Published online 27 July 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 3 **1**
https://doi.org/10.1093/nargab/lqab069

Careful feature selection is key in classification of Alzheimer's disease patients based on whole-genome sequencing data

Marlena Osipowicz, Bartek Wilczynski, Magdalena A. Machnicka* and for the Alzheimer's Disease Neuroimaging Initiative†

Institute of Informatics, Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, 02-097, Poland

Received February 09, 2021; Revised July 06, 2021; Editorial Decision July 13, 2021; Accepted July 20, 2021

ABSTRACT

Despite great increase of the amount of data from genome-wide association studies (GWAS) and whole-genome sequencing (WGS), the genetic background of a partially heritable Alzheimer's disease (AD) is not fully understood yet. Machine learning methods are expected to help researchers in the analysis of the large number of SNPs possibly associated with the disease onset. To date, a number of such approaches were applied to genotype-based classification of AD patients and healthy controls using GWAS data and reported accuracy of 0.65–0.975. However, since the estimated influence of genotype on sporadic AD occurrence is lower than that, these very high classification accuracies may potentially be a result of overfitting. We have explored the possibilities of applying feature selection and classification using random forests to WGS and GWAS data from two datasets. Our results suggest that this approach is prone to overfitting if feature selection is performed before division of data into the training and testing set. Therefore, we recommend avoiding selection of features used to build the model based on data included in the testing set. We suggest that for currently available dataset sizes the expected classifier performance is between 0.55 and 0.7 (AUC) and higher accuracies reported in literature are likely a result of overfitting.

INTRODUCTION

Alzheimer's disease (AD) is a neurodegenerative disorder for which both genetic background and environmental factors have been shown to influence risk. Some forms of the disease are caused by rare mutations in amyloid precursor protein (*APP*), presenilin 1 (*PSEN1*) and presenilin 2 (*PSEN2*) genes and inherited in an autosomal dominant manner. However, this heritable (familial) form of AD constitute only around 5% of all cases and only 5–10% of its occurrences can be explained by the presence of known mutations in *APP*, *PSEN1* and *PSEN2* genes (1). The great majority of AD occurrences represent the so called sporadic AD. The genetic background of the sporadic AD is no longer monogenic and its estimated contribution to the disease risk ranges from 58% to 79% (2). For many years the only known gene associated with AD risk was *APOE*. The *APOE* ε4 allele increases the disease risk and modifies the onset age (3,4). Until now around 30 different loci associated with AD risk have been identified, mainly by the use of genome-wide association studies (GWAS) (5–7).

Several machine learning approaches have been already applied to the problem of genotype-based classification of AD patients and healthy controls and reported classification accuracies range from 0.65 to 0.975 (8–13). However, since the influence of genotype on sporadic AD occurrence estimated in a twin study does not exceed 90% (2), higher accuracy of disease status prediction (classification of subjects into cases and controls) solely based on genetic data is not expected. There is, therefore, a possibility that very high accuracy of classification results from overfitting of the machine learning model. One possible reason for overfitting may be the presence of data used for model testing in the training set. We have observed that the highest AUC results are reported for studies in which initial pre-screening of SNPs was performed to preferentially include those that would be expected to correlate with the outcome. For example Briones and Dinu (9) apply logistic regression to the complete AD GWAS data set (prior to division into training and testing set) to pre-select SNPs associated with the

*To whom correspondence should be addressed. Tel: +48 22 55 44 583; Email: m.machnicka@mimuw.edu.pl

†Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

disease. A random forest classifier built based on these pre-selected SNPs reaches an average 10-fold cross-validation (CV) error of around 10%. In the approach by Nguyen *et al*. (11) random forest-based feature selection algorithm was applied in combination with random forest classifier also to an AD GWAS dataset. Similarly to the previous approach, the informative SNPs used to build the classifier were chosen from the complete dataset. The maximal AUC estimated from 5-fold CV was 0.975. Finally, application of allelic, genomic and regression tests to pre-select disease associated SNPs again based on complete GWAS data set and multifactor dimensionality reduction in 10-fold CV resulted in classification accuracy of 0.78 (12).

On the other hand Bayesian-network based models applied to AD GWAS dataset in 5-fold CV but with models build on training data only reach AUC around 0.7 (10). In a different approach genes known to be associated with AD based on meta-analysis results provided by the AlzGene database were used to pre-select SNPs from ADNI GWAS data set (8). In this case selected features were known to correlate with the outcome but not in the same data set as the one used for model building. The random forests for controls versus AD classification built on selected SNPs were characterized by out-of-bag error of around 0.42 to 0.52. Finally, when label propagation method was used to rank SNPs from a training subset of AD GWAS data set the maximal AUC of a *k*-nearest neighbor classification in 5-fold CV was around 0.75 (13).

It should be noted that the problem of patient classification based on genotype data is widely studied in literature, not only in the context of AD. In most of these contexts, some feature selection needs to be performed, as the number of features is usually much greater than the number of patients, which is an issue known in statistical machine learning as p>>n. Since the methods of feature selection meant for dealing with this problem are relatively new and complex, they require care in application to avoid overfitting.

