to real genome-wide datasets of Alzheimer's and Parkinson's disease and compared its performance of linear kernel SVM from LibSVM [65]. ts-RF performed better at prediction and was able to point 25 SNPs associated with Parkinson's disease that are located within gene regions studied by previous GWAS.

Silva et al. [66] showed that dimensionality reduction techniques based on random forest could effectively reduce dataset dimensions before conducting a cluster analysis of augmented GWAS data using a two-step machine learning approach. In the first step of dimension reduction, SNPs were ranked based on their relevance, and those with higher relevance underwent the second stage of analysis, which involved clustering. They tested the method on seroclearance GWAS in chronic hepatitis B while including the most significant SNPs in the clustering. The results included over 100 SNP sets which were associated with the phenotype of interest. SNPs were further detected and linked to HBsAg seroclearance with statistical significance based on Hamming distance-based association tests [67] in which a p-value for each predetermined causal SNP set was calculated. Knowing that statistically significant variants tend to cluster, the authors also investigated the functional relevance of SNPs found in the same SNP-set, as well as in individual SNPs followed by random forest and identified possible susceptible loci that could be otherwise ignored when only performing GWAS. The resulting SNP-sets from the cluster analyses were subsequently tested for trait-association and identified three susceptibility loci possibly associated with HBsAg seroclearance one of which was reported in the literature to be significantly associated with HBsAg seroclearance in patients who had received antiviral treatment.

Random forest was further combined with SVMs and k-nearest neighbour (kNN) clustering methods [68] by Gaudillo et al. and used for asthma genetic risk prediction. In their study they applied random forest to identify the SNPs with high implication to asthma. Following that, they trained kNN and SVM algorithms to classify the identified SNPs for their association to asthma. Recently, Díez Díaz et al. [69] proposed GASVeM that makes use of genetic algorithms together with SVMs to find out whether a certain biological pathway, assigned from a set of SNPs, can classify cases from controls. New frameworks using SVMs continue to be developed, while their performance is also shown to be heavily influenced by the heritability of the disease [70].

Recent research in Alzheimer's disease [71] used a hybrid feature selection approach based on association test, principal component analysis and the Boruta algorithm, to identify the most promising predictors. The selected features are then forwarded to a wide and deep neural network models to classify the Alzheimer's disease cases and healthy controls. In the first step, they conducted an association test to select the most signification SNPs influencing the disease, followed by a hybrid feature selection approach to reduce the number of features substantially. They subsequently used a selection process for neighbouring SNPs to generate a final set of SNPs. This set was then used to train wide and deep learning classification models for both cognitively normal individuals and those with Alzheimer's disease. Another method is DeepGWAS which uses a 14-layer deep neural network to enhance GWAS signals, using GWAS summary statistics, linkage disequilibrium information and brain related functional annotations. DeepGWAS was developed particularly for psychiatric diseases, starting with schizophrenia and outperformed XGBoost and logistic regression methods [72]. The range of applications using combinatory approaches continues to expand (Table 1).

2.3. Tools for SNP Discovery From Whole-Genome SNP Data

There is a growing number of efforts that use SVMs and neural networks narrow down the search space for GWAS. Additionally, there are tools designed to perform GWAS with no prior hypothesis or feature selection. Below we discuss algorithms and publicly available tools which have undergone internal benchmarking but warrant further testing in broader genetic epidemiological research (Table 2).

A method by Mieth et al. (2021), COMBI [73], employs a linear SVM which is trained and used as an indicator of importance and SNPs from each chromosome separately. This filtering step selects SNPs which contribute to phenotype classification with either high individual effects or effects in combination with the rest of SNPs, while removing results due to the correlation structure. At the application level, a phenotype vector and a genotype matrix which can be directly converted from a Plink [74] genotype object are generated. From these two objects, SVM weight vector is generated and used as importance measures.

Table 1. An overview of machine learning tools classified by application categories and machine learning approaches.

