genes

MDPI

Review

Statistical Learning Methods Applicable to Genome-Wide Association Studies on Unbalanced Case-Control Disease Data

Xiaotian Dai, Guifang Fu *, Shaofei Zhao and Yifei Zeng

Department of Mathematical Sciences, SUNY Binghamton University, Vestal, NY 13850, USA;
xdai@binghamton.edu (X.D.); szhao@math.binghamton.edu (S.Z.); yzeng@math.binghamton.edu (Y.Z.)
* Correspondence: gfu@binghamton.edu

Abstract: Despite the fact that imbalance between case and control groups is prevalent in genome-wide association studies (GWAS), it is often overlooked. This imbalance is getting more significant and urgent as the rapid growth of biobanks and electronic health records have enabled the collection of thousands of phenotypes from large cohorts, in particular for diseases with low prevalence. The unbalanced binary traits pose serious challenges to traditional statistical methods in terms of both genomic selection and disease prediction. For example, the well-established linear mixed models (LMM) yield inflated type I error rates in the presence of unbalanced case-control ratios. In this article, we review multiple statistical approaches that have been developed to overcome the inaccuracy caused by the unbalanced case-control ratio, with the advantages and limitations of each approach commented. In addition, we also explore the potential for applying several powerful and popular state-of-the-art machine-learning approaches, which have not been applied to the GWAS field yet. This review paves the way for better analysis and understanding of the unbalanced case-control disease data in GWAS.

Keywords: disease; GWAS; unbalanced case-control; genomic selection; genomic prediction

check for updates

Citation: Dai, X.; Fu, G.; Zhao, S.; Zeng, Y. Statistical Learning Methods Applicable to Genome-Wide Association Studies on Unbalanced Case-Control Disease Data. *Genes* **2021**, *12*, 736. https://doi.org/10.3390/genes12050736

Academic Editor: Tim Beck

Received: 23 March 2021
Accepted: 10 May 2021
Published: 13 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Over the past ten years, genome-wide association studies (GWAS) have shown great potential in investigating the biological and genetic etiology of disease, with the aims of providing better understanding, prevention, and treatment of diseases. As the cost of genotyping decreases, GWAS research moves to a new level, as phenome- wide association studies (PheWAS) enable thousands of phenotypes constructed from electronic health records (EHRs) and biobanks involving tens of millions of variants for hundreds of thousands of participants in large cohorts [1–4]. The PheWAS and large biobanks create new opportunities for detecting more scientific findings from the GWAS data [5,6]. However, most binary phenotypes have substantially fewer cases than controls [7].

Here we first give a few examples, from slight, moderate, to extreme imbalance ratios: The Wellcome Trust Case Control Consortium (WTCCC) provides a series of GWAS datasets that include 2000 case samples from each of seven common diseases [8] (e.g., type 1 diabetes [9], type 2 diabetes [10], coronary heart disease [10,11], bipolar disorder [12], rheumatoid arthritis [13,14], and so on). Their shared control has 3000 heathy samples (case-control ratio of 0.66). Dai et al. [15] analyzed a polycystic ovary syndrome (PCOS) affection status dataset consisting of 1043 cases and 3056 controls (with the case-control ratio of 0.34) and 731,442 SNPs.

The UK Biobank [1,16] is a very large study with over 400,000 participants from white British participants with European ancestry, which collected >1400 case-control disease phenotypes: colorectal cancer, prostate cancer, lung cancer, and Alzheimer's, disease to name a few. Most of their binary phenotypes have a case-control ratio lower than 1:100 [1,7,16]. The Michigan Genomics Initiative (MGI) is taking efforts to create a biorepository of genomic data and has involved >25,000 samples with >500,000 SNPs. It has already been noticed that multiple phenotypes have an extremely small number of cases, as small as 20 cases [3,4].

The imbalance of binary phenotypes poses serious challenges to traditional statistical methods, in terms of both genomic selection [2–4,7,17–26] and phenotypic prediction such as the prediction on disease status [24,27,28]. In this article, we review several statistical methods that have been applied in the unbalanced case-control GWAS data. We commented on the advantages and disadvantages of each method; in addition, we also introduced some state-of-the-art machine-learning methods that have great potential to be applied to solve the imbalance issues in the GWAS field. These methods have received a lot of attention in many other fields but they have not been applied to analyze GWAS data yet. Sun et al. [29] also wrote an overview of statistical learning methods that are suitable for classification of unbalanced data; however, two major differences are: we mainly focused on the GWAS application that was out of the scope of Sun et al.; the statistical learning approaches introduced in this article represent the newer developments than those in [29].

This review article is organized as follows. We first discuss why the imbalance causes an issue from a statistical aspect. Then we introduce the generalized linear mixed model association test in Section 3, and the Scalable and Accurate Implementation of Generalized mixed model in Section 4. They are both single-SNP methods related to logistic mixed model. In Section 5, we comment the Bayesian multiple Logistic Regression method [24] as a joint variable selection method for binary traits. In the remaining sections, we comment on multiple state-of-the-art machine-learning algorithms, the Support Vector Machine [30] in Section 6, AdaBoost in Section 7, and neural network in Section 8. In Section 9, we introduce the permutation-based significance test skills to facilitate the usage of machine-learning approaches into the GWAS field. Finally, we discuss other challenges of GWAS data analysis and summarize the advantages and limitations of these approaches in Section 10.

2. Why Does the Imbalance Cause an Issue from a Statistic Aspect?

The problems caused by an imbalance in case-control data can be summarized from four aspects [31]: (1) erroneously assuming that the accuracy metric (e.g., error rate) is appropriate; (2) erroneously assuming that the distribution of test statistic is the same between the case and control group; (3) erroneously assuming that the minority group has an adequate sample size; (4) erroneously assuming that the underlying asymptotic assumptions are still valid.

Firstly, the error rate has been widely used as an accuracy metric in the classification literature. However, it averages all observations without treating them differently, under the assumption that the samples in minority class have equal importance as those in majority class. As a result, it always favors the majority class [32]. For example, if the data contain 99% of the control (negative) and 1% of the case (positive), then predicting everything as negative will give us 99% accuracy. Statistically speaking, the classifier works correctly if the accuracy metric were appropriate. Drummond et al. [33] showed that it is usually very hard to outperform this simple classifier if the data are unbalanced.

However, in the GWAS field with binary traits, people care more about the cases (the disease status) than controls (healthy). Therefore, it is more serious to misclassify a case compared to misclassifying a control. To overcome the problem of error rate, weighted loss function or AUC (the area under the ROC curve) has been used [34]. The ROC curve is a plot of true positive rate vs. false positive rate under various thresholds, and usually a higher AUC stands for a better classifier. Liang et al. [35] showed that AUC is statistically consistent and better than the error rate under many scenarios. Hanley et al. [36] showed that AUC is actually equivalent to the Mann–Whitney statistic. Since the AUC is more related with rank statistic, it is invariant to the prior probabilities, which makes it a desired accuracy metric in evaluating the unbalanced data.

Secondly, ideally speaking, the training distribution should be the same as the test distribution, but in the unbalanced case, it is relatively more likely to get different distributions between the training and test data when randomly splitting an unbalanced data set.

