Genomic Machine Learning Meta-regression: Insights on Associations of Study Features With Reported Model Performance

Eric J. Barnett, Daniel G. Onete, Asif Salekin, and Stephen V. Faraone

Abstract—Many studies have been conducted with the goal of correctly predicting diagnostic status of a disorder using the combination of genomic data and machine learning. It is often hard to judge which components of a study led to better results and whether better reported results represent a true improvement or an uncorrected bias inflating performance. We extracted information about the methods used and other differentiating features in genomic machine learning models. We used these features in linear regressions predicting model performance. We tested for univariate and multivariate associations as well as interactions between features. Of the models reviewed, 46% used feature selection methods that can lead to data leakage. Across our models, the number of hyperparameter optimizations reported, data leakage due to feature selection, model type, and modeling an autoimmune disorder were significantly associated with an increase in reported model performance. We found a significant, negative interaction between model type and actual case/control labels. Researchers aim to build models that use real, generalizable differences between cases and controls to make each prediction, but in practice, models are free to take advantage of interactions between loci to find patterns that are statistically significantly different between cases of a disorder and healthy controls and to test for polygenicity. Across our models, the number of loci, researchers apply stringent genome-wide significance thresholds (most commonly \( p < 5 \times 10^{-8} \)) to reduce reporting false positive results.

I. INTRODUCTION

The genetic study of complex disorders has made great strides in the discovery of genome-wide significant genetic loci and substantial evidence for polygenicity [1]. These discoveries have generated new hypotheses about the etiology of these disorders and have motivated machine learning (ML) efforts to separate cases and controls using genome-wide association study (GWAS) data. While results from early genomic ML research had been promising, the potential pitfalls of such studies have limited their interpretation [2], [3]. Although best practices have previously been described, the methods, reporting, and overall study design for genomic ML studies vary so drastically that it is often difficult to compare and evaluate studies [4]. This between-study heterogeneity may contribute to distrust and underutilization of machine learning results.

To better appreciate the strengths and weaknesses of genomic ML research, one must understand the differences between ML analyses and traditional GWAS. GWAS seeks to determine loci that are statistically significantly different between cases of a disorder and healthy controls and to test for polygenicity [5]. Since these studies examine hundreds to millions of loci, researchers apply stringent genome-wide significance thresholds (most commonly \( p < 5 \times 10^{-8} \)) to reduce reporting false positive results.

In the ML analyses we review here, the primary goal is accurately predicting whether subjects are cases or controls genotype data from each individual. Towards achieving this goal, relying only on loci that meet the threshold for genome-wide significance limits the learning capability of ML models. For example, a schizophrenia GWAS found that while 108 genome wide significant loci were able to explain 3.4% of the variation on the liability scale, including loci that met the nominal significance threshold (0.05) increased the variation explained to 7% [6]. This effect may be more pronounced in ML models which take advantage of interactions between loci to find patterns that are useful in differentiating cases and controls since more loci give the models more potential to find patterns.

Including additional loci in ML models has some drawbacks. One of the most important aspects of generalizable ML models is avoiding overfitting, which becomes more difficult as the number of loci increases [7]. Overfitting occurs when a model learns patterns that are only present within the data used to train the model. In ML algorithms, the training process learns model parameters that minimize the difference between the predicted and actual case/control labels. Researchers aim to build models that use real, generalizable differences between cases and controls to make each prediction, but in practice, models are free to...
use whatever patterns best minimize that difference. If a model memorizes the noise specific to only the training data, the model is less motivated to learn patterns that may be more generalizable if using those patterns is less successful than memorizing training data noise. For many ML models, constraints are added to reduce the model’s ability to overfit, but overfitting is rarely completely avoided [7]. Each additional locus that a model has access to increases the probability of overfitting but also has the potential to add generalizable information the model can use to increase its ability to separate cases and controls.

Many ML researchers account for overfitting by testing the performance of their models on data that were not seen during training [8]. One of two methods is typically used: cross-validation within the training set and validation with data not used at all in the training process. In k-fold cross-validation, researchers randomly split the data into some number of subsets, called folds, and then complete the association analysis and modeling that detect and use loci that are different between cases and controls to best separate the two classes using all but one of the folds. The model is trained using data from k-1 folds and its accuracy is tested in the single fold that was not included in training. This process is repeated until the model has used each fold as the withheld subset. Then the results across all iterations are averaged.

