2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871234

# Adapting Random Forests to Predict Obesity-Associated Gene Expression

Jeremy Watts[1], Elexis Allen[2], Ahmad Mitoubsi[1], Anahita Khojandi[1] *Member, IEEE,* James Eales[3], Farideh Jalali-Najafabadi[4], and Theodore Papamarkou[5,6]

***Abstract*— Random forests (RFs) are effective at predicting gene expression from genotype data. However, a comparison of RF regressors and classifiers, including feature selection and encoding, has been under-explored in the context of gene expression prediction. Specifically, we examine the role of ordinal or one-hot encoding and of data balancing via oversampling in the prediction of obesity-associated gene expression. Our work shows that RFs compete with PrediXcan in the prediction of obesity-associated gene expression in subcutaneous adipose tissue, a highly relevant tissue to obesity. Additionally, RFs generate predictions for obesity-associated genes where PrediXcan fails to do so.**

## I. Introduction

Large-scale expression quantitative trait loci (eQTL) analyses have demonstrated that most genes have a genetic component to their expression [1]. By combining genotypes from multiple variants proximal to a gene, we can build models that can predict tissue-specific expression of a gene at the patient level. PrediXcan is one such genotype-based prediction method that can predict gene expression from individual-level genotype data [2], [3]. While there is empirical evidence that PrediXcan predicts the expression of a wide range of genes well, this is not universally the case for all genes [4], [5].

Machine learning approaches, such as random forests (RFs), have several desirable traits for bioinformatics, including the clear identification of important features, scalability, and robustness against noise [6], [7]. Consequently, RFs have been explored to predict gene expression [8], [9]. While these studies have shown great promise, comparative studies between RF regressors and classifiers have been carried out less extensively in the genetics literature.

The literature on how machine learning approaches generalize to many medically relevant genes is limited. One genetic association with far-reaching impact is obesity. Obesity is clinically linked to increased morbidity and mortality [10], [11]. Genetic effects on monogenic obesity have been well documented [10] and recent GWAS studies have discovered many novel single nucleotide polymorphisms (SNPs) associated with non-syndromic obesity [10], [11], [12].

In this work, we examine the capability of RFs to predict the gene expression of key obesity genes based on genetic variants. Specifically, each genetic variant is used as a potential feature in our models. Further, we compare these RFs to PrediXcan, a state-of-the-art gene expression prediction method [2]. PrediXcan uses predetermined weights, often from PredictDB (https://predictdb.org), making it a popular choice amongst researchers [13], [14]. We apply RFs and PrediXcan to subcutaneous adipose tissue samples from the Genotype-Tissue Expression (GTEx) project [1], [15].

## II. Data

In this study, we utilize the GTEx dataset [1], [15], which contains samples of up to 43 highly genotyped tissue types from more than 700 post-mortem donors. The GTEx database includes eQTL data, which represent gene-regulatory expression across multiple tissues [1], [15]. In this study, we examine the normalized expression and genetic variant data in subcutaneous adipose tissue to predict the expression of key obesity genes for 581 donors. The data were analyzed following GTEx's Data Use Certification Agreement. Ethical approval for GTEx was originally obtained from each study site's Institutional Review Board [1].

The role of genetics in obesity has been well-documented leading to the identification of many impactful genes [10], [11], [12]. The Centers for Disease Control (CDC) published an article identifying key genes associated with obesity [16]. These genes include *ADIPOQ* (ENSG00000181092), *FTO* (ENSG00000140718), *INSIG2* (ENSG00000125629), *LEP* (ENSG00000174697), *LEPR* (ENSG00000116678), *PCSK1* (ENSG00000175426), and *PPARG* (ENSG00000132170).

## III. Methods

We apply five RFs (ordinal or one-hot encoded regressors and classifiers and a balanced classifier) to predict gene expression using different feature selection and encoding

*The research is partially supported by NHGRI's Genomic Data Science Analysis, Visualization, and Informatics Lab-space's (AnVIL) Cloud Credits (AC2) pilot program.

*The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from dbGaP accession number phs000424.v8.p2 on 09/07/2021.