We have thus decided to explore the possibilities of applying machine-learning classification algorithms to single-nucleotide polymorphisms (SNPs) from whole-genome sequencing (WGS) data that provide approximately 40 times more features (sites of genomic variation) per sample. We applied Boruta feature selection algorithm (14) and random forest classifiers to two WGS and one GWAS datasets obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) and the Religious Orders Study and the Rush Memory and Aging Project (ROSMAP, www.radc.rush.edu) (15,16) and investigated the influence of machine learning study design and structure of analyzed sample on classification results.

MATERIALS AND METHODS

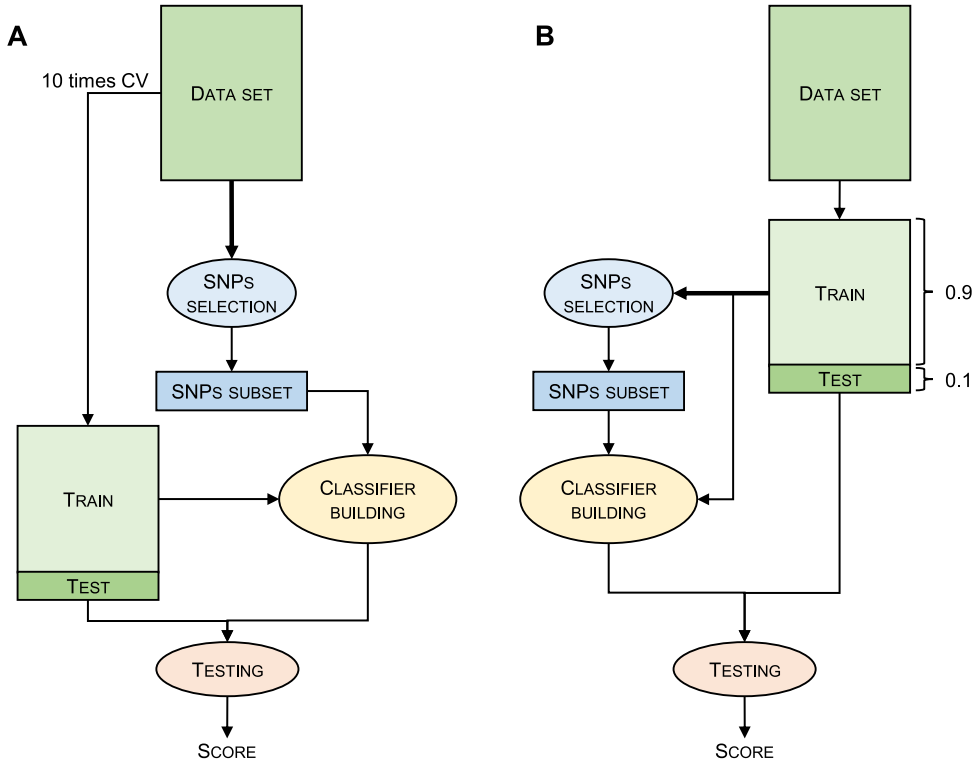
Datasets

Whole genome sequencing (WGS) data were obtained from two resources: the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) and the Religious Orders Study and the Rush Memory and Aging Project (ROSMAP, www.radc.rush.edu) (15,16). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org. SNP genotyping (GWAS) data were also obtained from ADNI.

The ADNI-WGS dataset includes results of whole-genome sequencing and variant calling for 808 patients (251 healthy, 235 AD and 322 with mild cognitive impairment, MCI). The ADNI-GWAS dataset includes SNP genotyping results for 793 patients (266 healthy, 219 AD and 308 MCI). MCI patients were not included in the analysis. The diagnosis for each patient was established based on 'DXSUM_PDXCONV_ADNIALL.csv' file that provides summary of diagnostic information for ADNI patients. Reference genome for ADNI dataset is hg19 (UCSC Genome Browser standard). The ROSMAP dataset consists three sub-datasets (Mayo, MSBB and Rosmap), which in total include results of whole-genome sequencing and variant calling for 1894 patients (503 healthy, 530 AD, 861 others). Only healthy and AD patients were included in the analysis. Diagnoses for participants from Mayo dataset were obtained from the 'MayoRNAseq_RNAseq_GenomeWide_Genotypes_Covariates.csv' file, diagnoses for MSBB dataset were obtained from 'MSBB_RNAseq.WES.WGS_sample_QC_info.csv' and 'MSBB_clinical.csv' files and diagnoses for Rosmap dataset were obtained from 'AMP-AD_rosmap_WGS_id_key.csv' and 'ROSMAP_clinical.csv' files. Reference genome for ROSMAP dataset is GRCh37 (Genome Reference Consortium standard).

