Predictive genetic panel for adult asthma using machine learning methods

Luciano Gama da Silva Gomes, PhD,[a] Álvaro Augusto Souza da Cruz, MD, PhD,[b] Maria Borges Rabêlo de Santana, PhD,[a] Gabriela Pimentel Pinheiro, PhD,[b] Cinthia Vila Nova Santana, PhD,[b] Carolina Barbosa Souza Santos, PhD,[b] Meher Preethi Boorgula, MS,[c] Monica Campbell, MS,[c] Adelmir de Souza Machado, MD, PhD,[a,b] Rafael Valente Veiga, PhD,[d] Kathleen C. Barnes, PhD,[c] Ryan dos Santos Costa, PhD,[a] and Camila Alexandrina Figueiredo, PhD[a] *Bahia, Brazil; Aurora, Colo; and Cambridge, United Kingdom*

Background: Asthma is a chronic inflammatory disease of the airways that is heterogeneous and multifactorial, making its accurate characterization a complex process. Therefore, identifying the genetic variations associated with asthma and discovering the molecular interactions between the omics that confer risk of developing this disease will help us to unravel the biological pathways involved in its pathogenesis.
Objective: We sought to develop a predictive genetic panel for asthma using machine learning methods.
Methods: We tested 3 variable selection methods: Boruta's algorithm, the top 200 genome-wide association study markers according to their respective *P* values, and an elastic net regression. Ten different algorithms were chosen for the classification tests. A predictive panel was built on the basis of joint scores between the classification algorithms.
Results: Two variable selection methods, Boruta and genome-wide association studies, were statistically similar in terms of the average accuracies generated, whereas elastic net had the worst overall performance. The predictive genetic panel was completed with 155 single-nucleotide variants, with 91.18% accuracy, 92.75% sensitivity, and 89.55% specificity using the support vector machine algorithm. The markers used range from known single-nucleotide variants to those not previously described in the literature. Our study shows potential in creating genetic prediction panels with tailored penalties per marker, aiding in the identification of optimal machine learning methods for intricate results.
Conclusions: This method is able to classify asthma and nonasthma effectively, proving its potential utility in clinical prediction and diagnosis. (J Allergy Clin Immunol Global 2024;3:100282.)

Key words: *Asthma, genetics, machine learning, single-nucleotide variants, prediction*

Asthma is usually an inflammatory condition of the airways that results in immune hyperreactivity.[1] This disease is heterogeneous and defined by a history of characteristics such as airway inflammation, smooth-muscle contraction, epithelial sloughing, mucous hypersecretion, bronchial hyperresponsiveness, and mucosal edema.[1]

The model of asthma as a single entity is no longer accepted, becoming a much more complex immune network of distinct and interrelated inflammatory pathways.[2,3]

Both genetics and environment contribute to asthma risk and interact in complex ways to influence asthma endotypes and immune processes.[4] Various studies have used strategies such as genomics, transcriptomics, and proteomics to understand this complexity.[5] Identifying genetic variations associated with asthma and discovering the molecular interactions between the omics that confer risk of developing this disease will help us to unravel the biological pathways involved in the pathogenesis of asthma, resulting in improved treatment.[6,7]

Genome-wide association studies (GWASs), for example, are an approach that use microarrays of single-nucleotide variant (SNV) chips.[8] The goal is to identify DNA variations associated with asthma or its characteristics by comparing individuals with the disease to those without. GWASs have become one of the most widely used genetic analysis methods in recent decades, mainly because they are large-scale population studies that analyze variations in the human genome without the need for extensive sequencing.[6]

Simultaneously, machine learning methods are tools that have quickly evolved and changed the way we approach clinical and laboratory data.[9,10] These techniques enable insights into big data sets that would be difficult or irresolvable with human processing, and can more easily uncover relationships between biological variables and clinical outcomes, especially in complex disease studies.[6] Depending on the purpose of the study, machine learning algorithms can be independent from, or complementary to, GWASs.[10]

In this work, we established a predictive genetic panel for asthma using machine learning methods, evaluating GWASs as a variable selection method.