For example, it is possible that the training data are highly unbalanced but that the test data are balanced; or in some extreme cases we may end up with no samples from the case group in training or test data. Under this circumstance, even when equipped with correct accuracy metric a classifier will not work well. To tackle this problem, sampling strategies such as over-sampling/under-sampling have been applied to make the data more balanced. However, over-sampling may increase the model and computation complexity, which is a burden for the GWAS data when millions of genetic variants are involved. Meanwhile, under-sampling may yield less information than we should have. See Zhou (2013) [37] for a detailed comparison on the performance of multiple sampling methods.

Thirdly, contamination may destroy the sample, and in particular for rare diseases it is always difficult to get enough case samples. When the number of minority classes is too small, we may have insufficient data for the classifier to learn, and thus yield bad results. As pointed out by Sammut and Webb [31], a ratio as low as 1:35 can make some methods inadequate for building a good model in some applications, while in some other situations, 1:10 may be tough to deal with. We should make different judgments based on different applications, datasets, sample sizes, and methods applied, etc.

Lastly, a very low case-control ratio in GWAS data may violate asymptotic assumptions of statistical inferences, such as that of the logistic regression models, which results in an inflated type I error rate [7]. For example, Chen et al. [2] assumes that the test statistic of genetic variants in a logistic mixed model asymptotically follows a Gaussian distribution under the null hypothesis, while the actual distributions are substantially different from Gaussian distribution when the case-control ratio is extremely unbalanced [7].

More specifically, unbalanced data may violate assumptions in statistical inferences. If the number of cases is drastically smaller than the number of controls, these cases may be viewed as outliers in most statistical models, and hence it leads to a higher variation for the estimation of coefficients. As a result, it shrinks the absolute value of test statistics and yields a larger p-value, which makes a truly influential variant insignificant.

Let us illustrate the idea using a simulation example. Suppose our sample size is 100 and we only have one true predictor X. In a balanced setting, we generate 50 controls from Uniform (0, 0.3) distribution, and remaining 50 cases from Uniform (0.7, 1). We then generate an intermediate variable $W = X + \varepsilon$ where ε follows standard normal distribution. Finally, we connect the response with the only true predictor through an indicator function as

$$Y = I(W > 0.6) = \begin{cases} 1, \text{ if } W > 0.6, \\ 0, \textit{ otherwise}. \end{cases}$$

In an unbalanced setting, we generate 90 controls from Uniform (0, 0.3), 10 cases from Uniform (0.7, 1), and follow the same procedure to generate the binary response Y. After the data are simulated, we use logistic regression to evaluate the significance of X in both balanced and unbalanced settings. We repeat 100 times and obtain the standard error and p-value of the coefficient of the true predictor X. From the results demonstrated in Table 1, we can see that even for this simple situation (with only one predictor), the p-value for the unbalanced data is almost ten times higher and it leads to a wrong conclusion that X is not significant.

Table 1. The mean (standard error) of the simulation example across 100 replications.

Simulation Settings	Standard Error	*p*-Value
Balanced data	0.5956 (0.0275)	0.0275 (0.0689)
Unbalanced data	0.9731 (0.1410)	0.1916 (0.2664)

3. Generalized Linear Mixed Model Association Test

Linear mixed models (LMMs) has become popular in GWAS for various biomedical traits because of its power in correcting for population structure and genetic relatedness [38–45]. Some fast algorithms have been proposed to estimate the model parameters and the variance component of LMMs to meet the ultrahigh dimensional need of the GWAS settings, such as the efficient mixed-model association (EMMA) [38], the efficient mixed models expedited (EMMAX) [39], and the fast linear mixed models (FaST-LMM) [41], to name a few. However, as a model with continuous response, LMMs are not designed for binary traits.

Chen et al. [2] applied logistic mixed models to analyze binary traits for GWAS data as follows

$$\text{logit}(\pi_i) = X_i\alpha + G_i\beta + b_i,$$

where $\pi_i = P(y_i = 1 \mid X_i, G_i, b_i)$; $y_i \in \{1, 0\}$ is the probability for a binary disease status phenotype to be a case for subject i, conditional on their covariates X_i, genotype G_i, and random effects b_i. X_i is a $1 \times p$ row vector of covariates for subject i, α is a $p \times 1$ column vector of fixed covariate effects, G_i is the genotype of a genetic variant for subject i, and β is the fixed genetic effect. The random effects $b = \{b_1, b_2, \ldots, b_n\} \sim N_n(0, \tau V)$, where τ is a scale variance component parameter and V is usually the genetic relationship matrix (GRM) estimated from a large number of genetic variants.

Chen et al. [2] also proposed a generalized linear mixed model association test (GMMAT) to select important genetic variants. The GMMAT score test is constructed based on the null hypothesis H_0:$\beta = 0$, which leads to the same null logistic mixed model for all genetic variants: $\text{logit}(\pi_i) = X_i\alpha + b_i$. They fit the null logistic mixed model using the penalized quasi-likelihood (PQL) method [46] and the efficient AI- REML algorithm [47] to estimate the variance components. The algorithm will iteratively estimate the fixed effects α and random effects b_i under the null hypothesis until the process reaches convergence. The score of each genetic variant under the null hypothesis is defined as $T = G(y - \hat{y})$, where $G = (G_1, G_2, ..., G_n)^T$ is the n 1 column vector of genotypes, $y = (y_1, y_2, ..., y_n)^T$ is the n 1 column vector of observed outcomes, and $\hat{y} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)^T$ is the estimated value of y under H_0. The asymptotic p-value of each genetic variant is obtained by assuming that the test statistic T asymptotically follows a Gaussian distribution.

4. Scalable and Accurate Implementation of GEneralized Mixed Model

Zhou et al. [7] claimed that the type I error rate of GMMAT test can still be inflated under the presence of unbalanced binary traits because a very low case-control ratio may violate asymptotic assumptions of logistic regression models. They proposed a Scalable and Accurate Implementation of Generalized mixed model (SAIGE) based on the saddle point approximation (SPA) [48,49] to conduct the score test.

The SAIGE method still adopted the logistic mixed model structure from the GMMAT, but it improved the variance component and the test statistic to better account for imbalance of binary traits. Specifically, Zhou et al. [7] applied a state-of-the-art pre-conditioned conjugate gradient (PCG) approach [50,51] to solve linear systems for a large cohort without requiring the computation of GRM, which achieves faster iterations for large n. In addition, the SAIGE improved the calculation of the test statistic T. Specifically, the GMMAT test assumes that T asymptotically follows a Gaussian distribution under the null hypothesis, which only considers the first two moments (mean and variance). However, the underlying distribution of T can be substantially different from Gaussian distribution when the case-control ratio is unbalanced. The saddle point approximation is used to approximate the distribution of T using the entire cumulant generating function (CGF). Zhou et al. [7] claimed that the approximated CGF can provide a more accurate p-value than the GMMAT does under the presence of unbalanced binary phenotypes.