The following WGS data curation steps were performed prior to the analysis: (i) SNPs were selected from vcf files using the SelectVariants method from GATK toolkit (version 4.0.11.0) (17), (ii) patients genotypes were encoded as 0 (homozygous reference), 1 or 2 or 3 (heterozygous or homozygous alternative, for different possible alternative alleles: 1 for first alternative allele, 2 for second and 3 for third) or -1 (missing data). No filtering based on the minor allele frequency was performed. SNPs common for ADNI-WGS and ROSMAP datasets were identified based on genomic positions. Population structure for ADNI-WGS and ROSMAP dataset was analyzed by PCA conducted with PLINK 1.9 software (18) using 1000 genomes data (19) as reference. Plots for the first four principal components are available as Supplementary Figure S1.

GWAS data were downloaded in the binary plink format. After its decompression using the PLINK 1.9 software (18) MAP file with information about SNPs and PED file with patients genotypes were obtained. Next, reference allele for each SNP was retrieved from the dbSNP b152 database (20). Then based on the PED files and reference allele values patients genotypes were encoded in the same way as for WGS data. No filtering based on missing data amount has been performed, leading to the following percentages of SNPs having > 5% of missing data in the analyzed datasets: ADNI GWAS – 7.5%, ADNI WGS – 26.7%, ROSMAP WGS – 24%.



C

Data set	#SNPs total	#SNPs shared	#Cases	#Controls
ADNI WGS	38,443,567	22,364,035	235	251
ROSMAP	51,511,212		530	503
ADNI GWAS	712,473	-	219	266

Figure 1. Analysis summary. (**A**) Feature (SNPs) selection before division of the data into training and test sets. (**B**) Feature selection after division of the data into training and test sets. (**C**) Datasets included in the study.

Feature selection

The python implementation of the Boruta feature selection algorithm (https://github.com/scikit-learn-contrib/boruta_py) was used to select a subset of SNPs expected to be relevant predictors of the disease status. Boruta algorithm (14) evaluates features importance using random forests and shadow features (copies of the real features with values shuffled across objects). During each iteration the algorithm calculates importance for every feature (the real and the shadow ones) and using a binomial test assesses which of the real features have significantly higher predictive value in the classification result than some chosen percentage of shadow features (90% in this analysis). Due to memory requirements feature selection was performed in batches containing 5000 consecutive SNPs. There were 7700 batches in the ADNI WGS and 11 695 batches in the ROSMAP dataset. Computation time for each batch was 3.61 min for ADNI and 3.86 min for ROSMAP. Relevant features selected from each batch were merged to generate the final set of selected features. Feature selection performed in batches required 1.9GB RAM for ADNI WGS and 5GB RAM for ROSMAP.

The feature selection step was performed either on the complete dataset (before division of data into training and test set, Figure 1A) or on the training set only (Figure 1B).

Classification

For classification and evaluation of classifier performance we used the scikit-learn package, version 0.20.1 (21). Classification was conducted with the random forest algorithm (sklearn.ensemble.RandomForestClassifier). In the first part of the analysis, where feature selection was conducted before division of the data into training and test sets (Figure 1A), we conducted 10-fold cross validation (sklearn.model_selection.Kfold). In the second part of the analysis, in which feature selection was conducted after division of the data into training and test sets (Figure 1B), the scores were estimated from the mean of results from three iterations of the whole algorithm in which the datasets were randomly split into training (90% of data) and test (10% of data) sets. The area under the ROC curve (AUC) calculated for the test data only was used as the evaluation metric (sklearn.metrics.roc_auc_score).

Code used to perform the presented analysis is available at https://github.com/regulomics/alzheimer_classification.

Selection of patients subsets based on genetic similarity

Each patient's genotype was expressed as a vector of 0, 1, 2, 3 and -1 values (as described in the Datasets section). Then the Hamming distance for each pair of pa-

Table 1. Numbers of selected features and classifiers performance for analysis with feature selection before test/train division. Mean values and standard deviations from a 10-fold CV are given

ADNI WGS		ROSMAP		ADNI GWAS	
AUC	#selected SNPs	AUC	#selected SNPs	AUC	#selected SNPs
0.99 ± 0.02	341 690	0.99 ± 0.01	369 012	0.98 ± 0.03	24 998

tients was computed (the difference between patients for each variant site was clipped to 0–1 range using numpy.clip method). Based on these distances UPGMA clustering was conducted using SciPy package, version 1.1.0 (function scipy.cluster.hierarchy.linkage with parameter method = 'average') and the largest cluster was selected (at dist > 0.1055 for ADNI-WGS and 0.096 for ROSMAP) and samples with almost identical genotypes (most likely family members at dist < 0.097 for ADNI-WGS and 0.093 for ROSMAP) were removed from it.

Functional annotation of relevant SNPs

Relevant SNPs common for ADNI-WGS and ROSMAP-WGS datasets were chosen using *intersect* tool from the BedTools toolkit (22). Selected SNPs were assigned to genes using Variant Effect Predictor via online interface (23) and gene list was submitted to the PANTHER Overrepresentation Test (Released 20190417, Annotation Version and Release Date: PANTHER version 14.1 Released 12 March 2019) (24). Overrepresentation of GO-Slim Biological Process terms was calculated with Fisher's Exact test.