In the hold-out method, researchers randomly split the data into either two (training and test) or three (training, validation, and test) subsets. The training subset is used for the association analysis and modeling. If present, the validation subset is used to tune the model by setting the optimal hyperparameters, which are all the options and configurations that are not trained by the model itself, to best predict the validation subset. Then, the test subset is used to measure and report model performance. Unlike in cross-validation where all folds are the same size, the hold-out method typically uses 60–80% of the data in the training subset, while the remaining data is split evenly between the validation and test subsets. Cross-validation is often used when data is limited since in this method the model has a chance to train using each person in the study. The hold-out method, while not allowing the model to train on each person, is thought to be the more conservative approach and less likely to produce an overfit model.

The value of external testing can easily be lost through methods that leak information about the test subset/fold into the training of a model [9]. When this occurs, the model can use that information to model the specific test data more accurately. Consequently, the test data no longer represent unseen data and no longer account for overfitting to the same degree. This results in a model that is biased in favor of the test data. This problem, called data leakage, is especially detrimental because it often goes unnoticed, leading researchers and their audiences to believe that their model performs exceptionally well even outside the training subset when instead they are observing model bias. The amount of data leakage caused through methodological issues can vary from slight leakages that may result in some overreporting of performance, to major leakages that cause the testing subset/fold to mimic the training performance with near-perfect prediction. One example of major data leakage is using the entire data set to select which of the many different features best separate cases and controls and only splitting the data into subsets during the actual modeling. In this situation, when the features were selected with the entire data set, information about the test data were directly leaked into the process and the features included in the model will perform well in the test data, but in unseen data will either be less predictive or entirely unpredictable if the features selected are specific to the data in the feature selection process. A less severe form of data leakage may occur when the dataset used in the study is a part of a larger GWAS meta-analysis that the study uses for feature selection. In this case, the influence of the data leakage is diluted by the other data within the GWAS meta-analysis and selecting features that are uniquely predictive in the test data are less likely. However, the test data are still used in the feature selection process and results will likely be inflated.

The choices researchers make regarding which ML models to train, loci to include, and methods to use for measuring performance are critical decisions that will determine the outcome and validity of their study. But since few studies compare ML variations in the same external data sets, comparisons between studies are difficult even within the same disorder. This leads to a potential dilemma: Is a model with a higher reported performance better than a lower performing model or more overfit to their data?

Here, we report a systematic review that extracted information on model performance, disorder, training size, ML methods, performance measurement methods, and reporting on common issues from all genomic ML papers found in our search. We use this information in a mixed-effects linear regression model predicting model accuracy as measured by the area under the receiver operating characteristic curve (AUC). Within each data set and study involving machine learning models, many factors, such as training size, model type, and disorder, likely have a real, generalizable impact on reported model performance. When looking at a group of studies, those factors may be harder to identify due to the variable effects of data leakage among the studies in the group. We hypothesized that data leakage would be associated with a significant increase in reported AUC. We sought to test this hypothesis and to identify other study features that lead to increased reported AUC. The purpose of our study is not to scrutinize any specific study or model, but to instead look at the current work in the field as a whole to understand what can be strengthened or avoided in future work to improve results.

II. METHODS

Search strategy and selection: To identify studies that used genotype data as input for machine learning models to predict any disorder, we searched PubMed using the key-words ‘(GWAS OR genotype[ti] OR SNP OR “risk score”) AND (classif∗ OR predict∗) AND (“machine learning” OR “data mining”[ti] OR “neural network∗” OR “random forest” OR “support vector machine” OR “deep learning”)’. The search produced 1435 studies (up to July 6, 2022).
We excluded studies that were not using genomic data in classifiers to predict complex genetic disorders in humans. We excluded studies that did not report results for genotype only models because the impact of the features investigated in our analysis may differ between genomic and multi-omic models, which would add even more heterogeneity to the broad set of included studies. We excluded studies that did not report testing performance outside the data used to train the model because these models are likely overfit to the data given the ability of machine learning models to learn the random noise within the training data. Since models that use all data on training can achieve near perfect prediction on those data regardless of other methodological features inclusion of these studies could mask other important features, especially when model complexity is high as is often the case in genomic models [3]. We also excluded studies that only used variants identified in previous studies, since these studies are generally focused on a small number of specific variants that have been thoroughly studied and validated and therefore do not face the data leakage due to feature selection issue addressed in this paper. We excluded studies that did not report AUC as a performance metric since combining different performance metrics in our analysis would limit interpretation and potentially bias results. After exclusion criteria, 55 studies remained [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63]. These studies provided accuracy statistics for 99 models because some studies modeled multiple disorders. Supplementary Fig. 1 shows the article selection procedure in a PRISMA diagram.