In simulation studies, Zhou et al. [7] showed that the GMMAT suffered from type I error inflation, but the SAIGE controlled the type I error rates effectively even when case-control ratios are extremely unbalanced (e.g., a case-control ratio of 1:99). Zhou et al. [7] also applied the SAIGE to the UK Biobank data, where most binary phenotypes have a case-control ratio around or lower than 1:100. For example, colorectal cancer has 4562 cases and 382,756 controls (with the case-control ratio of 0.012), glaucoma has 4462 cases and

397,761 controls (with the case-control ratio of 0.011), and thyroid cancer has 358 cases and 407,399 controls (with the case-control ratio of 0.0008). Although the analysis results of GMMAT suffered greatly from type I error inflation with no clear peaks in its Manhattan plot for the extremely unbalanced thyroid cancer phenotype (see Figure 1 of Zhou et al. [7]), the SAIGE approach successfully detected loci on Chromosome 9 that is well known for its association with thyroid cancer. Another data example with an extremely unbalanced case-control ratio is the MGI biorepository that is mentioned in the Introduction section. Dey et al. [3] analyzed four extremely unbalanced phenotypes from MGI by applying the SPA to approximate the p-value based on a test statistic that was also used in the SAIGE, and they reported a large number of significant markers whose nearby genes have been verified as truly associated with the corresponding phenotypes by other studies. Compared to SAIGE and GMMAT, the logistic regression model used by Dey et al. [3] did not exploit the random effect term.

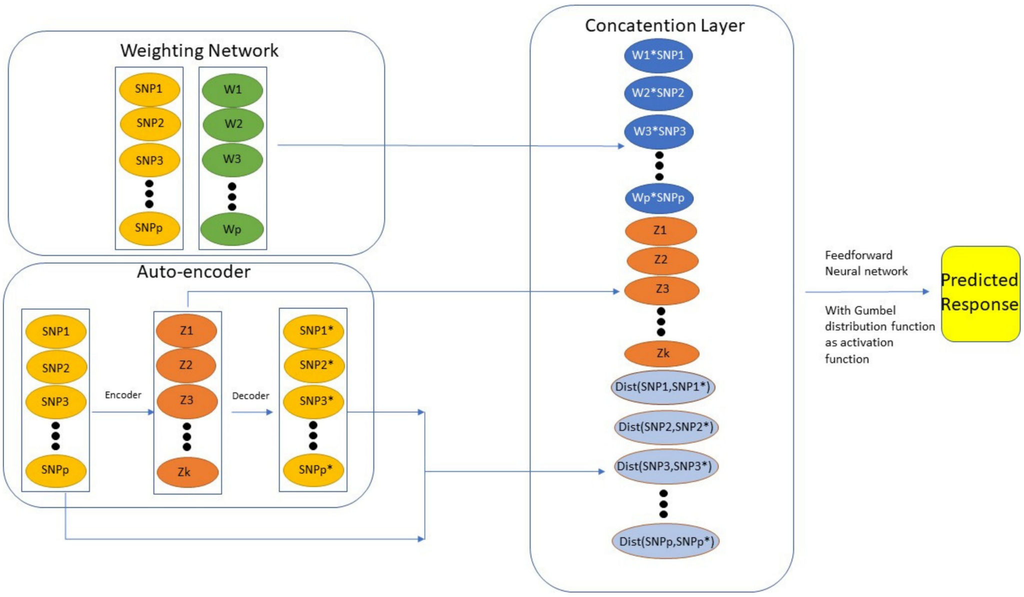

Figure 1. Overview of GEV-NN structure.

Both the GMMAT and SAIGE are single-SNP methods. The single-SNP method assesses the potential association of each genetic variant in isolation from the others. As a result, multiple limitations have been found for single-SNP methods: they inflate both false-positive and false-negative results [52]; they have limited detection and prediction ability because most complex diseases are actually polygenetic [24,53–56], where multiple variants affect the disease simultaneously but each one may have weak individual association [24,57]; they fail to differentiate potentially causative from non-causative variants if there exists a strong linkage disequilibrium (LD) between the noise variants and the truly influential ones.

In the remaining sections, we review several joint methods that have the advantages of considering multiple SNPs simultaneously.

5. Bayesian Multiple Logistic Regression Method

Various Bayesian approaches have been applied to the GWAS field to make genomic selections [57–63]. The benefits of Bayesian approaches lie in that they consider the polygenic effects of a large-scale of SNPs in GWAS using only one joint model and provide a feasible solution to estimate a large amount of unknown parameters. For example, Zhou, Carbonetto, and Stephens [57] proposed the Bayesian Sparse Linear Mixed Model (BSLMM) to select truly influential genes from the high-dimensional GWAS data with $p >> n$, where sparsity or regularization is expected. The selection is processed by putting sparsityenforcing priors on the regression coefficients so that the coefficients of non-influential genes can be forced to zero.

As an extension of BSLMM from the continuous to binary traits, a Bayesian multiple Logistic Regression method (B-LORE) was developed by setting point-normal priors on the regression coefficients:

$$p(\beta_i \mid \varphi_i, \sigma) = (1 - \varphi_i)\delta_0 + \varphi_i \, N(\beta_i \mid 0, \sigma^2),$$

where δ_0 is a constant, the hyperparameter φ_i controls the sparsity of the model, and σ^2 is the variance of the regression coefficients of influential genes. For binary traits, the likelihood function for the logistic regression model is maximized as follows:

$$\mathcal{L}(\beta) = \prod_{j=1}^{n} p_j^{y_j} \left(1 - p_j\right)^{1-y_j},$$

where $y_j \in \{1, 0\}$ is the observed phenotype for the jth subject, and p_j is the probability for a subject to have the disease given his/her genotypes and estimated regression coefficients $p_j = p\ (y_j = 1 \mid G, \hat{\beta})$. To estimate the hyperparameters φ_i and σ, Banerjee et al. [24] calculated the marginal likelihood function instead of the full likelihood function so that $\boldsymbol{\beta}$ can be integrated out in their posterior conditional distribution. This approach can greatly reduce the number of parameters to be estimated and can eventually improve its computational efficiency.

Unlike the p-values in single-SNP models, Bayesian approaches use a binary latent vector, say $\boldsymbol{c}$, to assess the importance of each SNP, with $c_i = 1$ meaning the ith SNP is influential and $c_i = 0$ meaning the ith SNP is not influential. See Banerjee et al. [24] for details about how the latent vector is incorporated and estimated through the joint regression model and MCMC sampling. Furthermore, Banerjee et al. [24] demonstrated that the B-LORE method outperforms other Bayesian variable selection methods through simulated data with case-control ratios as low as 0.25. Banerjee et al. [24] applied the B-LORE approach to the German Myocardial Infarction Family Study data with 6234 cases and 6848 controls of white European ancestry, which has a relatively balanced case-control ratio. For extremely unbalanced GWAS data, the priors of the B-LORE may require further adjustments.

However, it is also a legitimate concern that a Bayesian regression is much more computationally inefficient than the single-SNP GWAS methods due to the large number of parameters and the long iterative sampling process. There are many ways to improve the efficiency of a Bayesian algorithm, among which the most standard approach is to use conjugate priors and fast Gibbs sampling applied on the unknown parameters [64]. Some others tried to reduce the computational cost of a Bayesian regression from multiple aspects, such as the fast estimation of covariance matrices [65] and the use of marginal likelihood of predictors instead of a full model likelihood [66,67], to name a few.