Application categories	Applications and tools	Machine learning approach
Prioritization of top GWAS results	• GCDPipe [55] • DeepPerVar *[57]* • Mantis-ml [50] • RF and SVMSnp [51] • ts-RF [64] • clustering +RF [66] • DeepGWAS [72] • Methods developed prior to2021 [49]	Clustering SVM Random Forrest Neural Network
Epistasis detection among pre-selected SNPs	• DF-SNPs [61] • random forest [44] • DEOPEN [52] • K-means [60]	Clustering Random Forrest Neural Network
Variant prioritization	• ts-RF [64] • clustering, random forest [66] • random forest, SVM, kNN [68] • GASVeM [69] • Wide and Deep Learning [71] • GMStool [75]	SVM Random Forrest Neural Network
Hypothesis-free GWAS	• COMBI [73] • DeepCOMBI [76] • Deep Mixed Model [62] • GenNet [77] • GWANN [78] • SNVformer [ref] • GMStool [75] • MACLEAPS [70] • iMEGES [79]	SVM Neural Network
Polygenic Risk Score	• NNP [80] • DNN [81] • RF-GRS [82]	Random Forrest Neural Network

In a second step, SNPs with the higher scores selected undergo a chi2 based hypothesis test performed together with Westfall-Young [83] type threshold calibration for each SNP, based on the permutation distribution of the re-sampled p-values. By this way, using a pre-selected list of SNPs and relaxed p-value threshold the proportion of true positives in the data is ultimately increased. In the simulated dataset COMBI overperformed other SVM based algorithms, including previously mentioned from Roshan et al. [51]. Following that they used data from the 2007 WTCCC phase 1, consisting of 14,000 cases of seven common diseases and 3,000 shared controls. When compared to the standard p-value thresholding approach, COMBI detected twelve additional SNP, ten of which have already been replicated in later GWAS or meta-analyses of bipolar disorder, coronary artery disease, Crohn's disease and for type 2 diabetes.

DeepCOMBI [76] The authors of COMBI subsequently developed a "deep" extension of COMBI, called DeepCOMBI [76]. This extension was designed to identify SNPs associated with a trait of interest, leveraging genotypic and phenotypic data from GWAS. The methodology includes the construction of deep neural networks for phenotype prediction of any genotype and SNPs selection according to a threshold, followed by layer-wise relevance propagation application on the SNPs and the selection of the most relevant variants. Lastly, a hypothesis test is performed for each variant. In addition, layer-wise relevance propagation yields the relevant scores for each variant and the permutation test can guarantee the selection of novel SNPs based on their p-values. In their report DeepCOMBI showed a better performance compared to other methods and identified a higher number of significant SNPs with the lowest error rate.

GenNet [77] Applying fully connected networks to millions of SNPs requires an ample amount of computational time and memory. To overcome these limitations, developers of GenNet provided a novel framework for predicting phenotype from genotype [77]. GenNet uses neural network, as well as prior biological knowledge, to create groups of nodes that are connected among the layers, reducing the sum of learnable parameters that a fully connected neural network would need. Biological knowledge may include information on gene annotation, local and global pathways, exon annotation, chromosome annotation, as well as cell and tissue type expression. In this model, neurons represent biological entities, and the weights signify the effects between neurons, resulting in a biologically interpretable network. This method allows human biological input, via a straightforward framework with help of two other pieces of software, HASE [84] and ANNOVAR [85] embedded in for generating necessary files. The major drawback of the method is that any researcher can perform differently layer annotation, making it difficult for standardisation.

GMStool [75] The tool was developed and tested on soybean but can be easily applied to human GWAS with no modification. Overall workflow consists of three phases: preparation, marker selection and final modelling. The preparation phase includes preparation of data which are genotype matrix, phenotype file and a GWAS summary statistic file as the training set. The marker selection phase applies the forward selection method of regression analysis and sequentially selects SNP markers that increase the correlation rate between observed and predicted phenotypes on the validation set. The ridge regression best linear unbiased prediction and bootstrap trees methods are provided as learning models. The final modelling phase performs prediction modelling using ridge regression, random forest, deep neural network and convolution neural network models, using either only one of them, or all. Unfortunately, the current construction of the GMStool requires the use individual level data in addition to GWAS summary statistics, limiting the application areas of the method.

Deep Mixed Model [62] GWAS on moderately or cryptically related individuals have customised methods to correcting for relatedness, usually either by genetic components or mixed models. To account for relatedness in genome-wide deep learning application Wang et al. [62] proposed Deep Mixed Model which consists of two components. The first component acts as a confounding factor correction by using a convolutional neural network, while the second component uses Long-short Term Memory for genetic variants selection. The results from Deep Mixed Model applied on Alzheimer disease genome-wide datasets of 1,017 individuals were not directly comparable to literature because most findings in GWAS Catalog are conducted through univariate testing methods. Nonetheless, six out of 20 SNPs selected by Deep Mixed Model were associated with Alzheimer's disease.