[1]Jeremy Watts, Anahita Khojandi, and Ahmad Mitoubsi are with the Department of Industrial Engineering, University of Tennessee, Knoxville, TN, USA (email: gkm819@vols.utk.edu, khojandi@utk.edu).

[2]Elexis Allen is with the Department of Biomedical Engineering, University of Tennessee, Knoxville, TN, USA.

[3]James Eales is with the Division of Cardiovascular Sciences, University of Manchester, Manchester, UK.

[4]Farideh Jalali-Najafabadi is with the Centre for Genetics and Genomics Versus Arthritis, the Centre for Musculoskeletal Research, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK.

[5]Theodore Papamarkou is with the Department of Mathematics, University of Manchester, Manchester, UK.

[6]Theodore Papamarkou is also with the Department of Mathematics, University of Tennessee, Knoxville, TN, USA.

techniques. RFs are supervised machine learning models that use an ensemble of decision trees, and are generally shown to be very robust when applied to high-dimensional data, and not prone to overfitting [7]. Therefore, RFs are well-positioned to predict gene expression from genetic variants, which frequently entail more than $30,000$ features.

We apply RFs to two different versions of the GTEx dataset for comparison with PrediXcan. First, we perform 10-fold cross-validation (CV) on GTEx version 8. Second, each RF is separately trained on GTEx version 7 and then tested on the donors present in GTEx version 8 but not in GTEx version 7. The latter process is visualized in Figure 1. Note that we use this process to ensure that we avoid any potential overlap between training and testing data during the analysis. Each RF was implemented using scikit-learn's RF libraries in Python [17]. A similar training and testing approach is followed for PrediXcan; weights from PrediXcan defined for GTEx version 7 are used for testing PrediXcan on donors available only in GTEx version 8. This way, PrediXcan's testing and training data do not overlap.

### A. Random Forest Regressor and Classifier

In this study, we compare the predictive performance of RF regressors and classifiers in predicting obesity-associated gene expression. Gene expression is a normalized continuous variable, typically ranging from $-3$ to 3, which justifies the use of an RF regressor.

Much of clinical value in predicting gene expression comes from stratifying individuals' expression levels [18]. Therefore, an RF classifier may prove beneficial at predicting expression values at the extreme ends of the distribution. To apply an RF classifier, we discretize the continuous variable of gene expression. Specifically, we define five classes that group gene expression values. First, we define a class within one standard deviation (SD) from the mean, then between one and two SDs away from the mean (in both positive and negative directions), and finally beyond two SDs away from the mean (in both positive and negative directions). Once an RF predicts a class based on the genetic variants, we map the predicted class to a continuous value at the center of the corresponding class to compute $R^2$.

### B. Ordinal and One-hot Encoding

The genetic variant data from the GTEx project are provided in the VCF file format. We compare two different encoding methods. First, we apply the ordinal encoding mapping $.|.$ (missing data) $\rightarrow$ 0, $0|0$ (no alt alleles) $\rightarrow$ 1, $0|1$ and $1|0$ (one alt allele) $\rightarrow$ 2, $1|1$ (two alt alleles) $\rightarrow$ 3. Such encoding assigns higher importance to SNPs with more copies of the alternate allele.

As an alternative approach, one-hot encoding represents each variant as a binary vector. This encoding method is often a more efficient way to parse and store data. To accomplish this encoding, we use Keras' one-hot encoding library [19].

### C. Oversampling Approach for Data Balancing

One challenge in training a classification model is learning from unbalanced data [20]. This is the case in our constructed classes, where the expression values have been pre-normalized by GTEx. One way of ameliorating the class imbalance problem is to apply a domain-specific data oversampling technique in which noise is added to the sampled data. Similar 'oversampling with noise' techniques have been employed in the task of ordinal classification [20]. In this study, we randomly sample a genetic profile from an 'under-represented' class. We then randomly alter the ordinal encoding of 100 genetic variants from this profile, acting as noise in the sampling process. This profile maintains the same class representation as before the noise generation process. The oversampling process is repeated for each under-represented class.