Data extraction: We extracted the following data from each included study: disorder predicted, number of subjects used for training and testing, number of participants the model was trained on, highest AUC, method for testing/reporting model AUC, reporting on optimization, number of hyperparameters optimized, reporting on imputation, reporting on quality control procedures, and type of machine learning method with the highest AUC. We split studies that modeled multiple disorders such that each row of extracted data represented a single disorder from a single study.

Regression analysis: We fit linear mixed-effects models to test the individual and combined contribution of study variables to AUC using STATA17 [64]. We used this type of model to test the statistical significance of the included study features while accounting for the between-study heterogeneity present in this analysis of many different methods and disorders [65]. The standard error of each of our models was estimated using a clustered sandwich estimator clustering on PMID of the included studies, which adjusts standard errors for the lack of statistical independence of results within studies. We fit univariate models with AUC as the dependent variable and used the following as independent variables: data leakage through feature selection, disorder type, reporting of optimization, reporting of quality control procedures, reporting imputation usage, number of hyperparameter optimizations reported, disorder heritability, model type, testing method, and size of training dataset.

“Data leakage due to feature selection” was a binary feature scored as 1 if the data used to test model performance were used to select which features would be included in the model. It was scored zero otherwise. “Optimization reported” was a binary variable scored 1 if the authors reported any optimization of model hyperparameters. It was scored zero otherwise. “Quality control (QC) reported” was a binary variable scored 1 if authors reported the quality control procedures for their genotype data and zero otherwise. “Imputation reported” was a binary variable scored 1 if authors reported using imputed genotypes in their analysis and zero otherwise. “Number of hyperparameters optimizations” indicated the number of hyperparameters the authors reporting optimizing in the model with the highest AUC. “Disorder heritability” for each disorder included in these studies was gathered from heritability studies of those disorders based on twins or families [66], [15], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81], [82], [83], [84], [85], [86], [87], [88], [89], [90]. Training size was the total number of cases and controls used to train the model. “Model type” was a binary variable scored 1 if the model with the highest AUC was non-linear and 0 if it was linear. “Testing method” was a binary variable scored 1 if cross-validation was used to test model performance and 0 if a hold-out test subset was used to test model performance. Scatter plots of numerical variables and box plots of binary variables can be found in the Supplementary Results.

We corrected univariate p-values for multiple testing using Bonferroni correction based on the number of features tested across all univariate models. Variables were added to a multivariate model sequentially, ordered based on the p-value of the variables univariate model, and kept if the variable remained significant.

Among the possible interaction terms for the variables used in our models, we identified 4 interactions that could be reasonably hypothesized to impact prediction performance. These 4 terms were: data leakage through feature selection + training size, number of hyperparameter optimizations + training size, testing method + training size, and number of hyperparameter optimizations + testing method. We tested the potential interaction terms as individual additions to the multivariate model and kept the interactions and main effects of any interaction variables not already present in the final model if they remained significant after Bonferroni correction based on the number of interactions tested. We also performed sensitivity analyses removing studies with extreme values from models with the number of hyperparameter optimizations and training size variables. Sensitivity analysis results can be found in the Supplementary Results.

III. RESULTS

Among the 55 studies and 99 models included, 31 different disorders were modeled. The average AUC among all models was 0.75. Supplementary Fig. 5 shows a histogram of the frequency distribution of reported AUCs. Thirty-one percent (N = 31) of the models modeled eight autoimmune disorders and had a mean AUC of 0.80. Sixteen percent (N = 16) of the models...
modeled six psychiatric disorders and had a mean AUC of 0.73. The remaining 52 models modeled 16 disorders that did not fit into either of these groups and had a mean AUC of 0.73. The included studies and all data used in our models can be found in Supplementary Table I.