METHODS

Study population and genome-wide genotyping

This work was conducted using the ProAR (Programa para o Controle da Asma na Bahia - Salvador, Bahia, Brazil) cohort. The discovery population included 685 unrelated adults (349 asthma

From [a]Instituto de Ciências da Saúde and [b]Programa de Controle da Asma na Bahia (ProAR), Universidade Federal da Bahia, Salvador, Bahia; [c]the Department of Medicine, University of Colorado Denver, Aurora; and [d]the Laboratory of Lymphocyte Signalling and Development, The Babraham Institute, Cambridge.

Received for publication November 14, 2023; revised February 20, 2024; accepted for publication April 5, 2024.
Available online May 18, 2024.

Corresponding author: Camila Alexandrina Figueiredo, PhD, Federal University of Bahia, Institute of Health Sciences, Av Reitor Miguel Calmon, s/nSala 316 Salvador 40110100, Brazil. E-mail: cavfigueiredo@gmail.com; camilavf@ufba.br.

The CrossMark symbol notifies online readers when updates have been made to the article such as errata or minor corrections

2772-8293

https://doi.org/10.1016/j.jacig.2024.100282

Abbreviations used
ANN: Artificial neural network
FVC: Forced vital capacity
GWAS: Genome-wide association study
NB: Naive Bayes
ProAR: Programa de Controle da Asma na Bahia
RF: Random forest
SNV: Single-nucleotide variant
SVM: Support vector machine

cases, 336 controls) aged between 18 and 81 years. Individuals diagnosed with a chronic respiratory disease that affects the lower airways, such as active tuberculosis or sequelae of tuberculosis, cystic fibrosis, lung cancer, or chronic obstructive pulmonary disease, were excluded.

Asthma severity was classified according to the NIH-NHLBI Guidelines for the Diagnosis and Management of Asthma and the Global Initiative for Asthma guidelines, where the following were assessed: daily symptoms, limitations of daily activities, nighttime symptoms more than twice a week, use of bronchodilators more than twice a day, FEV_1 less than 60% of predicted, and number of exacerbations in the previous year.

We performed spirometry by using a portable computerized pulmonary function system (Ferraris KOKO Louisville, Colo) according to American Thoracic Society criteria. We evaluated spirometry before and after 15 minutes of 400 μg salbutamol inhalation. We also used the skin prick test kit (GREER Labs, ALK-Abelló, Horsholm, Denmark) to assess hypersensitivity status.

The skin prick test was conducted using allergens derived from *Dermatophagoides pteronyssinus*, *Blomia tropicalis*, dog epithelium, cat epithelium, *Blatella germanica*, *Periplaneta americana*, *Paspalum notatum*, and *Cynodon dactylon*. A positive skin prick test result was defined as the presence of a papule greater than or equal to 3 mm in diameter.

Genomic DNA was extracted from peripheral blood using the Gentra Puregene Blood Kit (Qiagen, Hilden, Germany). ProAR subjects were genotyped by using the Multi-Ethnic AMR/AFR-8 Kit BeadChip (Illumina), which was specifically designed to capture genetic variation in populations with a significant African and Native American genetic contribution.

Ethical approval was obtained through the Comitê de Ética em Pesquisa of the Unversidade Federal da Bahia and Comissão Nacional de Ética em Pesquisa (CONEP), Brazil (no. 15782/2010).

Data manipulation and quality control

The number of markers genotyped by the chip was 1,544,155. We used PLINK 1.9 to perform data quality control. Data exclusion criteria included more than 10% of data missing for a marker, markers with minor allele frequency less than 1%, and individuals with more than 10% of data lost. Repeated markers with identical features were identified in the raw database and excluded, leaving a final total of 1,009,762 SNVs (see Fig E1 in this article's Online Repository at www.jaci-global.org).