6. Support Vector Machine

The Support Vector Machine (SVM) [30] is a well-known machine-learning algorithm designed for classification of binary traits. It aims to locate an optimal hyperplane from the high-dimensional predictor space to separate the two classes of binary traits. The separation is achieved by maximizing the minimum of the distances of every data point to the hyperplane (defined as $r_j, j = 1, \ldots, n$). The result of SVM is interpretable because it tracks which genetic variants are used to construct the separating hyper-plane and classify binary traits. However, SVM may suffer from overfitting and loss of generalizability for high-dimensional data.

Marron et al. (2007) [68] proposed the Distance Weighted Discrimination (DWD) for classifications on high-dimensional and low sample-sized data ($p >> n$), which facilitates the applicability of SVM algorithm to case-control disease data in the GWAS fields. The DWD improves the standard SVM by locating the hyperplane that minimizes the sum of the reciprocals of r_j (i.e., $\min \sum_{j=1}^{n} \frac{1}{r_j}$). The DWD approach overcomes the challenges of high-dimensional classification by allowing all data points to have influences on the separating hyperplane, rather than considering only the point that is closest to the separating hyperplane.

However, as claimed by Qiao et al. [69], the DWD does not perform well on unbalanced data if the proportions of the two classes are quite different from each other. Qiao et al. [69] proposed a weighted sum of the reciprocals of r_j ($\min \sum_{j=1}^{n} \frac{w_j}{r_j}$) as an improved objective function to locate the separating hyperplane, where w_j is the weight of the jth sample. They demonstrated through simulations and real data examples that the weighted DWD yields accurate and robust prediction results under nonstandard situations such as unbalanced binary traits. Additionally, they also proved Fisher consistency for the weighted DWD approach to provide statistical guarantee (see Qiao et al. [69] for more detailed theoretical results). Since a naive classifier easily favors the majority class, a common strategy that address the imbalance is to assign a higher misclassification cost to the underrepresented minority class of the binary trait. Qiao and Liu [70] developed an optimal weighting scheme using the Bayes decision rule with Mean Within Group Error (MWGE) criterion to determine the weight, w_j, of each sample.

7. AdaBoost

Ensemble learning based on decision trees has been effective in achieving a balance between overfitting and under-fitting and also reducing variance of predictions through aggregating prediction results of multiple classifiers [71]. Another advantage of decision tree-based algorithms is that the hierarchical structures of decision trees naturally considers epistasis among variants without requiring an explicitly model structure [72]. For example, Bagging [73], Random forests [74], and AdaBoost [75,76] are some of the most well-known decision tree-based ensemble learning algorithms. In particular, the AdaBoost algorithm is known to be powerful for classifying unbalanced binary traits and capable of reducing bias of single classifiers [77]. It puts more weights on the subjects that are most often misclassified by the preceding decision trees. The AdaBoost algorithm is comprised of a series of "weak" decision trees, but the final model can still yield promising prediction performances as long as each tree can learn additional information from a subset of the subjects that are not achieving good results in preceding decision trees.

Despite the fact that the AdaBoost method already put more weight on misclassified subjects, it still treats subjects of the binary traits equally: weights of misclassified (correctly classified) subjects are increased (decreased) by the same percentage no matter they come from the majority class or from the minority class. In order to support the under-represented minority class and truly address the unbalanced case-control issue, Sun et al. [77] adjusted AdaBoost by assigning higher misclassification costs to the subjects coming from the minority class than those of the majority class. See Sun et al. [77] for details about the calculation of subject weights.

8. Neural Network

Frasca et al. [78] proposed a cost sensitive neural network (COSNet) method, which can handle unbalanced responses by utilizing a suitable Hopfield Network and learning parameter through a cost sensitive optimization procedure. They also introduced a regularized cost sensitive neural network (RCOSNet) by adding a regularization term into the energy function of the network that can deal with extremely unbalanced classification problems. These two characteristics make RCOSNet applicable to case-control disease GWAS datasets. Zhang et al. [79] proposed a stacked de-noising Auto-encoder neural network (SDAE) algorithm based on cost-sensitive oversampling. Cost- sensitive oversampling exploited misclassification cost as the weight of the original data and duplicate samples based on that weight to maintain balance between different classes [80]. They added noises into the input features using a Gaussian distribution, or Salt and pepper distribution, and hence improved the classification accuracy of the minority class compared to traditional stacked Auto-encoder neural network.

Munkhdalai et al. [81] proposed the Generalized Extreme Value distribution Neural Network (GEV-NN) that consists of three components: "Weighting Layer", "Auto-encoder Layer", and "Concatenation Layer". The "Weighting Layer" gives weight to each predictor by multiplying a factor in front to each predictor. The "Auto-encoder Layer" extracts important features out from samples. The "Concatenation Layer" combined the previous two components and feed the result to the final prediction function. In order to overcome the imbalance issue, they used a Gumbel distribution as an activation function in the network. Figure 1 illustrates the model structure proposed by Munkhdalai et al. [81]. By using the Auto-encoder, they generate efficient features (which in this case is the distance between original inputs and reconstructed inputs) for minority classes (as shown in the "Concatenation Layer" of Figure 1). See Munkhdalai et al. [81] for more technical details. Actually, Kweon et al. [82] has already compared the GEV-NN to some baseline methods such as logistic regression, random forest, AdaBoost, XGBoost, and Support Vector Machine using a health-related dataset to predict hypertension. GEV-NN achieved the best prediction performance in terms of a number of evaluation metrics including G-mean, AUC, Accuracy, Brier score, and F score.

Another well-known variable selection mechanism for the neural network was the so-called "dropout" proposed by Srivastava et al. [83], which is one of the most highly cited machine-learning research methods. In order to avoid overfitting, Srivastava et al. [83] proposed to randomly drop features (including original predictors and engineered features) from the neural network during training process and evaluate their impacts on predictions to select an optimal "thinned" network. The "dropout" strategy can be regarded as regularization of neural networks including GEV-NN and can be used for variable selection purposes. Therefore, the "dropout" strategy has great potential for selecting influential variants from high-dimensional GWAS data.

9. Significance Test

Performing the hypothesis test and statistical significance for each variant has been the core of the GWAS field, and, therefore, whether the p-value can be obtained is crucial to select significant genetic variants. However, the statistical significance study for the state-of-the-art machine-learning approaches is still under-developed. In order to facilitate wide applications of machine-learning approaches into the GWAS field, we refer the readers to the well-established permutation-based significance test skills, which have already been applied to some classifiers to obtain p-values [84,85]. For example, Chen et al. [86] performed the statistical significance test by permuting variable importance scores obtained from the random forests approach to obtain p-values [74]. As a non-parametric approach, the permutation-based statistical significance test can be applied to the SVM (Section 6), AdaBoost (Section 7), and Neural Network (Section 8) to obtain a p-value for each variant.

The distribution of a test statistic is empirically established by permuting the original data with a large amount of time. Then an empirical p-value of each variant is approximated by counting the fraction of the test statistic scores of permuted data that are larger than that of the original data [84,85]. The accuracy of p-values depends on the original data (whether there exist any real associations between variants and binary response) as well as the classifier itself (whether the classifier is able to discover these associations) [84]. Actually, prediction and significance studies are related. Specifically, a promising prediction performance of a machine-learning method can be an indicator of a good understanding of the dependency structure between the predictors and the response, which is very important for constructing a reliable and powerful significance test [84]. See Ojala and Garriga [84] for more details about the theoretical properties of permutation-based tests.