GWANN [78] Ashkenazy et al. (2022) [78] tried to exploit the ability of convolutional neural network in image recognition by developing and training a method for classification of variants associated to a trait of interest, using genomic data converted to image patterns. The model named GWANN, was trained using true positives and true negative data corresponding to trait association and finally makes prediction in a tested population. GWANN performance deteriorated when the simulated population did not accurately represent the tested data. For example, minor allele frequency less than 5% affected the pattern of SNP images, affecting the model's sensitivity. Therefore, parameters such as minor allele frequency, population structure, population size and sampling rate in the training populations need to be adjusted.

DeepWAS [86] Multivariate functional unit-wide association study (DeepWAS) was developed with the aim to only include SNPs that have been prioritised based on their risk potential. Genome-wide SNPs are first analysed for their functional roles and their association to specific cell lines and transcription factors using the deep learning model DeepSEA [87]. DeepWAS was able to identify and validate novel disease associated loci in multiple sclerosis, major depressive disorder and height that could not be identified in smaller cohort GWAS studies. It was also able to identify associations of SNPs within a functional unit relevant to a trait that typically missed in traditional GWAS. This methodology is ideal for any GWAS dataset if disease associated genetic conditions (cell-types effects, chromatin features) and its functional data are available. DeepWAS reduces the multiple testing burden of classical GWAS and makes regulatory information on a single SNP level readily available without requiring a second analysis step.

iMEGES [79] Integrated Mental-disorder GEnome Score (iMEGES), this method was developed as a deep learning tool for analysing whole genome/exome sequencing data, primarily for mental disorders [79]. In the first step, iMEGES prioritises variants based on non-coding and coding variants using tools EIGEN, CADD, DANN, GWAVA, FATHMM, known brain eQTLs from CommonMind and enhancer/promoters from PsychENCODE and Roadmap Epigenomics projects. In the second step genes are prioritised based on annotations for each variant from the first step of iMEGES.

Table 2 shows an overview of practical properties of these tools which are only internally benchmarked, requiring parallel assays for evaluating their analytical power over each other.

2.4. Applications Supporting PRS

While standard PRS is built upon linear models, below we summaries three methods which used nonlinear approaches to support disease prediction by GWAS based PRS. In principle, machine learning can be used to estimate PRS using classification scores achieved. In the breast cancer study by Badre et al. [81], the authors used deep neural network for breast cancer prediction and compared it to established statistical algorithms, via a combinatory design; firsts selecting SNPs by Plink and then building PRS either by deep neural network which they called neural network risk score or linear methods. Deep neural network outperformed best linear unbiased prediction methods [88].

Zhou et al. [80] developed deep neural network models for modelling Alzheimer's disease polygenic risk and compared them with the widely used weighted PRS and LASSO models. In their study they first selected the disease associated SNPs from a GWAS summary statistics and then predicted three different scenarios of training/validation splits. They considered the biological properties of variants, including gene and functional chromatin annotations to build seven-layer neural networks. Not the neural network risk score perform slightly better than weighted-PRS and LASSO, it also significantly associated with levels of the blood-based biomarkers of disease pathology.

Tree-based statistical learning methods were also tested for better PRS construction [82], showing that random forest and logic bagging outperform other tree based (logic regression, elastic net and RF-VIM) methods for predicting rheumatoid arthritis.

Table 2. Currently available tools that are designed for outcome prediction or gene/SNP discovery from genome-wide variation data.

Name	Method	Genotype matrix generation	Explainability/ Method for SNP relevance scores	Language
COMBI	Two-step method: 1) SVM training and selection of SNPs relevant for phenotype classification 2) Statistical testing	Not built-in. It requires a phenotype vector and a genotype matrix.	No/ SVM for SNP relevance scores	Matlab/Octave, R and Java
DeepCOMBI	Three-step method: 1) Training of a DNN for classification of subjects into their respective phenotypes 2) Calculation of SNP relevance scores (LRP) and SNP selection 3) Statistical testing	Not built-in. It requires a phenotype vector and a genotype matrix.	Yes/ relevance scores	Python
Deep Mixed Model	Two-component DL method: 1) One-dimensional CNN (confounding factor correction) 2) A LSTM model for selecting SNPs that contribute to residual phenotype in an epistatic manner	Not built-in. It requires genotype and phenotype matrices.	Not available	Python
DeepWAS	Integration method: 1) DL-based functional annotation of single GWAS SNPs for their regulatory effects on cell type-specific chromatin features (pre-trained DeepSEA network) 2) Association of regulatory SNPs with a disease/train into a multivariate setting (regularized regression models)	Not built-in. DeepSea requires vcf format.	Not available	R
GenNet	Use of NN with connections defined by prior biological knowledge to create groups of nodes across different layers to reduce the number of learnable parameters	Built-in	Built in as SNP, gene and pathway relevance scores based on relative weights	Python
GMStool	Three-step method: 1) Preparation of input files 2) Marker selection (RRB and/or BTS) 3) Prediction modelling (RRB, RF, DNN and/or CNN)	Not built-in. It requires genotype, phenotype, GWAS result and test list files.	Not available	R