To impute lost data, 3 different methods were tested. First, we used the mice package, but data imputation did not occur because of the wide correlation among variables, and this algorithm requires independence among them. We next tested imputation using linkage disequilibrium between adjacent markers, but this approach imputed only 10.11% of missing data.

The most viable method for imputation was to use the Naive Bayes (NB) algorithm (a supervised method) to predict the missing data. Imputation techniques using NB algorithms for missing data are already well known in the literature.[11] For this imputation, the variable with missing data was considered the outcome, whereas 200 other chromosomally adjacent variables were predictors.

One concern arising with any method of imputing missing data with respect to qualitative biological features is that the relative frequencies should not vary too much after imputation. We monitored this, and examples of relative genotypic frequency before and after imputation can be found in Table E1 (in the Online Repository available at www.jaci-global.org).

GWAS analysis

The GWAS was performed on the data after quality control. Analysis was performed using PLINK 1.9 with an additive genetic model, and logistic regression was performed using age, sex, body mass index, smoking, and genomic ancestry as covariables.[12]

Feature selection methods

In this work, we selected 3 feature selection methods: the Boruta algorithm (feature selection wrapper algorithm with random forest [RF] kernel), markers from GWASs according to their respective *P* values, and elastic net regression (embedded method).

Supervised methods for asthma prediction

We have chosen 10 different algorithms for classification: K-nearest neighbor, NB, artificial neural networks (ANNs), support vector machine (SVM), classification and regression trees, C5.0, bagging, adaptive boosting (AdaBoost), RF, and XGBoost. One crucial step in constructing a machine learning model is the selection of hyperparameters to arrive at the highest model performance possible (tuning process). All methods used, their hyperparameters, and their ranges are listed in Table E2 (in the Online Repository available at www.jaci-global.org). The definitive model was created after choosing the variable selection method. The data partitioning into training and testing data sets was conducted using the caret package. We allocated 80% (n = 549) of the original data to the training data set, whereas the model evaluation was performed on the remaining 20% (n = 136) of our test data set. To ensure optimal model generalization, we used 10-fold cross-validation (control parameters for train) during the training phase. Consequently, 10 models were created for each algorithm, and the average and SD of accuracies across the 10 folds were assessed, in addition to the best model with the highest accuracy. After this phase, the best model was evaluated using the test data set, which yielded more reliable results for accuracy, sensitivity, specificity, kappa coefficient, and F1 score.

Construction of the predictive genetic panel and algorithm performance evaluation

First, we assessed the accuracies of the models by scanning across the 3 feature selection methods mentioned above: Boruta, elastic net, and GWASs. Once the feature selection method was determined, we rebuilt all predictive models to ensure that the

results were consistent. Subsequently, knowing which algorithms generated a more informative set of features, we accessed the importance score of each feature through the varImP function. Each algorithm judges each feature in its own way, ranking the features with a range between 0 and 100. The 10 scores were added for each feature, finally giving a single score for each variable.

Because not all algorithms achieved satisfactory accuracy, the maximum value of a given feature chosen by an algorithm demonstrating low accuracy should not have the same judgment value as that given by an algorithm that performed well. To mitigate this discrepancy, scores were penalized as follows: if the algorithm was accurate up to 79%, we multiplied the variable importance scores by 1 (ie, kept the original value); if the accuracy presented was between 80% and 89%, we multiplied scores by 2, and if the test accuracy was above 90%, we multiplied the score by 3. Finally, we ranked markers in 2 ways: penalized and not penalized.

We used 2 spirometry variables, FEV_1 and FEV_1/forced vital capacity (FVC) ratio, to compare importance scores with the genetic markers, and thus observe how the genetic variables behaved in the face of well-established clinical variables for asthma diagnosis.