Of the models studied, 46% had methods with some degree of data leakage due to feature selection. Imputation and quality control procedures were reported in 46% and 91% of the models examined, respectively. Fifty-two percent of the models were non-linear while the remaining 48% were linear models. Cross-validation was used for testing in 65% of the models.

In univariate models the number of hyperparameter optimizations reported, data leakage due to feature selection, and model type were significantly associated with an increase in AUC in the test data after correcting for multiple testing (Table I). Modeling an autoimmune disorder was nominally associated with an increase in AUC but did not remain significant after correcting for multiple testing. Reporting on optimization, reporting on quality control procedures, reporting imputation usage, training size, testing method, and disorder heritability were not significant.

When interactions were entered one at a time, the interaction between data leakage due to feature selection and training size and the interaction between number of hyperparameter optimizations and training size were nominally associated with AUC (Table II). Only the interaction between data leakage due to feature selection and training size remained significant after Bonferroni correction for multiple testing.

Our final multivariate model included the number of hyperparameter optimizations, data leakage due to feature selection, autoimmune disorder modelling, training size, and the interaction between data leakage due to feature selection and training size (Table III). In this model, number of hyperparameter optimizations, data leakage due to feature selection, and modeling an autoimmune disorder were significantly associated with an increase in AUC after correcting for multiple testing. Based on the variable coefficients, our model estimated that the presence of data leakage due to feature selection leads to an overestimation of 0.23 in AUC. Models that classified autoimmune disorders saw an estimated AUC increase of 0.11 compared to all other disorders. The AUC predicted by the multivariate model had an R-squared of 0.58 in a regression model with the actual AUC values of the models examined.

### IV. DISCUSSION

Our analysis to determine which features of genomic ML studies lead to significant increases in AUC found evidence that, in many studies, methodological issues result in overreporting of model performance. Out of the studies investigated here, 44% had some form of data leakage due to feature selection. The most common form of data leakage due to feature selection was using the entire dataset for the GWAS and later splitting the dataset into multiple subsets or folds. Since the test data were used to

### TABLE I

| Feature                              | Coefficient | Std. Error | z     | P>|z|  
|--------------------------------------|-------------|------------|-------|-------|
| Number of optimizations              | 0.02        | 0.003      | 7.27  | <0.0001** |
| Data leakage: feature selection      | 0.17        | 0.05       | 3.60  | 0.0003** |
| Model type                           | 0.15        | 0.05       | 3.39  | 0.0007** |
| Autoimmune disorder                  | 0.06        | 0.03       | 2.35  | 0.019  |
| Reported any optimization            | 0.07        | 0.04       | 1.72  | 0.09   |
| Psychiatric disorder                 | -0.04       | 0.03       | -1.31 | 0.19   |
| Training size                        | -7 x 10^{-7}| 1 x 10^{-6} | -0.77 | 0.44   |
| Testing method                       | -0.03       | 0.05       | -0.52 | 0.60   |
| Reported imputation                  | -0.03       | 0.06       | -0.42 | 0.68   |
| Disorder heritability                | -0.04       | 0.1        | -0.33 | 0.74   |
| Reported quality control              | 0.03        | 0.07       | 0.38  | 0.71   |

*p<0.05 after multiple testing correction, **p<0.01 after multiple testing correction

### TABLE II

| Feature                              | z     | P>|z|  
|--------------------------------------|-------|-------|
| Data leakage: feature selection # training size | -2.50 | 0.01*  |
| Number of optimizations # training size  | -2.23 | 0.03   |
| Testing method # training size         | -0.48 | 0.63   |
| Testing method # number of optimizations| 0.18  | 0.86   |

*p<0.05 after Bonferroni multiple testing correction, **p<0.01 after Bonferroni multiple testing correction, # interaction
determine the loci that best separate cases from controls before ML modeling, even if researchers split the data into separate subsets during ML, information from the test data has already leaked into the process, which would lead to overestimates of accuracy. The coefficient of our multivariate model for this feature indicates that, on average, this form of data leakage increases the AUC by 0.23 after adjusting for other factors. For some applications, an increase in AUC of that magnitude would be enough to move models into a performance range that would falsely suggest clinical utility. Although this is the first demonstration of this problem for a collection of genomic ML studies, others have previously warned of the potential for this type of issue to interfere with results, studies, others have previously warned of the potential for this feature to interfere with studies.