## IV. Results

We compare the predictive performance (as determined through $R^2$, a standard metric in gene expression prediction studies [2], [3]) of five RFs and of PrediXcan at predicting gene expression. We first perform preliminary experiments to tune the RFs' hyper-parameters. Specifically, we determine the optimal number of decision trees by iteratively adding trees from 100 to 1500 with a step size of 100 trees. Further, we compare the optimal number of included features by keeping the most important features ranging from 500 to the maximum number of genetic variants for each gene (approximately $30,000$) with a step size of 500 features, as each gene may be influenced by a unique number of variants. The corresponding CV results on GTEx version 8 are presented in Figures 2 and 3, respectively. In general, these results indicate that increasing the number of decision trees does not dramatically impact predictive performance. Similarly, a relatively low number of features achieves overall comparative performance. Based on these results, in each of the following experiments, we set the number of decision trees to 100, and we perform feature pruning from 25 to 1000 features with a step size of 25 features.

### A. GTEx Version 8

We compare the five implemented RFs using 10-fold CV to predict gene expression in GTEx version 8 subcutaneous adipose tissue. These results are presented in Table I. The RFs are benchmarked against PrediXcan's elastic net and

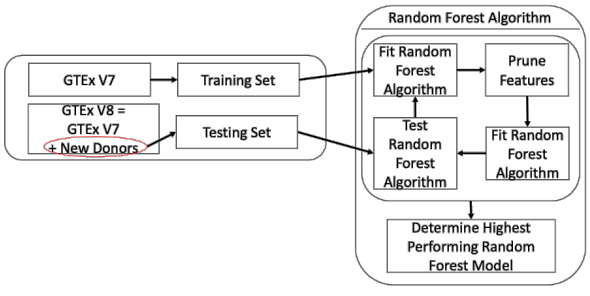


Fig. 1. Fitting a random forest (RF) to the GTEx data. The RF is trained on GTEx version 7 and tested on the donors present in GTEx version 8 but not in version 7. The same training/testing separation is performed for PrediXcan using the publicly available weights defined for GTEx version 7.

| Gene name | PrediXcan benchmark | | Random forest regressor | | Random forest classifier | | |
|---|---|---|---|---|---|---|---|
| | Elastic net | MASHR | Ordinal | One-hot | Ordinal | One-hot | Balanced |
| ADIPOQ | 0.030 (0.041) | 0.050 (0.066) | 0.031 (0.030) | 0.023 (0.018) | 0.034 (0.028) | 0.029 (0.039) | 0.028 (0.023) |
| FTO | **0.050 (0.035)** | 0.017 (0.017) | 0.020 (0.023) | 0.035 (0.040) | 0.022 (0.019) | 0.026 (0.032) | 0.036 (0.038) |
| INSIG2 | **0.080 (0.048)** | **0.060 (0.044)** | 0.035 (0.024) | 0.029 (0.034) | 0.025 (0.046) | 0.047 (0.061) | 0.033 (0.030) |
| LEP | - | 0.011 (0.007) | 0.026 (0.037) | 0.021 (0.027) | 0.039 (0.064) | 0.030 (0.045) | **0.040 (0.045)** |
| LEPR | 0.035 (0.022) | 0.016 (0.017) | **0.045 (0.041)** | 0.025 (0.029) | 0.028 (0.026) | 0.028 (0.030) | 0.036 (0.031) |
| PCSK1 | 0.052 (0.062) | 0.038 (0.051) | 0.036 (0.056) | 0.037 (0.060) | 0.031 (0.033) | 0.034 (0.042) | 0.045 (0.036) |
| PPARG | - | 0.028 (0.021) | 0.037 (0.053) | 0.020 (0.021) | 0.034 (0.047) | 0.041 (0.055) | 0.039 (0.035) |

TABLE I

MEAN $R^2$ VALUES OF RANDOM FORESTS AND OF PREDIXCAN BENCHMARKS RELATED TO THE PREDICTION OF OBESITY-ASSOCIATED GENE EXPRESSION (STANDARD DEVIATIONS ARE SHOWN IN PARENTHESES). EACH $R^2$ IS AVERAGED OVER 10-FOLD CV ON GTEX VERSION 8. MODELS WHICH EXHIBIT IMPROVEMENT IN GENE-SPECIFIC PREDICTIVE PERFORMANCE AT A SIGNIFICANCE LEVEL OF 10% ARE PRESENTED IN BOLD (SEE SECTION IV-A). SIGNIFICANCE IS DETERMINED VIA PAIRED T-TESTS IN THE DIFFERENCE OF $R^2$ VALUES BETWEEN MODEL PAIRS.