10. Conclusions

With the collection of a large-scale of diseases from participants in large cohorts such as biobanks and EHRs, it raises rapidly increasing demands on statistical and machine-learning methods driven by unbalanced case-control GWAS data analysis needs. In this article, we reviewed multiple methods that are designed to address the imbalance in binary traits, including GMMAT and SAIGE that were based on logistic mixed models, the Bayesian variable selection method B-LORE, and machine-learning approaches such as SVM, AdaBoost, and the neural network. Each method has its own advantages as well as limitations, as summarized in Table 2. It is impossible to find any method that is uniformly the best. Therefore, methods should be chosen according to the needs of different aims, backgrounds, and scientific questions for different datasets. For example, if one wants to build a high-performance disease risk prediction tool along with a ranking of the most influential genetic variants meanwhile taking care of the nonlinear and gene-gene interactions, the AdaBoost algorithm along with its variable importance measure will be an ideal option.

Table 2. A summarization of the methods evaluated from different aspects mentioned in the manuscript. Each method has its own advantages and limitations.

	Can the Method Be Applied to Genomic Selections?	**Can the Method Be Applied to Genomic Predictions?**	**Can the Method Handle Unbalanced Binary Response?**
GMMAT	√ GMMAT is designed for performing the significance test of each variant.	✘ GMMAT is a single-SNP method and is not good for prediction.	✘ Its significance test assumes a Gaussian distribution, which is not the case for unbalanced data.
SAIGE	√ SAIGE is designed for performing the significance test of each variant.	✘ SAIGE is a single-SNP method and is not good for prediction.	√ SAIGE use the entire cumulant generating function to approximate *p*- values.
B-LORE	√ B-LORE is a joint Bayesian variable selection regression method designed for high-dimensional variants.	√ B-LORE is a joint Bayesian regression and can be used for prediction.	✘ B-LORE cannot handle extremely unbalanced binary data.
SVM	√ SVM has not been widely used in GWAS field yet, but it has the potential to select important variants or use permutation-based testing to obtain significance.	√ SVM is a machine method with the strength of producing accurate prediction.	√ SVM with weighted DWD can handle extremely unbalanced binary data.
AdaBoost	√ AdaBoost has not been widely used in GWAS field yet, but it has the potential to select important variants or use permutation-based testing to obtain significance.	√ AdaBoost is a machine method with the strength of producing accurate prediction.	√ AdaBoost can handle extremely unbalanced binary data by assigning higher misclassification costs to the minority class.
Neural Network	√ Neural Network has not been widely used in GWAS field yet, but it has the potential to select important variants or use permutation-based testing to obtain significance.	√ Neural Network is a machine method with the strength of producing accurate prediction.	√ The RCOSNet and GEV-NN can handle extremely unbalanced binary data.

In addition to the unbalanced classification issue, there are several other common challenges and concerns in the GWAS literature, to name a few in the following: (1) The GWAS data easily involves millions of SNPs for only thousands of participants, the ultrahigh-dimensionality, "small *n* big *p*" or the curse of dimensionality issue, raise big challenges [87]. (2) LD is one of the most important, extensive, and widespread features in genomes, with 70–80% of genomes showing regions of high LD. As a result, it is difficult to separate the individual variants that are truly causative from those confounding spurious variants that are irrelevant to the phenotype but highly correlated with the causative loci due to LD [52]. (3) Epistasis is defined as nonlinear interactions among loci or among genes (GxG), has been gaining more and more attention for its substantial role in regulating biological traits [88–96]. (4) The underlying population structure acting as confounders in GWAS data [72]. (5) A method that is computationally efficient is desired in the GWAS field due to the extremely high volume of computational needs.

To overcome the ultrahigh dimensionality challenge, Carlsen et al. [52] proposed a two-stage framework to extensively eliminate a large amount of noise SNPs using feature screening skills (its theoretical sure screening consistency is guaranteed), and then applied a sophisticated model to analyze the remaining variables in depth. They demonstrated that the accuracy and speed of genomic selection from the whole-genome data using this two-stage approach outperformed the approaches that applied only logistic ridge regression model or only a single-SNP approach. This two-stage framework is flexible enough to bridge any machine-learning approaches introduced in this article with the sure independence screening (SIS) feature screening approaches so that the performance of machine-learning approaches is not affected much by the ultrahigh dimensionality.

In addition to genomic selection, phenotypic prediction such as prediction of disease status or population disease prevalence using GWAS data repositories have also attracted a lot of research attention lately [24,27,28,97,98]. For example, Banerjee et al. [24] tried to predict the risk of coronary artery disease (CAD) for participants with white European ancestry. We want to emphasize that prediction has been the focus and strength of machine-learning approaches. However, significance- and inference-related research is still under-developed in machine-learning fields. We hope that this article highlights the importance of incorporating machine-learning approaches into the GWAS field, so that significance- and inference-related research could be improved in machine-learning approaches in the future.

Author Contributions: X.D. and G.F. conceived the research. X.D. prepared the original manuscript. G.F. revised the manuscript. S.Z. wrote Section 2. Y.Z. wrote Section 8. All authors participated in the revision process. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study did not require ethical approval.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thank David Collins for revising the English. We thank the editors and three anonymous reviewers whose comments helped significantly improve this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sudlow, C.; Gallacher, J.; Allen, N.; Beral, V.; Burton, P.; Danesh, J.; Downey, P.; Elliott, P.; Green, J.; Landray, M.; 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.* **2015**, *12*, e1001779. [CrossRef]
2. Chen, H.; Wang, C.; Conomos, M.P.; Stilp, A.M.; Li, Z.; Sofer, T.; Szpiro, A.A.; Chen, W.; Brehm, J.M.; Celedón, J.C.; et al. Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *Am. J. Hum. Genet.* **2016**, *98*, 653–666. [CrossRef]
3. Dey, R.; Schmidt, E.M.; Abecasis, G.R.; Lee, S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *Am. J. Hum. Genet.* **2017**, *101*, 37–49. [CrossRef]
4. Fritsche, L.G.; Gruber, S.B.; Wu, Z.; Schmidt, E.M.; Zawistowski, M.; Moser, S.E.; Blanc, V.M.; Brummett, C.M.; Kheterpal, S.; Abecasis, G.R.; et al. Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. *Am. J. Hum. Genet.* **2018**, *102*, 1048–1061. [CrossRef] [PubMed]
5. MacArthur, J.; Bowler, E.; Cerezo, M.; Gil, L.; Hall, P.; Hastings, E.; Junkins, H.; McMahon, A.; Milano, A.; Morales, J.; et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **2017**, *45*, D896–D901. [CrossRef]
6. Visscher, P.M.; Wray, N.R.; Zhang, Q.; Sklar, P.; McCarthy, M.I.; Brown, M.A.; Yang, J. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **2017**, *101*, 5–22. [CrossRef] [PubMed]
7. Zhou, W.; Nielsen, J.B.; Fritsche, L.G.; Dey, R.; Gabrielsen, M.E.; Wolford, B.N.; LeFaive, J.; VandeHaar, P.; Gagliano, S.A.; Gifford, A.; et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic associa-tion studies. *Nat. Genet.* **2018**, *50*, 1335–1341. [CrossRef]
8. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared con-trols. *Nature* **2007**, *447*, 661. [CrossRef] [PubMed]