| Feature                        | Coefficient | z    | \( P>|z| \) |
|-------------------------------|-------------|------|-------------|
| Number of optimizations       | 0.015       | 3.68 | <0.001**    |
| Data leakage: feature selection | 0.23        | 6.14 | <0.001**    |
| Autoimmune disorder           | 0.11        | 5.07 | <0.001**    |
| Training size                 | 1.74 x 10^{-6} | 1.56 | 0.12        |
| Data leakage: feature selection # | -1.9 x 10^{-5} | -2.50 | 0.01*       |

\( *p < 0.05 \) after Bonferroni multiple testing correction, \( **p < 0.01 \) after Bonferroni multiple testing correction, #: interaction

Table III: Multivariate Mixed-Effects Linear Model Results

The type of disorder studied was significantly associated with AUC in our final multivariate model after Bonferroni correction. After adjusting for other significant predictors of AUC, models of autoimmune disorders had significantly higher AUCs compared to both psychiatric disorders and all other disorders. We were surprised by the lack of association between disorder heritability and AUC since disorders with a larger genetic component should theoretically be more predictable in genetic models, but the significance of a disorder type may help explain disorder heritability’s lack of significance in our multivariate models. The hypothesis that disorder heritability may be associated with an increase in AUC assumes that the difficulty of extracting and modeling the genetic components is equivalent in all disorders. If the accessibility and ease of modelling genetic risk differ between disorders, we would no longer expect heritability to be associated with AUC. While this conclusion could previously be inferred by comparing the effect sizes of individual loci in GWASs of different disorders, our analysis provides further evidence that the genetic information from some groups of disorders may collectively differ in accessibility and ease of modelling. Differences in accessibility and ease of modelling could be due to differences in genetic architecture or differences in measurement (e.g., differences in misclassification rates). For example, the risk for Celiac disease and other autoimmune disorders is largely within a single region, whereas the genetic

**TABLE III**

Multivariate Mixed-Effects Linear Model Results

- **Feature**
- **Coefficient**
- **z**
- **\( P>|z| \)**

| Feature                        | Coefficient | z    | \( P>|z| \) |
|-------------------------------|-------------|------|-------------|
| Number of optimizations       | 0.015       | 3.68 | <0.001**    |
| Data leakage: feature selection | 0.23        | 6.14 | <0.001**    |
| Autoimmune disorder           | 0.11        | 5.07 | <0.001**    |
| Training size                 | 1.74 x 10^{-6} | 1.56 | 0.12        |
| Data leakage: feature selection # | -1.9 x 10^{-5} | -2.50 | 0.01*       |

\( *p < 0.05 \) after Bonferroni multiple testing correction, \( **p < 0.01 \) after Bonferroni multiple testing correction, #: interaction

BARNETT et al.: GENOMIC MACHINE LEARNING META-REGRESSION: INSIGHTS ON ASSOCIATIONS OF STUDY FEATURES
risk for most psychiatric disorders is spread thinly across the genome [10], [93]. Another possibility is that the autoimmune models were consistently different from the other models in some untested way. However, many of the autoimmune models were part of studies that also modeled non-autoimmune disorders, which makes this possibility less likely.

The model type variable was associated with increased AUC in a univariate model but was not associated in multivariate models. Some studies directly comparing different model types demonstrated that non-linear models usually had better performance [14], [48]. For example, Botta et al. showed an AUC improvement from 0.73 in their best linear model to 0.77 in their best non-linear model. Other studies found no difference when comparing linear and non-linear models [33], [41]. Similarly, in studies comparing linear and non-linear models to polygenic risk score models, where a single weighted sum of risk is calculated for each person, in some cases the more complex linear and non-linear models were more accurate [10], while in other cases the standard polygenic risk score had better performance [10], [38]. The relative performance of models with differing complexity may depend on many factors, such as disorder, sample size, the number of features being modeled, and hyperparameter optimization. As shown in Supplementary Table II, model type was correlated with both data leakage due to feature selection (0.28) and number of hyperparameter optimizations (0.44). These correlations may be responsible for the model type variables association in univariate models.