We carried out a serial analysis to evaluate the evolution of test accuracies of predictive subpanels (from best to worst), built by adding 1 marker at a time, aiming to find the best panel with the smallest possible number of SNVs. Finally, we performed STRING analysis (https://string-db.org/) to functionally characterize our sets of genes and their networks of interactions.

Statistics

The Kolmogorov-Smirnov test was used to evaluate the normality of the numerical variables used. Means were compared by Mann-Whitney (for 2 independent means) or Kruskal-Wallis (for more than 2 means) test. Dunn's *post hoc* test was used to evaluate differences between means of accuracies generated by the feature selection methods. Fisher exact test was used to infer a relationship between the outcome and categorical variables. The χ^2 goodness-of-fit test was performed to assess whether there was a difference between genotypic proportions before and after imputation of missing data. We used R software, version 4.1.2 (on Linux). The caret package was used for the creation of models, to split train and test data, and to evaluate metrics such as accuracy, sensitivity, specificity, and kappa value. The F1 score was generated by using the MLmetrics package (China Pharmaceutical University, Nanjing, China). The receiver-operating characteristic and area under the curve curves were calculated by using the pROC package (Medical University Centre, Geneva, Switzerland). The assigned statistical significance level was 95%.

RESULTS

Sampling and clinical characterization

A total of 685 participants were evaluated, including 192 (55.01%) individuals with mild to moderate asthma and 157 patients with severe asthma (44.99%) from the ProAR cohort and 336 control subjects. In our sample, 67 (10.31%) patients with asthma-chronic obstructive pulmonary disease overlap syndrome were recorded. Among them, 54 (80.60%) were female. Numerical variables had a non-Gaussian distribution, except FVC.

Mean age differed significantly between asthma severity groups (Table I), although boxplots (see Fig E2, *B*, in this article's Online Repository at www.jaci-global.org) show parity in the general distribution between asthma cases and controls. Overall, patients with mild asthma were younger (mean age, 36 years) than patients with severe asthma (mean age, 48 years) (Fig E2, *A-C*).

Proportionally, there was a greater number of female subjects than males. Most subjects had no previous exposure to tobacco (63.36%). All spirometry measurements, which assess lung function, were significantly different between the asthmatic and nonasthmatic groups. Body mass index was higher in asthmatic patients, but as with age, the distributions are relatively even when looking at the boxplots; there was also a difference between mild (body mass index = 27.40) and severe (body mass index = 29.37) asthma (Table I and Fig E2, *D-F*). Blood eosinophil levels and total IgE levels were significantly higher in asthmatic patients than in controls. Among neutrophils, there was no statistical difference (Table I).

Imputation of missing data

For outcomes of lost data predictions, all goodness-of-fit tests were not significant. Therefore, the frequencies of the original genotypes are statistically equal to the genotypic frequencies after imputing missing data. The average *P* value obtained among all analyses was .9946; minimum *P* value = .8730 and maximum *P* value = .9999.

GWAS results

Among 1.5 million SNVs, we found 49,978 markers associated ($P < .05$) with asthma. Only 1 of them, rs2049388, closest gene *MCG4859*, showed suggestive association ($P = 5.78 \times 10^{-6}$; odds ratio, 0.48; 95% CI, 0.35-0.66; see Manhattan plot in Fig E3 in this article's Online Repository at www.jaci-global.org). rs2049388 is an intergenic SNV located on chromosome 7, at position 10421190 (GRCh38). The most frequent allele is A, the lowest frequency allele is G (minor allele frequency = 0.17), and the alleles are in Hardy-Weinberg equilibrium ($P = .66$).

The genotypic frequencies for healthy individuals were GG = 14 (4.14%), AG = 113 (33.43%), and AA = 211 (62.43%), whereas asthmatic genotypic frequencies were GG = 18 (2.38%), AG = 187 (24.74%), and AA = 551 (72.88%). The χ^2 test of independence obtained a *P* value of .002.