GWANN	1) SNP data is converted into a learnable image (matrix) 2) The constructed images, each representing a SNP, are classified as either associated or not-associated with the trait using a CNN.	Not built-in. It requires a VCF file with genotype data and a csv file with phenotype data.	Not available	Python
IMEGES	The Annovar input/bed format file			

* List of specifically designed tools for gene discovery or outcome prediction using machine learning. MACLEAPS [70] which is an SVM based tool from 2013 was not included as the links to the were not functional. SVM: Support Vector Machine, DNN: Deep Neural Network, SNP: Single-nucleotide polymorphism, LRP: layer-wise relevance propagation, CNN: convolutional neural network, LSTM: Long-short Term Memory, DL: Deep-learning, VCF: Variant Call Format, NN: Neural Network. RRB: ridge regression best linear unbiased prediction, BTS: bootstrap trees, RF: Random Forest, SNV: Single-nucleotide variant. All software are publicly available; COMBI: part of the GWASpi toolbox 2.0 (https://bitbucket.org/gwas_combi/gwaspi/), DeepCombi: https://github.com/AlexandreRozier/DeepCombi, Deep Mixed Model: https://github.com/HaohanWang/DMM, DeepWAS: https://github.com/ cellmapslab/DeepWAS, GenNet: https://github.com/ArnovanHilten/GenNet, GMSTool: https://github.com/JaeYoonKim72/GMStool, GWAAN: https://github.com/hubner-lab/GWANN, IMEGES: https://github.com/WGLab/iMEGES.

3. Limitations and Criticism of Machine Learning

While machine learning offers plethora of new tools when combined with countless combinations of multi-modal omics data, there are multiple concerns for its use in GWAS.

Explainabilty As previously mentioned, the primary use of GWAS has been to understand the biological factors underlying human traits and diseases, at the single nucleotide resolution. To this end, machine learning methods only focused on prediction, which cannot be used to identify molecular drug targets by default. However, the same methods can be very powerful in predicting and classifying diseases. Recently there has been considerable research dedicated to developing interpretability frameworks toward hypothesis free genome scans [77]. Applications such as GenNet and iMEGES are promising tools as their methods largely benefit from functional annotations across the human genome.

Comparability So-called interpretable machine learning applications provide feature importance score reflecting the importance or relevance of variables in the prediction model [77]. However, they can neither be translated into effect estimates nor p-values which constitute the summary statistics tables in large repositories. Thus, there is limited comparability between data accumulated in conventional GWAS repositories and those generated by machine learning.

Standardisation and data accumulation GWAS methodology has been developed via rigorous consortia work for almost two decades. Standards related to study design, sample size, replication, population stratification, and meta-analyses have been integrated into practical workflows. Currently there is lack of standardisation for best practices in applying machine learning to human genetics. Since the field is still in its early stages, it requires guidelines to define the best approaches.

Data imbalance Commonly overlooked problem in machine learning is addressing data imbalance. Machine learning methods often require same number of cases and controls [77]. As we discussed above, most biobanks are designed in population based settings and in most study designs case numbers are lower than controls. This brings a discrepancy in study design and add more complication on comparison to conventional GWAS results. Nevertheless, in terms of study power, same limitations of GWAS also exist in machine learning, study power depends on sample size [89] as well as the heritability of the disease [70].

4. Future Prospects

Here we emphasize the two important drivers of the field, growing number of biobanks and fast developing new AI methods.

4.1. Multimodal Omics Databases

One of the most important applications of machine learning in the medical field is the development of multimodal AI models necessary for the integration of omics data across different modalities from biobanks and initiatives [90], These studies are designed to include hundreds of thousands of individuals with in-depth genetic and health information that are regularly enriched with new omics layers and follow-up measurements. The data generated are high-dimensional and multi-layered as they incorporate a massive collection of "omics" (genomic, transcriptomic, proteomic, metabolic or microbiome) along with electronic health records and study specific other measurements. The best known longitudinal population-based biobanks include the UK Biobank [91], the China Kadoorie Biobank [92], the Estonian Biobank [93] and the Lifelines Biobank [94]. The use of this data through implementation of AI methods has allowed high-throughput analysis and has led to new discoveries in the medical field [71] and shown to improve prediction in comparison to unimodal approach [90].