9. Cooper, J.D.; The Type I Diabetes Genetics Consortium; Walker, N.M.; Smyth, D.J.; Downes, K.; Healy, B.C.; Todd, J.A. Follow-up of 1715 SNPs from the Wellcome Trust Case Control Consortium genome-wide association study in type I diabetes families. *Genes Immun.* **2009**, *10*, S85–S94. [CrossRef]
10. Maller, J.B.; The Wellcome Trust Case Control Consortium; McVean, G.; Byrnes, J.; Vukcevic, D.; Palin, K.; Su, Z.; Howson, J.M.M.; Auton, A.; Myers, S.; et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **2012**, *44*, 1294–1301. [CrossRef] [PubMed]
11. Reilly, M.P.; Li, M.; He, J.; Ferguson, J.F.; Stylianou, I.M.; Mehta, N.N.; Burnett, M.S.; Devaney, J.M.; Knouff, C.W.; Thompson, J.R.; et al. Identification of ADAMTS7 as a novel locus for coronary atherosclerosis and association of ABO with my-ocardial infarction in the presence of coronary atherosclerosis: Two genome-wide association studies. *Lancet* **2011**, *377*, 383–392. [CrossRef]
12. Holmans, P.; Green, E.K.; Pahwa, J.S.; Ferreira, M.A.; Purcell, S.M.; Sklar, P.; Owen, M.J.; O'Donovan, M.C.; Craddock, N. Gene Ontology Analysis of GWA Study Data Sets Provides Insights into the Biology of Bipolar Disorder. *Am. J. Hum. Genet.* **2009**, *85*, 13–24. [CrossRef]
13. Thomson, W.; Barton, A.; Ke, X.; Eyre, S.; Hinks, A.; Bowes, J.; Donn, R.; Symmons, D.; Hider, S.; Bruce, I.N.; et al. Rheu-matoid arthritis association at 6q23. *Nat. Genet.* **2007**, *39*, 1431. [CrossRef]
14. Eyre, S.; Bowes, J.; Diogo, D.; Lee, A.; Barton, A.; Martin, P.; Zhernakova, A.; Stahl, E.; Viatte, S.; McAllister, K.; et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat. Genet.* **2012**, *44*, 1336–1340. [CrossRef] [PubMed]
15. Dai, X.; Fu, G.; Reese, R. Detecting PCOS susceptibility loci from genome-wide association studies via iterative trend corre-lation based feature screening. *BMC Bioinform.* **2020**, *21*, 1–15. [CrossRef] [PubMed]
16. Bycroft, C.; Freeman, C.; Petkova, D.; Band, G.; Elliott, L.T.; Sharp, K.; Motyer, A.; Vukcevic, D.; Delaneau, O.; O'Connell, J.; et al. Genome-wide genetic data on 500,000 UK Biobank participants. *BioRxiv* **2017**, 166298. [CrossRef]
17. Wang, H.; Smith, K.P.; Combs, E.; Blake, T.; Horsley, R.D.; Muehlbauer, G.J. Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. *Theor. Appl. Genet.* **2011**, *124*, 111–124. [CrossRef]
18. Cortes, A.; Hadler, J.; Pointon, J.P.; Robinson, P.C.; Karaderi, T.; Leo, P.; Cremin, K.; Pryce, K.; Harris, J.; Lee, S.; et al. Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related lo-ci. *Nat. Genet.* **2013**, *45*, 730.
19. Dawson, J.C.; Endelman, J.B.; Heslot, N.; Crossa, J.; Poland, J.; Dreisigacker, S.; Manès, Y.; Sorrells, M.E.; Jannink, J.-L. The use of unbalanced historical data for genomic selection in an international wheat breeding program. *Field Crop. Res.* **2013**, *154*, 12–22. [CrossRef]
20. Fakiola, M.; Strange, A.; Cordell, H.J.; Miller, E.N.; Pirinen, M.; Su, Z.; Mishra, A.; Mehrotra, S.; Monteiro, G.R.; Band, G.; et al. Common variants in the HLA-DRB1–HLA-DQA1 HLA class II region are associated with susceptibility to visceral leishmaniasis. *Nat. Genet.* **2013**, *45*, 208–213. [CrossRef] [PubMed]
21. Fingerlin, T.E.; Murphy, E.; Zhang, W.; Peljto, A.L.; Brown, K.K.; Steele, M.P.; Loyd, J.E.; Cosgrove, G.P.; Lynch, D.; Groshong, S.; et al. Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat. Genet.* **2013**, *45*, 613–620. [CrossRef]
22. Liu, J.Z.; The UK-PSCSC Consortium; Hov, J.R.; Folseraas, T.; Ellinghaus, E.; Rushbrook, S.M.; Doncheva, N.T.; Andreassen, O.A.; Weersma, R.K.; Weismüller, T.J.; et al. Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. *Nat. Genet.* **2013**, *45*, 670–675. [CrossRef]
23. Ma, C.; Blackwell, T.; Boehnke, M.; Scott, L.J. the GoT2D Investigators Recommended Joint and Meta-Analysis Strategies for Case-Control Association Testing of Single Low-Count Variants. *Genet. Epidemiol.* **2013**, *37*, 539–550. [CrossRef] [PubMed]
24. Banerjee, S.; Zeng, L.; Schunkert, H.; Söding, J. Bayesian multiple logistic regression for case-control GWAS. *PLoS Genet.* **2018**, *14*, e1007856. [CrossRef]
25. Li, Y.; Levran, O.; Kim, J.; Zhang, T.; Chen, X.; Suo, C. Extreme sampling design in genetic association mapping of quantita-tive trait loci using balanced and unbalanced case-control samples. *Sci. Rep.* **2019**, *9*, 1–9.
26. Zhang, X.; Basile, A.O.; Pendergrass, S.A.; Ritchie, M.D. Real world scenarios in rare variant association analysis: The impact of imbalance and sample size on the power in silico. *BMC Bioinform.* **2019**, *20*, 46. [CrossRef] [PubMed]
27. Barr, R.G.; Avilés-Santa, L.; Davis, S.M.; Aldrich, T.K.; Ii, F.G.; Henderson, A.G.; Kaplan, R.C.; LaVange, L.; Liu, K.; Loredo, J.S.; et al. Pulmonary Disease and Age at Immigration among Hispanics. Results from the Hispanic Community Health Study/Study of Latinos. *Am. J. Respir. Crit. Care Med.* **2016**, *193*, 386–395. [CrossRef]
28. Schubach, M.; Re, M.; Robinson, P.N.; Valentini, G. Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants. *Sci. Rep.* **2017**, *7*, 2959. [CrossRef]
29. Sun, Y.; Wong, A.K.C.; Kamel, M.S. Classification of imbalanced data: A review. *Int. J. Pattern Recognit. Artif. Intell.* **2009**, *23*, 687–719. [CrossRef]
30. Vapnik, V. *The Nature of Statistical Learning Theory, Number 401–403*; Springer Science & Business Media: Berlin, Germany, 2013.
31. Sammut, C.; Webb, G.I. *Encyclopedia of Machine Learning*; Springer Science & Business Media: Berlin, Germany, 2011.
32. Xue, J.-H.; Titterington, D.M. Do unbalanced data have a negative effect on LDA? *Pattern Recognit.* **2008**, *41*, 1558–1571. [CrossRef]
33. Drummond, C.; Holte, R.C. Severe Class Imbalance: Why Better Algorithms Aren't the Answer. In *Proceedings of the Computer Vision*; Springer Science and Business Media LLC: Berlin, Germany, 2005; Volume 3720, pp. 539–546.