RESULTS

Feature selection before division into training and testing set results in severe overfitting

Using feature selection can lead to surprisingly high rates of prediction accuracy. By performing feature selection on the complete ADNI-WGS, ADNI-GWAS and ROSMAP datasets we have selected approximately 350 thousand relevant SNPs for each WGS dataset and 25 thousand for GWAS dataset (Table 1). These selected SNPs have been used to build random forest classifiers within a 10-fold cross-validation procedure (Figure 1A). The classifiers performance expressed as the area under the ROC curve (AUC) was around 0.98 for all three datasets (Table 1). That high classification accuracy is unexpected since there is a non-negligible environmental influence on the onset of AD even when identical twins are considered.

Models trained with globally selected features show no apparent utility between patient cohorts. Next, we performed an analysis where we used random forest models built on one data set to classify patients from the second dataset. Since ADNI-WGS and ROSMAP data are characterized by a substantial difference in the total number of SNPs (Figure 1C), we have conducted this analysis only on SNPs shared between ADNI-WGS and ROSMAP datasets. Feature selection in these experiments was again performed on complete data sets, with the starting SNPs positions set restricted to SNPs shared between ADNI-WGS and ROSMAP. The classification performance in this setup was significantly lower than when the analysis was performed on one dataset (Table 2). This result suggests that the models were overfitted to the datasets from which the features have been selected and carry little information about global associations between SNPs and disease status.

Table 2. Numbers of selected features and classifiers performance for analysis with feature selection before test/train division on SNPs shared between ADNI-WGS and ROSMAP-WGS datasets. Mean values and standard deviations from a 10-fold CV are given

Test set	ADNI WGS		ROSMAP	
Training set	AUC	#selected SNPs	AUC	#selected SNPs
ADNI WGS	0.99 ± 0.01	257 634	0.50 ± 0.02	257 634
ROSMAP	0.50 ± 0.04	185 252	0.99 ± 0.01	185 252

Using proper precautions in feature selection leads to models with lower performance, but much more reproducible between datasets. Subsequently, we have investigated if performing feature selection on the complete dataset (before division of data into training and test set) could have influenced the performance of classifiers built on selected SNPs. To do so we have built classifiers on features selected from the training set only (Figure 1B). AUC for these classifiers ranged from 0.56 ± 0.02 for ADNI-WGS to 0.67 ± 0.06 for ADNI-GWAS (Table 3). This suggests that performing feature selection on the complete dataset may result in overfitting of the classifier due to inclusion of test set data in the feature selection procedure.

The performances of classifiers obtained for SNPs shared between ADNI-WGS and ROSMAP datasets were comparable to the results obtained for complete datasets (Table 3) suggesting that these values are not particularly dependent on patient population and that for other datasets with comparable sample and SNP number (500–1000 samples and ∼20 million of SNP) we would expect similar results. On the other hand, classifier performance on ADNI-GWAS dataset, which contains approximately 40 times fewer SNPs than ADNI-WGS, was significantly better than performance of any of the classifiers built based on WGS data (Table 3), suggesting that indeed pre-selection of SNPs to a smaller group of variants seems to be helpful in classification. It is difficult to verify whether this is simply due to the lower number of variables, or indeed there were some selection criteria in the design of the SNP arrays that are helping in the classification accuracy.

Population structure influences classification results

We have analyzed the genetic similarity between patients from the ADNI-WGS and ROSMAP datasets and found that some individuals are either closely related (high pairwise similarity) or originate from a distinct population (low similarity). We have thus repeated the analysis on the subset of patients which are neither too closely nor too distantly related (for the detailed description of selection criteria see the Materials and Methods section): 444 for the ADNI-WGS dataset (222 cases and 222 controls) and 809

Table 3. Numbers of selected features and classifiers performance for analysis with feature selection after test/train division. Mean values and standard deviations from three repetitions are given

	ADNI WGS		ROSMAP		ADNI GWAS	
SNPs set	AUC	#selected SNPs	AUC	#selected SNPs	AUC	#selected SNPs
All from each dataset	0.56 ± 0.02	334 886 ± 15 762	0.58 ± 0.02	358 181 ± 4640	0.67 ± 0.06	24 610 ± 667
Common for ADNI WGS and ROSMAP WGS	0.51 ± 0.09	258 758 ± 13 288	0.54 ± 0.04	182 348 ± 5622	-	-

Table 4. Numbers of selected features and classifiers performance for analysis with feature selection after test/train division performed on ROSMAP and ADNI-WGS subsets of patients chosen based on genetic similarity. Mean values and standard deviations were calculated from two and four repetitions for ADNI-WGS and ROSMAP, respectively

ADNI WGS		ROSMAP	
AUC	#selected SNPs	AUC	#selected SNPs
0.63 ± 0.03	201 484	0.55 ± 0.09	160 992 ± 211

for the ROSMAP dataset (426 cases and 383 controls). This resulted in a small increase of the classifiers performance (Table 4).