4.2. Opportunities of Large Language models and Foundation Models

Genomic sequences are vast repositories of complex biological data containing distant semantic relationships which may not be fully captured by traditional AI methods although ideal for foundation models. In traditional AI, most of the compute resources were spent on training models for specific tasks. To train such models, we need large amounts of labelled data (e.g. outcomes) which is often expensive, especially in the healthcare field. On the other hand, foundation models are large deep neural networks pre-trained on diverse data from a range of problems using *self-supervised learning* [95,96], which does not require expensive human labels. Once these foundation models are pre-trained they can be finetuned for downstream tasks which are specific to a particular problem using relatively little labelled data or in some cases no labelled data. Therefore, foundation models have been transforming the AI landscape in natural language processing, computer vision and multimodal analysis including the field of omics. Foundation models started to emerge in natural language processing around 2018 and in 2023, multimodal foundation models appeared in healthcare, radiology [97].

The self-supervised learning principles which are behind these foundation models are usually based on simple principles. Typically, words are converted to vector representation using simple neural network embeddings. Then, the job of the deep neural network in self-supervised learning is to recover words masked randomly from the context. For example, BERT [95] masks 15% of the words randomly and recover these words at the output. In addition, BERT predicts whether two sentences are next to each other or not. On the other hand, GPT like model simply predicts the next word in the sentence. If the deep neural network is unable to predict the right words, their weights are updated using back propagation algorithms [98]. When applied in genomics, DNA or RNA string can be considered as text document with characters in DNA or words in proteins enabling foundation models to capture complex local and distant semantic relations.

The complexities of genetic information pose unique challenges, such as high dimensionality and the need for significant computational power, which have so far hindered the widespread adoption of foundation models in this area with relatively few publications applying basic concepts of foundation models to genomic data [99] [100,101]. For example, Santiesteban et al [101] showed that foundation models combining transcriptomics and histopathology data through self-supervised learning significantly improves survival prediction. As the volume of omics data continues to grow in biobanks and computational capabilities advance, the full spectrum of foundation models' capabilities is likely to bring a new era of scientific discovery and innovation in biomedicine.

5. Conclusions

Broad range of applications under the machine learning umbrella offer solution for some of the problems in GWAS, however, application of these methods carelessly may also mitigate their benefits. We believe the benefits of this new interdisciplinary area will increase by building a common language and aims and through collaborative efforts.

Conflict of interest The authors declare no conflict of interest.

References

1. Visscher, P.M. et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet 101, 5-22 (2017).
2. Watanabe, K. et al. A global overview of pleiotropy and genetic architecture in complex traits. Nat Genet 51, 1339-1348 (2019).
3. GWAS catalogue.
4. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. Nat Genet 50, 1593-1599 (2018).
5. Frontini, M. et al. Genome-wide association of rice response to blast fungus identifies loci for robust resistance under high nitrogen. BMC Plant Biol 21, 99 (2021).
6. Young, B.C. et al. Panton-Valentine leucocidin is the key determinant of Staphylococcus aureus pyomyositis in a bacterial GWAS. Elife 8(2019).
7. Tibbs Cortes, L., Zhang, Z. & Yu, J. Status and prospects of genome-wide association studies in plants. Plant Genome 14, e20077 (2021).
8. Plassais, J. et al. Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. Nat Commun 10, 1489 (2019).
9. Wang, K. et al. The Chicken Pan-Genome Reveals Gene Content Variation and a Promoter Region Deletion in IGF2BP1 Affecting Body Size. Mol Biol Evol 38, 5066-5081 (2021).
10. Ramirez, A.H. et al. The All of Us Research Program: Data quality, utility, and diversity. Patterns (N Y) 3, 100570 (2022).
11. Claussnitzer, M. et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N Engl J Med 373, 895-907 (2015).
12. Ng, M.C. et al. Implication of genetic variants near TCF7L2, SLC30A8, HHEX, CDKAL1, CDKN2A/B, IGF2BP2, and FTO in type 2 diabetes and obesity in 6,719 Asians. Diabetes 57, 2226-33 (2008).
13. Lambert, J.C. et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet 45, 1452-8 (2013).
14. Lagou, V. et al. GWAS of random glucose in 476,326 individuals provide insights into diabetes pathophysiology, complications and treatment stratification. Nat Genet 55, 1448-1461 (2023).
15. Reay, W.R. & Cairns, M.J. Advancing the use of genome-wide association studies for drug repurposing. Nat Rev Genet 22, 658-671 (2021).
16. Ochoa, D. et al. Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. Nat Rev Drug Discov 21, 551 (2022).
17. Ochoa, D. et al. The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. Nucleic Acids Res 51, D1353-D1359 (2023).
18. Ghoussaini, M. et al. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. Nucleic Acids Res 49, D1311-D1320 (2021).
19. Genin, E. et al. APOE and Alzheimer disease: a major gene with semi-dominant inheritance. Mol Psychiatry 16, 903-7 (2011).
20. Ni, G. et al. A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts. Biol Psychiatry 90, 611-620 (2021).
21. International Schizophrenia, C. et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460, 748-52 (2009).
22. Demirkan, A. et al. Genetic risk profiles for depression and anxiety in adult and elderly cohorts. Mol Psychiatry 16, 773-83 (2011).
23. Lewis, C.M. & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. Genome Med 12, 44 (2020).