Our study has several limitations that may have limited our ability to detect associations between the included features and AUC. Like all meta-analytic regressions, we were only able to model study characteristics as they were reported within each study. In some studies, information about the test data may have been leaked in inadvertent and unreported ways that we could not examine. The reported AUCs from some models were better than the theoretical maximum AUCs for the disorders and nearly perfect, which could reflect label leakage [94]. Label leakage is a more direct form of data leakage where the model is given access to class labels, which the model can then use to make a prediction of those same labels. Label leakage is almost always accidental and unreported, so we were unable to study it. Supplementary Fig. 5 shows the peak caused by the near-perfect models in the frequency distribution of reported AUCs. Data leakage due to feature selection can also lead to near perfect classification and is likely responsible for the cluster of models with near perfect prediction [58]. Given the large effect size of data leakage due to feature selection on AUC, it is possible that this bimodal distribution reflects studies without severe data leakage in lower range of AUCs and studies with severe data leakage in the upper range of AUCs. Publication bias may be partially responsible for the observed distribution as well, since studies with inappropriate methods that lead to inflated results are more likely to be published compared to studies with inappropriate methods that lead to null results.

If any confounding study characteristics are present, it is possible that the real effect of included variables could differ from our results. For example, disorder heritability and training size, which we would expect to influence model performance but were not significant in our analysis, indicate the limitations of representing study characteristics in our analysis. Disorder heritability and training size may have had their true effect masked by unreported data leakage, inappropriate method choices within studies, or some other confounding factor. For example, if an inappropriate modeling strategy was selected, leading to a low AUC, that would make it more difficult to detect effects. Our clustered sandwich estimator usage limits but does not eliminate the possibility of an unreported study characteristic in a single study confounding our results. Alternatively, the variables we included could be acting in part as proxies for unreported or untested study characteristics. However, any potential proxy relationships may also have importance for understanding current practices and informing future guidelines.

The studies used in this analysis are heterogeneous due, in part, to the inclusion of any human disorder. These disorders likely have differing genetic complexities and optimal prediction performances, which may have limited our ability to detect differences in AUC based on the features of the study but also highlights the strength of the features that were detected despite this heterogeneity. We were unable to use specific disorders as a feature in our analysis due to the limited number of studies with each disorder. Excluding papers that did not use AUC as a performance metric limited the number of studies we could include in this analysis. Some relevant studies may have been excluded from our analysis because they did not match our search query. Our study was also limited to using what was reported in these studies, which may have limited our ability to fully assess data leakage and optimization. The data leakage features we used in this study were binary, but in practice different methods have different types and severities of data leakage which would be difficult to accurately represent in this type of analysis.

Our analysis of genomic machine learning studies has implications for defining best practices for genomic ML studies. Although such studies may eventually lead to clinically action-able risk calculators, publications that overestimate results will not be replicated, which could lead the field to prematurely abandon ML. We found data leakage due to feature selection to be significantly associated with AUC and a significant, negative interaction between data leakage and training size. The former leads to increased AUCs that likely overestimate the models’ true performance outside the training data while the latter limits or hides the expected positive effect of increasing training size. Perhaps our most important finding is that 46% of studies used feature selection methods that can lead to data leakage, which suggests a need to promote better machine learning practices in the field as is being done by the MLPsych consortium [95]. If genomic machine learning methods and results are to be improved and trusted, researchers must recognize and avoid these issues. Thorough best practice guidelines that promote the avoidance of data leakage and other common issues will be critical as the field grows and advances.

Acknowledgment

This project has received funding from the European Union’s Horizon 2020 research and innovation programme grant agreement No 965381.
Financial Disclosures: In the past year, Dr. Faraone received income, potential income, travel expenses continuing education support and/or research support from Aardvark, Akili, Genomind, Ironshore, KenPharm/Corium, Noven, Ondosis, Osuka, Rhodes, Supernus, Takeda, Tris and Vallon. With his institution, he has US patent US20130217707 A1 for the use of sodium-hydrogen exchange inhibitors in the treatment of ADHD. In previous years, he received support from: Alcorba, Arbor, Aveksham, CogCubed, Eli Lilly, Enzymotec, Impact, Janssen, Lundbeck/Takeda, McNeil, NeuroLifeSciences, NovoRance, Novartis, Pfizer, Shire, and Sunovion. He also receives royalties from books published by Guilford Press: Straight Talk About Your Child’s Mental Health; Oxford University Press: Schizophrenia: The Facts; and Elsevier: ADHD: Non-Pharmacologic Interventions. He is also Program Director of Nature, vol. 20, no. 8, pp. 467–484, 2019.