Model built for the classification of one dataset carries information relevant for the second dataset

To assess whether information used to classify one dataset can be successfully used for classification of a completely different patient set, we have conducted two types of analyses. First, we have performed feature selection and random forest building on the training set selected from the first dataset (e.g. 90% of patients from ADNI-WGS) and used the resulting random forest to classify the patients from the other dataset (e.g. ROSMAP). In the second approach, we have again selected relevant features based on the training subset of the first dataset and then used the locations of the chosen SNPs to build a new random forest model on the data on SNP variants and diagnose labels from patients from the second dataset. In both cases, we assessed the performance of the resulting classifier on the patients from the second dataset that were not used in training of the tested classifier. The AUC values for both approaches dropped down slightly compared to the results on one dataset but they were still well above 0.5 suggesting that all the models carry some universal information about genetic differences between AD cases and controls that did not depend on the patient sub-population of the ROSMAP and ADNI cohorts. The results are presented in Table 5.

Selected relevant SNPs reflect known aspects of AD biology

We have identified 10 176 relevant SNPs common to ADNI-WGS and ROSMAP by intersecting feature selection results for these two datasets for one of the three repetitions of the analysis. These SNPs were assigned to 5075 genes for which the PANTHER Overrepresentation Test for GO-Slim Biological Process terms was calculated. Significantly overrepresented terms (FDR < 0.05) are listed in Table 6. They show that identified SNPs represent functional terms related with synapse formation and function. Synaptic loss

Table 5. Cross classification based on features chosen and/or random forest built on different dataset, after train/test division (mean values and standard deviations from two repetitions)

Analysis type	Training set	Test set	AUC	#selected SNPs
SNPs selected and random forest built on training set	ADNI-WGS	ROSMAP	0.55 ± 0.01	243 799
	ROSMAP	ADNI-WGS	0.53 ± 0.04	183 528
SNPs selected on training set, random forest built on test set	ADNI-WGS	ROSMAP	0.55 ± 0.03	243 799
	ROSMAP	ADNI-WGS	0.57 ± 0.01	183 528

is one of the major features of AD and is believed to be correlated with cognitive impairment (25,26). Several terms related with formation of cell–cell junctions, including adherens junctions, are also overrepresented. Since this type of cell junctions is expected to be present between endothelial cells forming the blood–brain barrier this finding is in agreement with known blood-brain barrier dysfunctions in neurodegenerative disorders (reviewed in (27)).

DISCUSSION

We have studied the potential of applying random forest classifiers with Boruta feature selection for building predictive models for identifying patients likely to develop AD based on their genotype. This has been done before, on different datasets and with varying level of accuracy. Importantly, some studies have reported prediction accuracy significantly exceeding the expected level of possible accuracy given the fact that even among identical twins, there should be a non-negligible environmental contribution to the onset of AD. Since most of the previously published methods used some form of feature pre-selection, usually based on overall SNP association with the class variable, we expected that this might be the underlying reason behind the probable overfitted models.

Using two different strategies to apply feature selection, we were able to reproduce the overfitting effect on two well-known datasets provided by the ADNI and ROSMAP projects. In both cases we were able to create classifiers giving seemingly perfect prediction scores, even though they were properly cross-validated (except the feature selection stage). The culprit is in the fact that the feature selection was done on the whole dataset, and only selected features that we verified to perform well on the whole sample set, giving the classifier an artificially easy task. The fact that the classifiers were overfitted is corroborated by the results of the

Table 6. Gene Ontology terms overrepresentation for SNPs selected from ADNI-WGS and ROSMAP-WGS

PANTHER GO-Slim Biological Process	# in *Homo sapiens* - REFLIST (20996)	#	expected	Fold Enrichment	raw *P* value	FDR
Cell–cell adhesion via plasma-membrane adhesion molecules (GO:0098742)	52	29	7.24	4.01	3.77E-08	7.53E-06
Modulation of chemical synaptic transmission (GO:0050804)	75	32	10.44	3.07	8.79E-07	1.13E-04
Adherens junction organization (GO:0034332)	23	16	3.2	5	5.55E-06	4.99E-04
Cell–cell junction assembly (GO:0007043)	28	17	3.9	4.36	1.07E-05	8.35E-04
Synaptic transmission, glutamatergic (GO:0035249)	44	19	6.12	3.1	1.24E-04	5.85E-03
Actin cytoskeleton organization (GO:0030036)	179	47	24.91	1.89	2.16E-04	9.45E-03
Cellular calcium ion homeostasis (GO:0006874)	151	41	21.01	1.95	3.39E-04	1.45E-02
Cell morphogenesis involved in neuron differentiation (GO:0048667)	110	32	15.31	2.09	6.83E-04	2.51E-02
Second-messenger-mediated signaling (GO:0019932)	191	47	26.58	1.77	9.26E-04	3.08E-02
Synapse assembly (GO:0007416)	15	9	2.09	4.31	1.40E-03	4.11E-02

attempt of predicting AD in patients from one study based on the seemingly perfect classifier trained on the other one. This approach led to a drastic reduction in predictive power (AUC 0.99 lowered to AUC ∼0.5).