34. Chawla, N.V.; Japkowicz, N.; Kotcz, A. Special issue on learning from imbalanced data sets. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 1–6. [CrossRef]
35. Ling, C.X.; Huang, J.; Zhang, H. AUC: A statistically consistent and more discriminating measure than accuracy. *Ijcai* **2003**, *3*, 519–524.
36. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [CrossRef]
37. Zhou, L. Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. *Knowl.-Based Syst.* **2013**, *41*, 16–25. [CrossRef]
38. Kang, H.M.; Zaitlen, N.A.; Wade, C.M.; Kirby, A.; Heckerman, D.; Daly, M.J.; Eskin, E. Efficient control of population struc-ture in model organism association mapping. *Genetics* **2008**, *178*, 1709–1723. [CrossRef]
39. Kang, H.M.; Sul, J.H.; Service, S.K.; Zaitlen, N.A.; Kong, S.-Y.; Freimer, N.B.; Sabatti, C.; Eskin, E. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **2010**, *42*, 348–354. [CrossRef] [PubMed]
40. Zhang, Z.; Ersoz, E.; Lai, C.-Q.; Todhunter, R.J.; Tiwari, H.K.; Gore, M.; Bradbury, P.J.; Yu, J.; Arnett, D.K.; Ordovas, J.M.; et al. Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **2010**, *42*, 355–360. [CrossRef] [PubMed]
41. Lippert, C.; Listgarten, J.; Liu, Y.; Kadie, C.M.; Davidson, R.I.; Heckerman, D. FaST linear mixed models for genome-wide association studies. *Nat. Methods* **2011**, *8*, 833–835. [CrossRef] [PubMed]
42. Yang, J.; Lee, S.H.; Goddard, M.E.; Visscher, P.M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.* **2011**, *88*, 76–82. [CrossRef]
43. Svishcheva, G.R.; Axenovich, T.I.; Belonogova, N.M.; Van Duijn, C.M.; Aulchenko, Y.S. Rapid variance components–based method for whole-genome association analysis. *Nat. Genet.* **2012**, *44*, 1166–1170. [CrossRef]
44. Zhou, X.; Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **2012**, *44*, 821–824. [CrossRef]
45. Loh, P.-R.; Tucker, G.J.; Bulik-Sullivan, B.K.; Vilhjálmsson, B.J.; Finucane, H.K.; Salem, R.M.; Chasman, D.I.; Ridker, P.M.; Neale, B.M.; Berger, B.; et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **2015**, *47*, 284–290. [CrossRef]
46. Breslow, N.E.; Clayton, D.G. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* **1993**, *88*, 9–25.
47. Gilmour, A.R.; Thompson, R.; Cullis, B.R. Average information REML: An efficient algorithm for variance parameter esti-mation in linear mixed models. *Biometrics* **1995**, 1440–1450. [CrossRef]
48. Imhof, J.P. Computing the distribution of quadratic forms in normal variables. *Biometrika* **1961**, *48*, 419–426. [CrossRef]
49. Kuonen, D. Miscellanea. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **1999**, *86*, 929–935. [CrossRef]
50. Hestenes, M.; Stiefel, E. Methods of conjugate gradients for solving linear systems. *J. Res. Natl. Inst. Stand. Technol.* **1952**, *49*, 409. [CrossRef]
51. Kaasschieter, E. Preconditioned conjugate gradients for solving singular systems. *J. Comput. Appl. Math.* **1988**, *24*, 265–275. [CrossRef]
52. Carlsen, M.; Fu, G.; Bushman, S.; Corcoran, C. Exploiting Linkage Disequilibrium for Ultrahigh-Dimensional Genome-Wide Data with an Integrated Statistical Approach. *Genetics* **2016**, *202*, 411–426. [CrossRef] [PubMed]
53. Hoggart, C.J.; Whittaker, J.C.; De Iorio, M.; Balding, D.J. Simultaneous Analysis of All SNPs in Genome-Wide and Re-Sequencing Association Studies. *PLoS Genet.* **2008**, *4*, e1000130. [CrossRef]
54. Weeks, D.E.; Lathrop, G. Polygenic disease: Methods for mapping complex disease traits. *Trends Genet.* **1995**, *11*, 513–519. [CrossRef]
55. Van Rheenen, W.; Peyrot, W.J.; Schork, A.J.; Lee, S.H.; Wray, N.R. Genetic correlations of polygenic disease traits: From the-ory to practice. *Nat. Rev. Genet.* **2019**, *20*, 567–581. [CrossRef]
56. Wald, N.J.; Old, R. The illusion of polygenic disease risk prediction. *Genet. Med.* **2019**, *21*, 1705–1707. [CrossRef] [PubMed]
57. Zhou, X.; Carbonetto, P.; Stephens, M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genet.* **2013**, *9*, e1003264. [CrossRef] [PubMed]
58. Servin, B.; Stephens, M. Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genet.* **2007**, *3*, e114. [CrossRef]
59. Guan, Y.; Stephens, M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* **2011**, *5*, 1780–1815. [CrossRef]
60. Li, J.; Das, K.; Fu, G.; Li, R.; Wu, R. The Bayesian lasso for genome-wide association studies. *Bioinformatics* **2010**, *27*, 516–523. [CrossRef] [PubMed]
61. Carbonetto, P.; Stephens, M. Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies. *Bayesian Anal.* **2012**, *7*, 73–108. [CrossRef]
62. Bottolo, L.; Chadeau-Hyam, M.; Hastie, D.I.; Zeller, T.; Liquet, B.; Newcombe, P.; Yengo, L.; Wild, P.S.; Schillert, A.; Ziegler, A.; et al. GUESS-ing Polygenic Associations with Multiple Phenotypes Using a GPU-Based Evolutionary Stochastic Search Algorithm. *PLoS Genet.* **2013**, *9*, e1003657. [CrossRef]
63. Liquet, B.; Bottolo, L.; Campanella, G.; Richardson, S.; Chadeau-Hyam, M. R2GUESS: A Graphics Processing Unit-Based R Package for Bayesian Variable Selection Regression of Multivariate Responses. *J. Stat. Softw.* **2016**, *69*, 1–32. [CrossRef]