Data Availability: All data produced in the present study and the code used in our analyses are in the online supplement.

REFERENCES

[1] N. R. Wray et al., “Research review: Polygenic methods and their application to psychiatric traits,” J. Child Psychol. Psychiatry, vol. 55, no. 10, pp. 1068–1087, 2014.

[2] N. R. Wray et al., “Pinfalls of predicting complex traits from SNPs,” Nature Rev. Genet., vol. 14, no. 7, pp. 507–515, 2013.

[3] S. Whalen et al., “Navigating the pitfalls of applying machine learning in genomics,” Nat. Rev. Genet., vol. 23, no. 3, pp. 169–181, 2022.

[4] M. W. Libbrecht and W. S. Noble, “Machine learning applications in genetics and genomics,” Nat. Rev. Genet., vol. 16, no. 6, pp. 321–332, 2015.

[5] V. Tam et al., “Benefits and limitations of genome-wide association studies,” Nat. Rev. Genet., vol. 20, no. 8, pp. 467–484, 2019.

[6] Psychiatric Genomics Consortium, “Biological insights from 108 schizophrenia-associated genetic loci,” Nature, vol. 511, no. 7510, pp. 421–427, 2014.

[7] X. Yang, “An overview of overfitting and its solutions,” in IOP Conference Series: Journal of Physics: Conference Series. Bristol, U.K.: IOP Publishing, 2019.

[8] G. C. Cawley and N. L. C. Talbot, “On over-fitting in model selection and subsequent selection bias in performance evaluation,” J. Mach. Learn. Res., vol. 11, pp. 2079–2107, 2010.

[9] S. Kaufman et al., “Leakage in data mining: Formulation, detection, and avoidance,” ACM Trans. Knowl. Discov. Data, vol. 6, no. 4, 2012, Art. no. 12.

[10] G. Abraham et al., “Utilizing deep learning and genome wide association studies for epistatic-driven preterm birth classification in African-American women,” IEEE/ACM Trans. Comput. Biol. Bioinform, vol. 17, no. 2, pp. 668–678, 2020.

[11] J. Gaudilio et al., “Machine learning approach to single nucleotide polymorphism-based asthma prediction,” PLoS One, vol. 14, no. 12, 2019, Art. no. e0225574.

[12] Y. Guo et al., “Machine learning derived risk prediction of anorexia nervosa,” BMC Med. Genomic., vol. 9, 2016, Art. no. 4.

[13] J. Kang et al., “Improved risk prediction for Crohn’s disease with a multi-focus approach,” Hum. Mol. Genet., vol. 20, no. 12, pp. 2435–2442, 2011.

[14] S. Kinreich et al., “Predicting risk for Alcohol Use Disorder using longitudinal data with multimodal biomarkers and family history: A machine learning study,” Mol. Psychiatry, vol. 26, no. 4, pp. 1133–1141, 2021.

[15] C. Kooperberg, M. LeBlanc, and V. Obenchain, “Risk prediction using genome-wide association studies,” Genet Epidemiol., vol. 34, no. 7, pp. 454–652, 2010.

[16] N. Krautenbacher et al., “Asthma in farm children is more determined by genetic polymorphisms and in non-farm children by environmental factors,” Pediatr. Allergy Immunol., vol. 32, no. 2, pp. 295–304, 2021.

[17] S. Lee et al., “Machine learning on genome-wide association studies to predict the risk of radiation-associated contralateral breast cancer in the WECARE Study,” PLoS One, vol. 15, no. 2, 2020, Art. no. e0226157.

[18] L. Liu et al., “Deep learning model reveals potential risk genes for ADHD, especially Ephrin receptor gene EPH5A5,” Brief. Bioinf., vol. 22, 2021, Art. no. bbab207.