The quantile-quantile plot (Fig E3) of the *P* values illustrates that the observed significant associations were beyond those expected. Moreover, the estimated genomic inflation factor (λ) was 1.01, indicating that the population's genetic structure had an insignificant impact on association results.

Genetic prediction of asthma: Feature selection and machine learning algorithms

For this analysis, we selected the 200 best markers from GWASs according to their respective *P* values. Boruta selected 155 variables (both accepted and undefined). The elastic net algorithm selected 18,271 variables, too many to drive predictions. Therefore, elastic net variables with *B* (absolute value) greater than 1.0 were selected, leaving 74 remaining features (Fig E1).

Table II summarizes the results of the classification algorithms among the 3 variable selection methods. It contains training

TABLE I. Social and clinical characterization of subjects

Characteristic	No asthma (n = 336 [49.1%])		Asthma (n = 349 [50.9%])		*P* value
Age (y)	44.00	(34.50-54.00)	41.00	(28.00-52.00)	.008
Sex					.017
Male	48	(14.29)	75	(21.49)	
Female	288	(85.71)	274	(78.51)	
Smoke					.193
No	211	(62.80)	223	(67.78)	
Yes	125	(37.20)	106	(32.22)	
Pre-BD FEV_1 %	87.13	(80.00-95.49)	70.28	(59.73-81.87)	<.001
Pre-BD FEV_1 (cutoff 80%)					<.001
<80%	86	(25.60)	259	(74.21)	
≥80%	250	(74.40)	90	(25.79)	
Δ FEV_1 (L)	0.07	(0.01-0.13)	0.24	(0.11-0.38)	<.001
Pre-BD FVC %	86.06	(77.71-94.72)	81.11	(72.53-89.26)	<.001
Pre-BD FEV_1/FVC %	101.23	(96.43-106.25)	85.80	(78.35-95.18)	<.001
Pre-BD $FEF_{25\%-75\%}$	94.40	(79.77-113.7)	49.06	(36.33-70.24)	<.001
Eosinophils/mL	150.00	(92.00-241.00)	264.00	(145.50-421.25)	<.001
Eosinophils (categorical)					<.001
<200/mL	221	(66.37)	119	(36.06)	
≥200/mL	112	(33.63)	211	(63.94)	
Neutrophils/mL	3227.00	(2432.00-4206.00)	3533.00	(2467.50-4479.25)	.100
Lymphocytes/mL	2008.00	(1646.00-2335.00)	2011.00	(1692.75-2354.75)	.700
Monocytes/mL	377.00	(297.00-473.00)	445.00	(351.00-544.00)	<.001
Basophils/mL	61.00	(49.00-79)	68.50	(55.00-87.00)	<.001
Total IgE (UI/mL)	118.20	(33.92-339.38)	291.59	(122.04-553.58)	<.001
Skin prick test					<.001
Nonreactive	179	(53.27)	93	(26.65)	
Reactive	80	(23.81)	191	(54.73)	
Not tested	77	(22.92)	65	(18.62)	
BMI	26.16	(22.96-29.64)	27.64	(23.92-31.87)	.002

Values are expressed as n (%) with *P* value obtained by Fisher exact test, or median (interquartile range) with *P* value obtained by Mann-Whitney test. The skin prick test *P* value represents the difference between reactive and nonreactive subjects.
BMI, Body mass index; *pre-BD*, prebronchodilator.

accuracy corresponding to the best test accuracy, the best test accuracy, the mean of training accuracies across the 10 folds, the SD of training accuracies across the 10 folds, as well as sensitivity, specificity, kappa, and F1 score corresponding to the model with the best test accuracy. The best predictive model found using the variables selected by Boruta and GWAS was the SVM using the polynomial kernel, obtaining test accuracies of 90.00% and 94.00%, respectively. SVM showed great sensitivity, specificity, kappa, and F1 score. At this time, SVM appeared to be the best model for the predictive panel, but other algorithms, including bagging, AdaBoost, RF, and XGBoost, were also promising (Table II).