24. O'Sullivan, J.W. et al. Polygenic Risk Scores for Cardiovascular Disease: A Scientific Statement From the American Heart Association. Circulation 146, e93-e118 (2022).
25. Martin, A.R. et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet 100, 635-649 (2017).
26. Kachuri, L. et al. Principles and methods for transferring polygenic risk scores across global populations. Nat Rev Genet (2023).
27. Gola, D. et al. Population Bias in Polygenic Risk Prediction Models for Coronary Artery Disease. Circ Genom Precis Med 13, e002932 (2020).
28. Richmond, R.C. & Davey Smith, G. Mendelian Randomization: Concepts and Scope. Cold Spring Harb Perspect Med 12(2022).
29. van Rheenen, W., Peyrot, W.J., Schork, A.J., Lee, S.H. & Wray, N.R. Genetic correlations of polygenic disease traits: from theory to practice. Nat Rev Genet 20, 567-581 (2019).
30. Bergen, S.E. & Petryshen, T.L. Genome-wide association studies of schizophrenia: does bigger lead to better results? Curr Opin Psychiatry 25, 76-82 (2012).
31. Degroeve, S., De Baets, B., Van de Peer, Y. & Rouze, P. Feature subset selection for splice site prediction. Bioinformatics 18 Suppl 2, S75-83 (2002).
32. Bucher, P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J Mol Biol 212, 563-78 (1990).
33. Heintzman, N.D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39, 311-8 (2007).
34. Segal, E. et al. A genomic code for nucleosome positioning. Nature 442, 772-8 (2006).
35. Mathieu, A., Leclercq, M., Sanabria, M., Perin, O. & Droit, A. Machine Learning and Deep Learning Applications in Metagenomic Taxonomy and Functional Annotation. Front Microbiol 13, 811495 (2022).
36. Costea, P.I. et al. Enterotypes in the landscape of gut microbial community composition. Nat Microbiol 3, 8-16 (2018).
37. Callahan, B.J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 13, 581-3 (2016).
38. Statnikov, A. et al. A comprehensive evaluation of multicategory classification methods for microbiomic data. Microbiome 1, 11 (2013).
39. Hie, B., Zhong, E.D., Berger, B. & Bryson, B. Learning the language of viral evolution and escape. Science 371, 284-288 (2021).
40. Ramakrishnan, G. et al. Understanding structure-guided variant effect predictions using 3D convolutional neural networks. Front Mol Biosci 10, 1204157 (2023).
41. Huang, X., Rymbekova, A., Dolgova, O., Lao, O. & Kuhlwilm, M. Harnessing deep learning for population genetic inference. Nat Rev Genet (2023).
42. Saba Moeinizade, G.H., Lizhi Wang. A REINFORCEMENT LEARNING APPROACH TO
43. RESOURCE ALLOCATION IN GENOMIC SELECTION. (2021).
44. Chen, X. & Ishwaran, H. Random forests for genomic data analysis. Genomics 99, 323-9 (2012).
45. Lunetta, K.L., Hayward, L.B., Segal, J. & Van Eerdewegh, P. Screening large-scale association study data: exploiting interactions using random forests. BMC Genet 5, 32 (2004).
46. Cortes, C. & Vapnik, V. Support-vector networks. Machine Learning 20, 273-297 (1995).
47. Gurney, K. An Introduction to Neural Networks, (CRC Press, 1997).
48. Alzubaidi, L. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021).
49. Montesinos-Lopez, O.A. et al. A review of deep learning applications for genomic selection. BMC Genomics 22, 19 (2021).
50. Nicholls, H.L. et al. Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci. Front Genet 11, 350 (2020).
51. Vitsios, D. & Petrovski, S. Mantis-ml: Disease-Agnostic Gene Prioritization from High-Throughput Genomic Screens by Stochastic Semi-supervised Learning. Am J Hum Genet 106, 659-678 (2020).
52. Roshan, U., Chikkagoudar, S., Wei, Z., Wang, K. & Hakonarson, H. Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. Nucleic Acids Res 39, e62 (2011).