When we repeated the training experiments with the same feature selection methodology, but now performed only on the training set, we saw a dramatically different picture. In this case, we were able to predict the patients outcome on a much lower level (AUC ∼0.6), but the prediction accuracy remained relatively unchanged when we switched to the other patients' set. These results indicate that indeed, combining the random forest classifier with Boruta feature selection can lead to learning valuable information about genetic causes of AD, regardless of patient subpopulation. We were also very impressed by the fact that the functional annotation of genes carrying mutations used by the trained classifiers to predict AD cases, were significantly enriched in many functional categories (such as neuronal development) lending further confidence in our results.

Given that the overall prediction quality on the SNP data is rather low, and we do not expect it to increase in other datasets of comparable size, we have experimented with selecting sub-populations from our datasets, to train classifiers tuned to more genetically homogenous populations. This approach seems to be promising, as the results on the sub-samples of patients were slightly higher than on the whole.

The main conclusions of this work should not depend on the choice of the specific classification algorithm or quality-control filtering of the input dataset. Based on our experience other modern classical machine learning methods, for example logistic regression and linear discriminant analysis, are expected to give similar results as random forests. We also verified that removal of variants and samples with very low call rates and variants with extreme departures from Hardy–Weinberg equilibrium from the input datasets does not change the results significantly (data not shown).

For late-onset diseases, such as AD, the signal for association of genetic factors with the diagnosis can be quenched by wrong assignment of disease status to some of the control subjects who can convert to cases at a later point. To account for this phenomenon the controls should be at least age-matched or even older than cases. The use of centenarians as 'extreme controls' can strengthen the association signals as it was shown for AD and type II diabetes (28,29).

For diseases with complex, polygenic genetic background such as AD, polygenic risk score (PRS) analysis is widely used for assessment of genetic risk for an individual taking effects of many genetic markers into account (reviewed in (30)). The reported prediction accuracies of logistic regression models based on PRS for sporadic forms of AD are typically described by AUC ∼0.6–0.65 (31–33) and are thus similar to prediction accuracies reported in this study. The most prominent difference between our approach and PRS analysis is in the number of selected markers, ranging from ∼20 to ∼200 000 in the PRS analysis and from 10 000 to 350 000 in feature selection with the Boruta algorithm, depending on significance thresholds used. Specific comparison of prediction accuracies obtained for ADNI WGS in our study and by Leonenko *et al*. (33) suggests that our approach does not lead to higher prediction accuracies than PRS analysis. However, as pointed by Escott-Price *et al*. (34), PRS analysis exhibits similar susceptibility to overfitting effects when test data are used during the selection of markers as we describe here for random forest based models. This shows that our main conclusions have broad significance for the field of genotype-based prediction of phenotypes regardless of the type of statistical model used.

CONCLUSION

We have showed that overfitting the data due to premature feature selection is likely the reason behind surprisingly high performance of some published attempts at classification of AD patients based on their genotype. Any approach using full dataset for feature selection is giving an unfair advantage to the training process and artificially inflating the estimate of accuracy based on cross-validation that is done after feature selection. We also show that there is a way to combine random forest classification with feature selection that avoids this problem and allows us to obtain classifiers with comparable performance on unrelated patients subsets. The prediction accuracy we obtain using this scheme is relatively low (AUC ∼0.6 for WGS SNP data and AUC ∼0.7 for GWAS SNP data); however, it is likely to be the limit possible to obtain on the datasets with the currently typical sizes of ∼1000 patients and controls. We expect these limits to increase as the number of patient samples is significantly increased (35,36) (at least to tens of thousands); however, we do not expect this to reach 0.9, based on the

estimates of limited contribution of the genetic background to the AD risk and non-negligible role of environmental factors. However, we see some promise in training specialized classifiers for curated sub-populations of patients, where maybe even for smaller numbers of samples we can get the increased accuracy. Overall, we expect that with all the large-scale genomic sequencing efforts underway in many developed countries, the cohort sizes will increase to tens or hundreds of thousands of patients allowing for the model prediction accuracy to reach the levels of expected heritability of AD. Additionally, as the sequencing efforts are now collecting more and more of the environmental metadata, we might be constructing next-generation models that exceed the quality of classifiers based solely on the genomic information.

CODE REPOSITORY

Code used to perform the presented analysis is available at: https://github.com/regulomics/alzheimer_classification

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

For ADNI written informed consent was obtained from all participants at the time of enrollment for imaging and genetic sample collection and protocols of consent forms were approved by each participating sites' Institutional Review Board (IRB).

All participants of the ROSMAP study signed the informed consent and Anatomical Gift Act upon enrollment. Both studies were approved by the institutional review board of Rush University Medical Center.