The large difference in accuracy between training and testing is evidence of elastic net feature selection. Overall, the models fit the training data well. However, they could not extrapolate good predictions with new data (Table II). Figs E4 to E6 (in the Online Repository available at www.jaci-global.org) complement these metrics with receiver-operating characteristic curves for each feature selection method.

We observed that 2 methods of variant selection, Boruta and GWAS, were statistically similar when comparing the averages of the generated test accuracies (81.03% and 79.85%), whereas the elastic net method had the worst overall performance (52.87%) and was significantly different from the other 2 (see Fig E7 in this article's Online Repository at www.jaci-global.org). At this point, Boruta was considered the best feature selection method, because it had a higher overall average with less variability among tests.

Using the variables selected by Boruta and the sum of the importance scores of each SNV in descending order (from best to worst), it was possible to observe the performance of each algorithm when adding SNVs 1 by 1 to form sequenced predictive models. We evaluated the scores with and without the penalized system and noticed a subtle difference in the results (Fig 1).

The SVM algorithm used the lowest number of markers to reach an accuracy of 80%, with 50 SNVs needed to reach an accuracy of 80.15% for the nonpenalized system and 55 SNVs to reach an accuracy of 80.88% for the penalized system. Conversely, the classification and regression tree algorithm showed poor test accuracies regardless of the number of SNVs added to the model.

Overall, good predictive accuracies commonly appeared when more than 100 markers were included. Furthermore, we observed that 80% accuracies typically appear first in the penalized system, implying that the penalty creates better predictive accuracies (Fig 1).

When evaluating the panel's interaction with spirometry-related variables, we observed that obstruction (FEV_1/FVC ratio) and FEV_1 had the highest scores and appeared first in all algorithms (see Fig E8 in this article's Online Repository at www.jaci-global.org). The mean initial accuracy among the algorithms was 69.71% ± 3.51%; that is, most algorithms are approximately 70% accurate using only FEV_1 or the FEV_1/FVC ratio. Sequenced analyses without the 2 spirometry features start at a predictive

TABLE II. The summary of performance of algorithms to choose among 3 variable selection methods