53. Liu, Q., Xia, F., Yin, Q. & Jiang, R. Chromatin accessibility prediction via a hybrid deep convolutional neural network. Bioinformatics 34, 732-738 (2018).
54. Kelley, D.R., Snoek, J. & Rinn, J.L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res 26, 990-9 (2016).
55. Lee, D. et al. A method to predict the impact of regulatory variants from DNA sequence. Nat Genet 47, 955-61 (2015).
56. Pinakhina, D., Loboda, A., Sergushichev, A. & Artomov, M. Gene, cell type, and drug prioritization analysis suggest genetic basis for the utility of diuretics in treating Alzheimer disease. HGG Adv 4, 100203 (2023).
57. Mountjoy, E. et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Nat Genet 53, 1527-1533 (2021).
58. Wang, Y. & Chen, L. DeepPerVar: a multi-modal deep learning framework for functional interpretation of genetic variants in personal genome. Bioinformatics 38, 5340-5351 (2022).
59. Bureau, A. et al. Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol 28, 171-82 (2005).
60. Garcia-Magarinos, M., Lopez-de-Ullibarri, I., Cao, R. & Salas, A. Evaluating the ability of tree-based methods and logistic regression for the detection of SNP-SNP interaction. Ann Hum Genet 73, 360-9 (2009).
61. Leem, S., Jeong, H.H., Lee, J., Wee, K. & Sohn, K.A. Fast detection of high-order epistatic interactions in genome-wide association studies using information theoretic measure. Comput Biol Chem 50, 19-28 (2014).
62. Xie, Q. et al. Decision forest analysis of 61 single nucleotide polymorphisms in a case-control study of esophageal cancer; a novel method. BMC Bioinformatics 6 Suppl 2, S4 (2005).
63. Wang, H., Yue, T., Yang, J., Wu, W. & Xing, E.P. Deep mixed model for marginal epistasis detection and population stratification correction in genome-wide association studies. BMC Bioinformatics 20, 656 (2019).
64. Motsinger-Reif, A.A., Dudek, S.M., Hahn, L.W. & Ritchie, M.D. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. Genet Epidemiol 32, 325-40 (2008).
65. Nguyen, T.T., Huang, J., Wu, Q., Nguyen, T. & Li, M. Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. BMC Genomics 16 Suppl 2, S5 (2015).
66. Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. . ACM Transactions on Intelligent Systems and Technology 2:27:1--27:27(2011).
67. Silva, P.P. et al. A machine learning-based SNP-set analysis approach for identifying disease-associated susceptibility loci. Sci Rep 12, 15817 (2022).
68. Wang, C., Kao, W.H. & Hsiao, C.K. Using Hamming Distance as Information for SNP-Sets Clustering and Testing in Disease Association Studies. PLoS One 10, e0135918 (2015).
69. Gaudillo, J. et al. Machine learning approach to single nucleotide polymorphism-based asthma prediction. PLoS One 14, e0225574 (2019).
70. Díez Díaz, F. et al. GASVeM: A New Machine Learning Methodology for Multi-SNP Analysis of GWAS Data Based on Genetic Algorithms and Support Vector Machines. Mathematics 9, 654 (2021).
71. Mittag, F. et al. Use of support vector machines for disease risk prediction in genome-wide association studies: concerns and opportunities. Hum Mutat 33, 1708-18 (2012).
72. Alatrany, A.S., Khan, W., Hussain, A., Al-Jumeily, D. & Alzheimer's Disease Neuroimaging, I. Wide and deep learning based approaches for classification of Alzheimer's disease using genome-wide association studies. PLoS One 18, e0283712 (2023).
73. Li, Y. et al. DeepGWAS: Enhance GWAS Signals for Neuropsychiatric Disorders via Deep Neural Network. Res Sq (2023).
74. Mieth, B. et al. Combining Multiple Hypothesis Testing with Machine Learning Increases the Statistical Power of Genome-wide Association Studies. Sci Rep 6, 36671 (2016).
75. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559-75 (2007).
76. Jeong, S., Kim, J.Y. & Kim, N. GMStool: GWAS-based marker selection tool for genomic prediction from genomic data. Sci Rep 10, 19653 (2020).
77. Mieth, B. et al. DeepCOMBI: explainable artificial intelligence for the analysis and discovery in genome-wide association studies. NAR Genom Bioinform 3, lqab065 (2021).
78. van Hilten, A. et al. GenNet framework: interpretable deep learning for predicting phenotypes from genetic data. Commun Biol 4, 1094 (2021).