| Gene name | PrediXcan benchmark | Random forest regressor | | Random forest classifier | | |
|---|---|---|---|---|---|---|
| | Elastic net - V7 | Ordinal | One-hot | Ordinal | One-hot | Balanced |
| ADIPOQ | (0.000, 0.052) | (0.014, 0.179) | (0.004, 0.102) | (0.003, 0.143) | (0.004, 0.174) | (0.000, 0.075) |
| FTO | (0.000, 0.080) | (0.027, 0.257) | (0.011, 0.204) | (0.000, 0.161) | (0.000, 0.139) | (0.000, 0.087) |
| INSIG2 | (0.035, 0.194) | (0.001, 0.102) | (0.003, 0.181) | (0.014, 0.160) | (0.026, 0.258) | (0.000, 0.012) |
| LEP | - | (0.000, 0.097) | (0.001, 0.083) | (0.014, 0.197) | (0.016, 0.201) | (0.000, 0.000) |
| LEPR | (0.001, 0.111) | (0.000, 0.079) | (0.006, 0.148) | (0.004, 0.169) | (0.002, 0.146) | (0.009, 0.161) |
| PCSK1 | (0.000, 0.073) | (0.000, 0.052) | (0.016, 0.170) | (0.006, 0.149) | (0.000, 0.192) | (0.000, 0.000) |
| PPARG | - | (0.002, 0.126) | (0.012, 0.110) | (0.001, 0.129) | (0.000, 0.090) | (0.067, 0.279) |

TABLE II

90% BIAS-CORRECTED AND ACCELERATED (BCA) BOOTSTRAP CONFIDENCE INTERVALS FOR $R^2$ VALUES OF RANDOM FORESTS AND OF PREDIXCAN RELATED TO THE PREDICTION OF OBESITY-ASSOCIATED GENE EXPRESSION. THE TESTING DATA ARE EXTRACTED FROM GTEX VERSION 8, CORRESPONDING TO DONORS NOT IN GTEX VERSION 7. EACH BCA CONFIDENCE INTERVAL IS ESTIMATED FROM 1,000 BOOTSTRAP SAMPLES FROM THE TESTING SET. MODELS WHICH EXHIBIT IMPROVEMENT IN GENE-SPECIFIC PREDICTIVE PERFORMANCE AT A SIGNIFICANCE LEVEL OF 10% ARE PRESENTED IN BOLD (SEE SECTION IV-B). PREDIXCAN'S ELASTIC NET MODEL DOES NOT HAVE PREDETERMINED WEIGHTS FOR *LEP* AND *PPARG*.

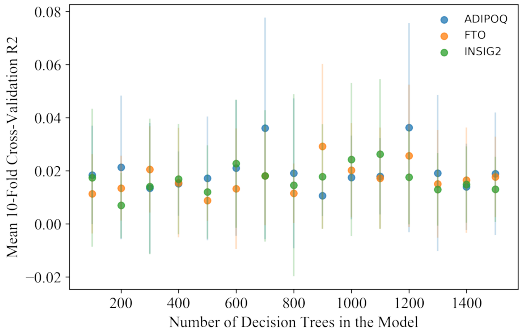


Fig. 2. Sample of the performance of the random forest regressor for incremental amounts of decision trees for *ADIPOQ*, *FTO*, and *INSIG2*. Each point represents the mean $R^2$ averaged over 10-fold CV, where the error bars represent the standard deviation.