This study was approved by the Research Ethics Committee of the Faculty of Mathematics, Informatics and Mechanics, University of Warsaw.

CONSENT FOR PUBLICATION

Not applicable.

DATA AVAILABILITY

The datasets analyzed during the current study are available in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) and the Religious Orders Study and the Rush Memory and Aging Project (ROSMAP, www.radc.rush.edu).

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

This study analyses data collected within the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Religious Orders Study and the Rush Memory and Aging Project (ROSMAP). ROSMAP data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161, R01AG15819, R01AG17917, and U01AG46152. Data collection and sharing for ADNI dataset was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Authors' contributions: M.O. performed data preprocessing, feature selection and classification analysis and participated in results interpretation. B.W. participated in results interpretation and study design. M.M. participated in results interpretation and study design and performed gene ontology analysis. All authors contributed in writing the manuscript and read and approved the final manuscript.

FUNDING

Fundacja na rzecz Nauki Polskiej [POWROTY/REINTEGRATION, POIR.04.04.00-00-3E86/17-00], Polish National Science Center [DEC-2015/16/W/NZ2/00314].

Conflict of interest statement. None declared.

REFERENCES

1. Van Cauwenberghe,C., Van Broeckhoven,C. and Sleegers,K. (2016) The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genet. Med.*, **18**, 421–430.
2. Gatz,M., Reynolds,C.A., Fratiglioni,L., Johansson,B., Mortimer,J.A., Berg,S., Fiske,A. and Pedersen,N.L. (2006) Role of genes and environments for explaining Alzheimer disease. *Arch. Gen. Psychiatry*, **63**, 168–174.
3. Corder,E.H., Saunders,A.M., Strittmatter,W.J., Schmechel,D.E., Gaskell,P.C., Small,G.W., Roses,A.D., Haines,J.L. and Pericak-Vance,M.A. (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, **261**, 921–923.
4. Saunders,A.M., Strittmatter,W.J., Schmechel,D., St. George-Hyslop,P.H., Pericak-Vance,M.A., Joo,S.H., Rosi,B.L.,