64. George, E.I.; McCulloch, R.E. Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.* **1993**, *88*, 881–889. [CrossRef]
65. Ishwaran, H.; Rao, J.S. Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Stat.* **2005**, *33*, 730–773. [CrossRef]
66. Shang, Z.; Clayton, M.K. Consistency of Bayesian linear model selection with a growing number of parameters. *J. Stat. Plan. Inference* **2011**, *141*, 3463–3474. [CrossRef]
67. Narisetty, N.N.; He, X. Bayesian variable selection with shrinking and diffusing priors. *Ann. Stat.* **2014**, *42*, 789–817. [CrossRef]
68. Marron, J.S.; Todd, M.J.; Ahn, J. Distance-Weighted Discrimination. *J. Am. Stat. Assoc.* **2007**, *102*, 1267–1271. [CrossRef]
69. Qiao, X.; Zhang, H.H.; Liu, Y.; Todd, M.J.; Marron, J.S. Weighted distance weighted discrimination and its asymptotic prop-erties. *J. Am. Stat. Assoc.* **2010**, *105*, 401–414. [CrossRef] [PubMed]
70. Qiao, X.; Liu, Y. Adaptive Weighted Learning for Unbalanced Multicategory Classification. *Biometrics* **2008**, *65*, 159–168. [CrossRef]
71. Kittler, J.; Hatef, M.; Duin, R.P.W.; Matas, J. On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 226–239. [CrossRef]
72. Fu, G.; Dai, X.; Symanzik, J.; Bushman, S. Quantitative gene-gene and gene-environment mapping for leaf shape variation using tree-based models. *New Phytol.* **2017**, *213*, 455–469. [CrossRef]
73. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
74. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
75. Friedman, J.; Hastie, T.; Tibshirani, R. Additive logistic regression: A statistical view of boosting. *Ann. Stat.* **2000**, *28*, 337–407. [CrossRef]
76. Hastie, T.; Rosset, S.; Zhu, J.; Zou, H. Multi-class AdaBoost. *Stat. Interface* **2009**, *2*, 349–360. [CrossRef]
77. Sun, Y.; Kamel, M.S.; Wong, A.K.; Wang, Y. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognit.* **2007**, *40*, 3358–3378. [CrossRef]
78. Frasca, M.; Bertoni, A.; Re, M.; Valentini, G. A neural network algorithm for semi-supervised node label learning from un-balanced data. *Neural Netw.* **2013**, *43*, 84–98. [CrossRef] [PubMed]
79. Zhang, C.; Gao, W.; Song, J.; Jiang, J. An imbalanced data classification algorithm of improved autoencoder neural network. In Proceedings of the 2016 Eighth International Conference on Advanced Computational Intelligence (ICACI), Chiang Mai, Thailand, 14–16 February 2016; pp. 95–99.
80. Elkan, C. *The Foundations of Cost-Sensitive Learning. International Joint Conference on Artificial Intelligence*; Lawrence Erlbaum Associates Ltd.: Mahwah, NJ, USA, 2001; Volume 17, pp. 973–978.
81. Munkhdalai, L.; Munkhdalai, T.; Ryu, K.H. GEV-NN: A deep neural network architecture for class imbalance problem in binary classification. *Knowl.-Based Syst.* **2020**, *194*, 105534. [CrossRef]
82. Kweon, S.; Kim, Y.; Jang, M.J.; Kim, Y.; Kim, K.; Choi, S.; Chun, C.; Khang, Y.H.; Oh, K. Data resource profile: The Korea na-tional health and nutrition examination survey (KNHANES). *Int. J. Epidemiol.* **2014**, *43*, 69–77. [CrossRef] [PubMed]
83. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural net-works from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
84. Ojala, M.; Garriga, G.C. Permutation Tests for Studying Classifier Performance. In Proceedings of the 2009 Ninth IEEE International Conference on Data Mining, Miami, FL, USA, 6 December 2009; pp. 908–913. [CrossRef]
85. Modarres, R.; Good, P. Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. *J. Am. Stat. Assoc.* **1995**, *90*, 384. [CrossRef]
86. Chen, X.; Liu, C.-T.; Zhang, M.; Zhang, H. A forest-based approach to identifying gene and gene gene interactions. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 19199–19203. [CrossRef]
87. Qian, J.; Tanigawa, Y.; Du, W.; Aguirre, M.; Chang, C.; Tibshirani, R.; Rivas, M.A.; Hastie, T. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLoS Genet.* **2020**, *16*, e1009141. [CrossRef]
88. Yang, Q.; Khoury, M.J.; Sun, F.; Flanders, W.D. Case-Only Design to Measure Gene-Gene Interaction. *Epidemiology* **1999**, *10*, 167–170. [CrossRef]
89. Howard, T.D.; Koppelman, G.H.; Xu, J.; Zheng, S.L.; Postma, D.S.; Meyers, D.A.; Bleecker, E.R. Gene-Gene Interaction in Asthma: IL4RA and IL13 in a Dutch Population with Asthma. *Am. J. Hum. Genet.* **2002**, *70*, 230–236. [CrossRef] [PubMed]
90. Peng, D.Q.; Zhao, S.P.; Nie, S.; Li, J. Gene-gene interaction of PPARγ and ApoE affects coronary heart disease risk. *Int. J. Cardiol.* **2003**, *92*, 257–263. [CrossRef]
91. Dong, C.; Chu, X.; Wang, Y.; Wang, Y.; Jin, L.; Shi, T.; Huang, W.; Li, Y. Exploration of gene–gene interaction effects using entropy-based methods. *Eur. J. Hum. Genet.* **2007**, *16*, 229–235. [CrossRef] [PubMed]
92. Cordell, H.J. Detecting gene–gene interactions that underlie human diseases. *Nat. Rev. Genet.* **2009**, *10*, 392–404. [CrossRef] [PubMed]
93. Yung, L.S.; Yang, C.; Wan, X.; Yu, W. GBOOST: A GPU-based tool for detecting gene–gene interactions in genome–wide case control studies. *Bioinformatics* **2011**, *27*, 1309–1310. [CrossRef]
94. Howson, J.M.; Cooper, J.D.; Smyth, D.J.; Walker, N.M.; Stevens, H.; She, J.-X.; Eisenbarth, G.S.; Rewers, M.; Todd, J.A.; Akolkar, B.; et al. Evidence of Gene-Gene Interaction and Age-at-Diagnosis Effects in Type 1 Diabetes. *Diabetes* **2012**, *61*, 3012–3017. [CrossRef] [PubMed]
95. Van Steen, K. Travelling the world of gene-gene interactions. *Brief. Bioinform.* **2011**, *13*, 1–19. [CrossRef]

96. Fathima, N.; Narne, P.; Ishaq, M. Association and gene–gene interaction analyses for polymorphic variants in CTLA-4 and FOXP3 genes: Role in susceptibility to autoimmune thyroid disease. *Endocrine* **2019**, *64*, 591–604. [CrossRef] [PubMed]
97. Damen, J.A.A.G.; Hooft, L.; Schuit, E.; Debray, T.P.; Collins, G.S.; Tzoulaki, I.; Lassale, C.M.; Siontis, G.C.M.; Chiocchia, V.; Roberts, C.; et al. Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ* **2016**, *353*, i2416. [CrossRef] [PubMed]
98. Farzadfar, F. Cardiovascular disease risk prediction models: Challenges and perspectives. *Lancet Glob. Health* **2019**, *7*, e1288–e1289. [CrossRef]