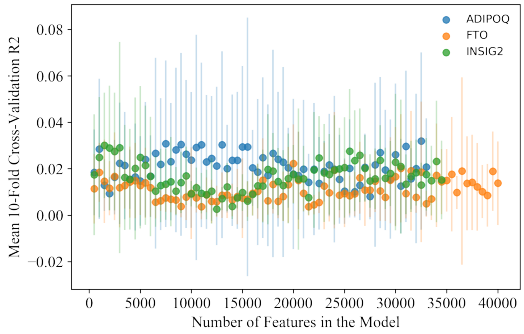


Fig. 3. Sample of the performance of the random forest regressor for incremental amounts of features for *ADIPOQ*, *FTO*, and *INSIG2*. Each point represents the mean $R^2$ averaged over 10-fold CV, where the error bars represent the standard deviation.

MASHR models, which use predetermined weights previously trained on the entirety of GTEx version 8 [13]. Note that the PrediXcan's elastic net model does not have publicly available predetermined weights for the genes *LEP* and *PPARG* in subcutaneous adipose tissue.

These results show that there is no universal winner between RFs and PrediXcan's methods based on paired $t$-tests on the differences in $R^2$ between associated model pairs at a 10% level of statistical significance. For *LEPR*, the ordinal RF regressor and the balanced classifier outperform PrediXcan's MASHR model; for *LEP*, the balanced RF classifier outperforms PrediXcan's MASHR model; for *FTO*, PrediXcan's elastic net outperforms PrediXcan's MASHR and the ordinal RF regressor and classifier; for *INSIG2*, PrediXcan's elastic net outperforms PrediXcan's MASHR model and all RFs. Finally, for *LEPR* the ordinal RF regressor outperform both the one-hot encoded regressor and classifier.

### B. GTEx Version 7

We again compare the five implemented RFs and PrediXcan's methods in terms of gene expression predictions on the testing data. As discussed in Section III and depicted in Figure 1, to avoid overlap between training and testing data, we use PrediXcan's predetermined weights from GTEx version 7 to predict expression for a testing set that only contains the new donors added in GTEx version 8. Similarly, the RFs are trained on GTEx version 7 and are tested on the donors added in GTEx version 8 [14]. We draw $1,000$ bootstrap samples from the testing set, each with sample size equal to the sample size of the testing set. For each of the $1,000$ bootstrap samples, we compute the $R^2$ of each model. We then determine bias-corrected and accelerated (BCa) confidence intervals for $R^2$ over the bootstrap samples from the testing sets as a range of predictive performance [21]. These confidence intervals are presented in Table II. To utilize this approach, we translate variants from human genome reference build b37 to b38.

We compute 90% BCa confidence intervals for the differences in $R^2$ between model pairs as a metric of statistical significance. The predictive performance of RFs is statistically similar to PrediXcan at a significance level of 10%, that is based on the 90% BCa confidence intervals for the differences in $R^2$. Moreover, at a significance level of 10% the balanced RF classifier is outperformed on *INSIG2* (by PrediXcan, and the one-hot RF classifier), *LEP* (by ordinal and one-hot RF classifiers), and *PCSK1* (by ordinal and one-hot RF classifiers).

## V. Discussion and Future Work

In this study, we compare RF regressors and classifiers in the prediction of obesity-associated gene expression. The RFs are benchmarked against PrediXcan's state-of-the-art genotype-based prediction algorithms. Our results show, that in general the RFs perform comparably and, in some cases, outperform PrediXcan. Finally, we find that ordinal and one-hot encoding perform similarly for the majority of genes.

This work is subject to certain limitations. Due to the pre-trained nature of PrediXcan, it is not trivial to make a fair and direct comparison. Using the GTEx version 8 dataset, PrediXcan gains an advantage from having previously been trained on the entirety of GTEx, meaning it has had exposure to the testing set. Although GTEx presents one of the most deeply sampled genetic populations, testing on a completely novel/separate obesity genetic dataset would be ideal for benchmarking.

RFs have shown to be effective in predicting gene expression for select gene/tissue pairs [8], [9]. This work expands the literature to encompass key obesity genes while additionally exploring the predictive merits of regressors, classifiers, encoding, and the balancing of training sets. To this end, future work may include developing an RF that trims features (genetic variants) based on their positional proximity relatively to one another, for a biologically-inspired approach to parameter pruning.

## References


[1] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, *et al.*, "The genotype-tissue expression (gtex) project," *Nature genetics*, vol. 45, no. 6, pp. 580–585, 2013.