79. Nimrod Ashkenazy, M.F., Ofer M. Shir, Sariel Hübner. GWANN: Implementing deep learning in genome wide association studies. BioRxiv (2022).
80. Khan, A., Liu, Q. & Wang, K. iMEGES: integrated mental-disorder GEnome score by deep neural network for prioritizing the susceptibility genes for mental disorders in personal genomes. BMC Bioinformatics 19, 501 (2018).
81. Zhou, X. et al. Deep learning-based polygenic risk analysis for Alzheimer's disease prediction. Commun Med (Lond) 3, 49 (2023).
82. Badre, A., Zhang, L., Muchero, W., Reynolds, J.C. & Pan, C. Deep neural network improves the estimation of polygenic risk scores for breast cancer. J Hum Genet 66, 359-369 (2021).
83. Lau, M., Wigmann, C., Kress, S., Schikowski, T. & Schwender, H. Evaluation of tree-based statistical learning methods for constructing genetic risk scores. BMC Bioinformatics 23, 97 (2022).
84. Peter H. Westfall, S.S.Y. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment, (1993).
85. Roshchupkin, G.V. et al. HASE: Framework for efficient high-dimensional association analyses. Sci Rep 6, 36076 (2016).
86. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164 (2010).
87. Arloth, J. et al. DeepWAS: Multivariate genotype-phenotype associations by directly integrating regulatory information using deep learning. PLoS Comput Biol 16, e1007616 (2020).
88. Zhou, J. & Troyanskaya, O.G. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 12, 931-4 (2015).
89. Maier, R. et al. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. Am J Hum Genet 96, 283-94 (2015).
90. Wei, Z. et al. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. Am J Hum Genet 92, 1008-12 (2013).
91. Laurie Prélot, H.D., Mila D. Anasanti, Zhanna Balkhiyarova, Matthias Wielscher, Loic Yengo, Beverley Balkau, Ronan Roussel, Sylvain Sebert, Mika Ala-Korpela, Philippe Froguel, Marjo-Riitta Jarvelin, Marika Kaakinen, Inga Prokopenko. Machine Learning in Multi-Omics Data to Assess Longitudinal Predictors of Glycaemic Health. BioRxiv (2018).
92. Sudlow, C. et al. 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, e1001779 (2015).
93. Chen, Z. et al. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. Int J Epidemiol 40, 1652-66 (2011).
94. Leitsalu, L. et al. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int J Epidemiol 44, 1137-47 (2015).
95. Scholtens, S. et al. Cohort Profile: LifeLines, a three-generation cohort study and biobank. Int J Epidemiol 44, 1172-80 (2015).
96. Jacob Devlin, M.-W.C., Kenton Lee, Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv (2018).
97. Atito, S., Awais, M. & Kittler, J. Sit: Self-supervised vision transformer. ArXiv.
98. Moor, M. et al. Foundation models for generalist medical artificial intelligence. Nature 616, 259-265 (2023).
99. Rumelhart, D.E., Hinton, G.E. & Williams, R.J. Learning representations by back-propagating errors. Nature 323, 533-536 (1986).
100. Kieran Elmes, D.B.-P., Neşet Özkan Tan, Trung Bao Nguyen, Nicholas Sumpter, Megan Leask, Michael Witbrock, Alex Gavryushkin. SNVformer: An Attention-based Deep Neural Network for GWAS Data. (2022).
101. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R.V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics 37, 2112-2120 (2021).
102. Santiesteban, S., Awais, M., Song, Y. & Kittler, J. Multimodal Self-Supervised Learning for Pan-Cancer Survival Prediction using Histology-Genomic Data." Open review CVPR (2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.