Gusella,J.F., Crapper-MacLachlan,D.R., Alberts,M.J. *et al.* (2012) Association of apolipoprotein E allele 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology*, **43**, 1467–1472.
5. Jansen,I.E., Savage,J.E., Watanabe,K., Bryois,J., Williams,D.M., Steinberg,S., Sealock,J., Karlsson,I.K., Hägg,S., Athanasiu,L. *et al.* (2019) Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.*, **51**, 404–413.
6. Lambert,J.-C., Ibrahim-Verbaas,C.A., Harold,D., Naj,A.C., Sims,R., Bellenguez,C., Jun,G., DeStefano,A.L., Bis,J.C., Beecham,G.W. *et al.* (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.*, **45**, 1452–1458.
7. Kunkle,B.W., Grenier-Boley,B., Sims,R., Bis,J.C., Damotte,V., Naj,A.C., Boland,A., Vronskaya,M., van der Lee,S.J., Amlie-Wolf,A. *et al.* (2019) Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.*, **51**, 414–430.
8. Araújo,G.S., Souza,M.R.B., Oliveira,J.R.M. and Costa,I.G. (2013) Random forest and gene networks for association of SNPs to Alzheimer's disease. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, **8213**, 104–115.
9. Briones,N. and Dinu,V. (2012) Data mining of high density genomic variant data for prediction of Alzheimer's disease risk. *BMC Med. Genet.*, **13**, 7.
10. Jiang,X., Cai,B., Xue,D., Lu,X., Cooper,G.F. and Neapolitan,R.E. (2014) A comparative analysis of methods for predicting clinical outcomes using high-dimensional genomic datasets. *J. Am. Med. Informatics Assoc.*, **21**, e312–e319.
11. Nguyen,T.-T., Huang,J., Wu,Q., Nguyen,T. and Li,M. (2015) Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. *BMC Genomics*, **16**, S5.
12. Sherif,F.F., Zayed,N., Fakhr,M., Wahed,M.A. and Kadah,Y.M. (2017) Integrated higher-order evidence-based framework for prediction of higher-order epistasis interactions in Alzheimer's disease. **11**, 16–24.
13. Stokes,M.E., Barmada,M.M., Kamboh,M.I. and Visweswaran,S. (2014) The application of network label propagation to rank biomarkers in genome-wide Alzheimer's data. *BMC Genomics*, **15**, 282.
14. Kursa,M.B. and Rudnicki,W.R. (2010) Feature selection with the boruta package. *J. Stat. Softw.*, **36**, 1–13.
15. Bennett,D.A., Buchman,A.S., Boyle,P.A., Barnes,L.L., Wilson,R.S. and Schneider,J.A. (2018) Religious orders study and rush memory and aging project. *J. Alzheimer's Dis.*, **64**, S161–S189.
16. De Jager,P.L., Ma,Y., McCabe,C., Xu,J., Vardarajan,B.N., Felsky,D., Klein,H.U., White,C.C., Peters,M.A., Lodgson,B. *et al.* (2018) Data descriptor: A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Sci. Data*, **5**, 180142.
17. Van der Auwera,G.A., Carneiro,M.O., Hartl,C., Poplin,R., del Angel,G., Levy-Moonshine,A., Jordan,T., Shakir,K., Roazen,D., Thibault,J. *et al.* (2013) From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. **43**, 11.10.1–11.10.33.
18. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A.R., Bender,D., Maller,J., Sklar,P., de Bakker,P.I.W., Daly,M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
19. 1000 Genomes Project Consortium, Auton,A, Brooks,LD, Durbin,RM, Garrison,EP, Kang,HM, Korbel,JO, Marchini,JL, McCarthy,S, McVean,GA and Abecasis,G.R. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
20. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, **29**, 308–311.
21. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
22. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
23. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R.S., Thormann,A., Flicek,P. and Cunningham,F. (2016) The ensembl variant effect predictor. *Genome Biol.*, **17**, 122.
24. Mi,H., Muruganujan,A., Ebert,D., Huang,X. and Thomas,P.D. (2019) PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.*, **47**, D419–D426.
25. Terry,R.D., Masliah,E., Salmon,D.P., Butters,N., DeTeresa,R., Hill,R., Hansen,L.A. and Katzman,R. (1991) Physical basis of cognitive alterations in alzheimer's disease: Synapse loss is the major correlate of cognitive impairment. *Ann. Neurol.*, **30**, 572–80.
26. DeKosky,S.T. and Scheff,S.W. (1990) Synapse loss in frontal cortex biopsies in Alzheimer's disease: correlation with cognitive severity. *Ann. Neurol.*, **27**, 457–464.
27. Sweeney,M.D., Sagare,A.P. and Zlokovic,B.V. (2018) Blood-brain barrier breakdown in Alzheimer disease and other neurodegenerative disorders. *Nat. Rev. Neurol.*, **14**, 133–150.
28. Garagnani,P., Giuliani,C., Pirazzini,C., Olivieri,F., Bacalini,M.G., Ostan,R., Mari,D., Passarino,G., Monti,D., Bonfigli,A.R. *et al.* (2013) Centenarians as super-controls to assess the biological relevance of genetic risk factors for common age-related diseases: a proof of principle on type 2 diabetes. *Aging (Albany. NY).*, **5**, 373–385.
29. Tesi,N., van der Lee,S.J., Hulsman,M., Jansen,I.E., Stringa,N., van Schoor,N., Meijers-Heijboer,H., Huisman,M., Scheltens,P., Reinders,M.J.T. *et al.* (2019) Centenarian controls increase variant effect sizes by an average twofold in an extreme case–extreme control analysis of Alzheimer's disease. *Eur. J. Hum. Genet.*, **27**, 244–253.
30. Chasioti,D., Yan,J., Nho,K. and Saykin,A.J. (2019) Progress in polygenic composite scores in Alzheimer's and other complex diseases. *Trends Genet.*, **35**, 371–382.
31. Escott-Price,V., Sims,R., Bannister,C., Harold,D., Vronskaya,M., Majounie,E., Badarinarayan,N., Morgan,K., Passmore,P., Holmes,C. *et al.* (2015) Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain*, **138**, 3673–3684.
32. Cruchaga,C., Del-Aguila,J.L., Saef,B., Black,K., Fernandez,M.V., Budde,J., Ibanez,L., Deming,Y., Kapoor,M., Tosto,G. *et al.* (2018) Polygenic risk score of sporadic late-onset Alzheimer's disease reveals a shared architecture with the familial and early-onset forms. *Alzheimer' Dement.*, **14**, 205–214.
33. Leonenko,G., Baker,E., Stevenson-Hoare,J., Sierksma,A., Fiers,M., Williams,J., De Strooper,B. and Escott-Price,V. (2021) Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores is most accurate when using all genetic information. Research Square doi: https://doi.org/10.21203/rs.3.rs-137252/v1, 18 January 2021, preprint: not peer reviewed.
34. Escott-Price,V., Myers,A.J., Huentelman,M. and Hardy,J. (2017) Polygenic risk score analysis of pathologically confirmed Alzheimer disease. *Ann. Neurol.*, **82**, 311–314.
35. Figueroa,R.L., Zeng-Treitler,Q., Kandula,S. and Ngo,L.H. (2012) Predicting sample size required for classification performance. *BMC Med. Inform. Decis. Mak.*, **12**, 8.
36. Sordo,M. and Zeng,Q. (2005) On Sample Size and Classification Accuracy: A Performance Comparison. In: Oliveira,J.L., Maojo,V., Martín-Sánchez,F. and Pereira,A.S. (eds). *Biological and Medical Data Analysis. ISBMDA 2005. Lecture Notes in Computer Science*. Vol. **3745**, Springer, Berlin, Heidelberg, pp. 193–201.