[19] F. Mittag et al., “Use of support vector machines for disease risk prediction in genome-wide association studies: Concerns and opportunities,” Hum. Mutat., vol. 33, no. 12, pp. 1708–1718, 2012.

[20] M. Muneeb and A. Henschel, “Eye-color and Type-2 diabetes phenotype prediction from genotype data using deep learning methods,” BMC Bioinf., vol. 22, no. 1, 2021, Art. no. 198.

[21] T. T. Nguyen et al., “Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests,” BMC Genomic., vol. 2, no. Suppl 2, 2015, Art. no. S5.

[22] L. R. Pal et al., “CAGH4 Crohn’s exome challenge: Marker SNP versus exome variant models for assigning risk of Crohn disease,” Hum. Mutat., vol. 38, no. 9, pp. 1225–1234, 2017.

[23] G. Paré, S. Mao, and W. Q. Deng, “A machine-learning heuristic to improve gene score prediction of polygenic traits,” Sci. Rep., vol. 7, no. 1, 2017, Art. no. 12665.

[24] L. Pirozzi et al., “Data mining approaches for genome-wide association of mood disorders,” Psychiatr. Genet., vol. 22, no. 2, pp. 55–61, 2012.

[25] A. Romagnoni et al., “Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data,” Sci. Rep., vol. 9, no. 1, 2019, Art. no. 10351.

[26] B. L. Romero-Rosales et al., “Improving predictive models for Alzheimer’s disease using GWAS data by incorporating misclassified samples modeling,” PLoS One, vol. 15, no. 4, 2020, Art. no. e0232103.

[27] C. Sinoquet, “A method combining a random forest-based technique with the modeling of linkage disequilibrium through latent variables, to run multilocus genome-wide association studies,” BMC Bioinf., vol. 19, no. 1, 2018, Art. no. 106.

[28] E. Skalidas et al., “Predicting the diagnosis of autism spectrum disorder using gene pathway analysis,” Mol. Psychiatry, vol. 19, no. 4, pp. 504–510, 2014.

[29] Y. V. Sun et al., “Classification of rheumatoid arthritis status with candidate gene and genome-wide single-nucleotide polymorphisms using random forests,” BMC Proc., vol. 1, no. Suppl 1, 2007, Art. no. S62.

[30] M. Thomas et al., “Genome-wide modeling of polygenic risk score in colorectal cancer risk,” JMIR Med. Inform., vol. 10, no. 3, pp. 432–444, 2020.

[31] H. Wang and P. Avillich, “Diagnostic classification and prognostic prediction using common genetic variants in autism spectrum disorder: Genotype-based deep learning,” JIMR Med. Inform., vol. 9, no. 4, 2021, Art. no. e24754.

[32] H. Y. Wang et al., “Machine learning-based method for obesity risk evaluation using single-nucleotide polymorphisms derived from next-generation sequencing,” J. Comput. Biol., vol. 25, no. 12, pp. 1347–1360, 2018.
Eric J. Barnett received the BS degree in biology and chemistry from Guilford College, and the PhD degree from SUNY Upstate Medical University. He is a post-doctoral associate with the Department of Psychiatry, SUNY Upstate Medical University in Syracuse, NY, USA. He received the World Federation of ADHD Young Scientist Award in 2021.

Daniel G. Onete received the bachelor of Medical Sciences with an Honors Specialization in physiology from the University of Western Ontario in Canada, and the master of Public Health from Queen’s University. He is a graduate from the Norton College of Medicine, SUNY Upstate Medical University, in Syracuse, NY, USA. He is a member of the American Society of Anesthesiology.

Asif Salekin received the PhD degree. He is an assistant professor in computer science from Syracuse University, NY, USA, where he directs the Ubiquitous and Intelligent Sensing (UIS Lab). He is also affiliated with Aging Studies Institute, Syracuse University, and an assistant professor (voluntary track) with SUNY Upstate Medical University, Syracuse, NY, USA.

Stephen V. Faraone received the PhD degree. He is a distinguished professor and vice chair for research with the Department of Psychiatry, Norton College of Medicine, SUNY Upstate Medical University. He is the president of the World Federation of ADHD and a leading expert in ADHD research.