[2] E. R. Gamazon, H. E. Wheeler, K. P. Shah, S. V. Mozaffari, K. Aquino-Michaels, R. J. Carroll, A. E. Eyler, J. C. Denny, D. L. Nicolae, N. J. Cox, *et al.*, "A gene-based association method for mapping traits using reference transcriptome data," *Nature genetics*, vol. 47, no. 9, pp. 1091–1098, 2015.

[3] A. N. Barbeira, S. P. Dickinson, R. Bonazzola, J. Zheng, H. E. Wheeler, J. M. Torres, E. S. Torstenson, K. P. Shah, T. Garcia, T. L. Edwards, *et al.*, "Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics," *Nature communications*, vol. 9, no. 1, pp. 1–20, 2018.

[4] M. Wainberg, N. Sinnott-Armstrong, N. Mancuso, A. N. Barbeira, D. A. Knowles, D. Golan, R. Ermel, A. Ruusalepp, T. Quertermous, K. Hao, *et al.*, "Opportunities and challenges for transcriptome-wide association studies," *Nature genetics*, vol. 51, no. 4, pp. 592–599, 2019.

[5] A. V. Mikhaylova and T. A. Thornton, "Accuracy of gene expression prediction from genotype data with PrediXcan varies across and within continental populations," *Frontiers in Genetics*, vol. 10, p. 261, 2019.

[6] Y. Qi, "Random forest for bioinformatics," in *Ensemble machine learning*, pp. 307–323, Springer, 2012.

[7] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[8] R. Díaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, vol. 7, no. 1, pp. 1–13, 2006.

[9] Y. Kong and T. Yu, "A deep neural network model using random forest to extract feature representation for gene expression data classification," *Scientific reports*, vol. 8, no. 1, pp. 1–9, 2018.

[10] A. J. Walley, J. E. Asher, and P. Froguel, "The genetic contribution to non-syndromic human obesity," *Nature Reviews Genetics*, vol. 10, no. 7, pp. 431–442, 2009.

[11] H. Choquet and D. Meyre, "Genetics of obesity: what have we learned?," *Current genomics*, vol. 12, no. 3, pp. 169–179, 2011.

[12] B. M. Herrera, S. Keildson, and C. M. Lindgren, "Genetics and epigenetics of obesity," *Maturitas*, vol. 69, no. 1, pp. 41–49, 2011.

[13] PredictDB Team, "GTEx v8 models on eQTL and sQTL." `https://predictdb.org/post/2021/07/21/gtex-v8-models-on-eqtl-and-sqtl/`, 2021.

[14] PedictDB Team, "GTEx v7 expression models." `https://predictdb.org/post/2017/11/29/gtex-v7-expression-models/`, 2017.

[15] GTEx Consortium, K. G. Ardlie, D. S. Deluca, A. V. Segrè, T. J. Sullivan, T. R. Young, E. T. Gelfand, C. A. Trowbridge, J. B. Maller, T. Tukiainen, *et al.*, "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans," *Science*, vol. 348, no. 6235, pp. 648–660, 2015.

[16] Centers for Disease Control and Prevention, "Genes and obesity." `https://www.cdc.gov/genomics/resources/diseases/obesity/obesedit.htm`. Last accessed 1/4/22.

[17] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

[18] T. G. Richardson, K. O'Nunain, C. L. Relton, and G. D. Smith, "Harnessing whole genome polygenic risk scores to stratify individuals based on cardiometabolic risk factors and biomarkers at age 10 in the lifecourse," *Arteriosclerosis, Thrombosis, and Vascular Biology*, pp. ATVBAHA–121, 2021.

[19] F. Chollet, "Keras." `https://github.com/fchollet/keras`, 2015.

[20] I. Domingues, J. P. Amorim, P. H. Abreu, H. Duarte, and J. Santos, "Evaluation of oversampling data balancing techniques in the context of ordinal classification," in *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2018.

[21] T. J. DiCiccio and B. Efron, "Bootstrap confidence intervals," *Statistical science*, vol. 11, no. 3, pp. 189–228, 1996.