Boruta								
Algorithm	**Train***	**Test†**	**Mean‡**	**SD**	**Sensitivity**	**Specificity**	**Kappa**	**F1 score**
KNN	0.84	0.79	0.77	0.06	0.88	0.69	0.57	0.81
NB	0.93	0.82	0.84	0.07	0.97	0.67	0.65	0.85
ANN	0.87	0.78	0.79	0.05	0.97	0.58	0.56	0.82
SVM	**0.93**	**0.90**	**0.85**	**0.06**	**0.91**	**0.90**	**0.81**	**0.91**
Bagging	0.91	0.84	0.85	0.03	0.87	0.81	0.68	0.85
AdaBoost	0.93	0.86	0.86	0.04	0.90	0.82	0.72	0.87
CART	0.69	0.55	0.60	0.08	0.57	0.54	0.10	0.56
C5.0	0.89	0.85	0.83	0.05	0.74	0.96	0.69	0.83
RF	0.91	0.88	0.85	0.04	0.88	0.87	0.75	0.88
XGBoost	0.89	0.84	0.85	0.05	0.84	0.84	0.68	0.84
Mean	0.88	0.81	0.81	0.05	0.85	0.77	0.62	0.82
SD	0.07	0.10	0.08	0.01	0.12	0.14	0.20	0.10
200 GWASs								
Algorithm	**Train**	**Test**	**Mean**	**SD**	**Sensitivity**	**Specificity**	**Kappa**	**F1 score**
KNN	0.91	0.86	0.86	0.03	0.94	0.78	0.72	0.87
NB	0.69	0.65	0.64	0.06	0.99	0.31	0.30	0.74
ANN	0.91	0.51	0.79	0.16	1.00	0.00	0.00	0.67
SVM	**0.98**	**0.94**	**0.93**	**0.04**	**0.90**	**0.99**	**0.88**	**0.94**
Bagging	0.93	0.85	0.85	0.04	0.90	0.81	0.71	0.86
AdaBoost	0.91	0.89	0.88	0.03	0.87	0.91	0.78	0.89
CART	0.64	0.66	0.57	0.04	0.74	0.58	0.32	0.69
C5.0	0.91	0.82	0.84	0.04	0.86	0.79	0.65	0.83
RF	0.91	0.90	0.87	0.03	0.88	0.91	0.79	0.90
XGBoost	0.96	0.90	0.89	0.04	0.86	0.94	0.79	0.89
Mean	0.87	0.80	0.81	0.05	0.89	0.70	0.59	0.83
SD	0.11	0.14	0.11	0.04	0.07	0.32	0.29	0.09
Elastic Net								
Algorithm	**Train**	**Test**	**Mean**	**SD**	**Sensitivity**	**Specificity**	**Kappa**	**F1 score**
KNN	0.82	0.55	0.76	0.07	0.87	0.22	0.09	0.66
NB	0.81	0.49	0.67	0.08	0.10	0.90	0.00	0.17
ANN	0.95	0.49	0.81	0.11	0.19	0.81	0.01	0.27
SVM	0.93	0.54	0.87	0.03	0.57	0.51	0.07	0.55
Bagging	**0.84**	**0.59**	**0.75**	**0.05**	**0.81**	**0.36**	**0.17**	**0.67**
AdaBoost	0.91	0.51	0.83	0.05	0.48	0.55	0.03	0.50
CART	0.75	0.54	0.70	0.04	0.59	0.48	0.07	0.57
C5.0	0.80	0.54	0.68	0.05	0.75	0.33	0.08	0.63
RF	0.87	0.51	0.80	0.05	0.57	0.46	0.03	0.54
XGBoost	0.87	0.51	0.81	0.05	0.61	0.42	0.03	0.56
Mean	0.85	0.53	0.77	0.06	0.55	0.50	0.06	0.51
SD	0.06	0.03	0.07	0.02	0.25	0.21	0.05	0.16

Boldface indicates higher test accuracy.
CART, Classification and regression tree; *KNN*, K-nearest neighbor.
*Training accuracy corresponding to the best test accuracy.
†Best test accuracy.
‡Mean is the average of train accuracies among the 10-folds. SD is the standard deviation of train accuracies among the 10-folds.

mean of 57.72% ± 10.46% for both penalized and nonpenalized systems.

The inclusion of spirometry variables also hastened the 80% accuracy prediction and reduced the accuracy variation for ANN. However, this method was not useful for bagging, AdaBoost, and RF, because it reduced their cumulative predictive power.

The predictive genetic panel created with SVM is composed of 155 variants, including 81 (52.26%) intronic variants, 62 (40.00%) intergenic variants, 7 (4.52%) missense variants, 4 (2.58%) noncoding transcript variants (pseudogenes), and 1 (0.66%) 3′ untranslated region variant. Table E3 (in the Online Repository available at www.jaci-global.org) compiles these genetic markers and relevant information, such as related genes, variant type, alleles, genotypes, and frequencies.

Table III summarizes definitive prediction using the panel of 155 SNVs. Seven models were obtained with accuracies above 80%: NB, SVM, bagging, AdaBoost, C5.0, RF, and XGBoost (Fig 2). With a test accuracy of 91.18%, sensitivity of 92.75%, and specificity of 89.55%, the model generated by SVM was the best predictive model (Table III). Although the ANN model has low accuracy, this test demonstrated 100% sensitivity, predicting all patients with asthma.

The frequencies of SNVs from each chromosome were verified, and it was observed that chromosome 6 had the

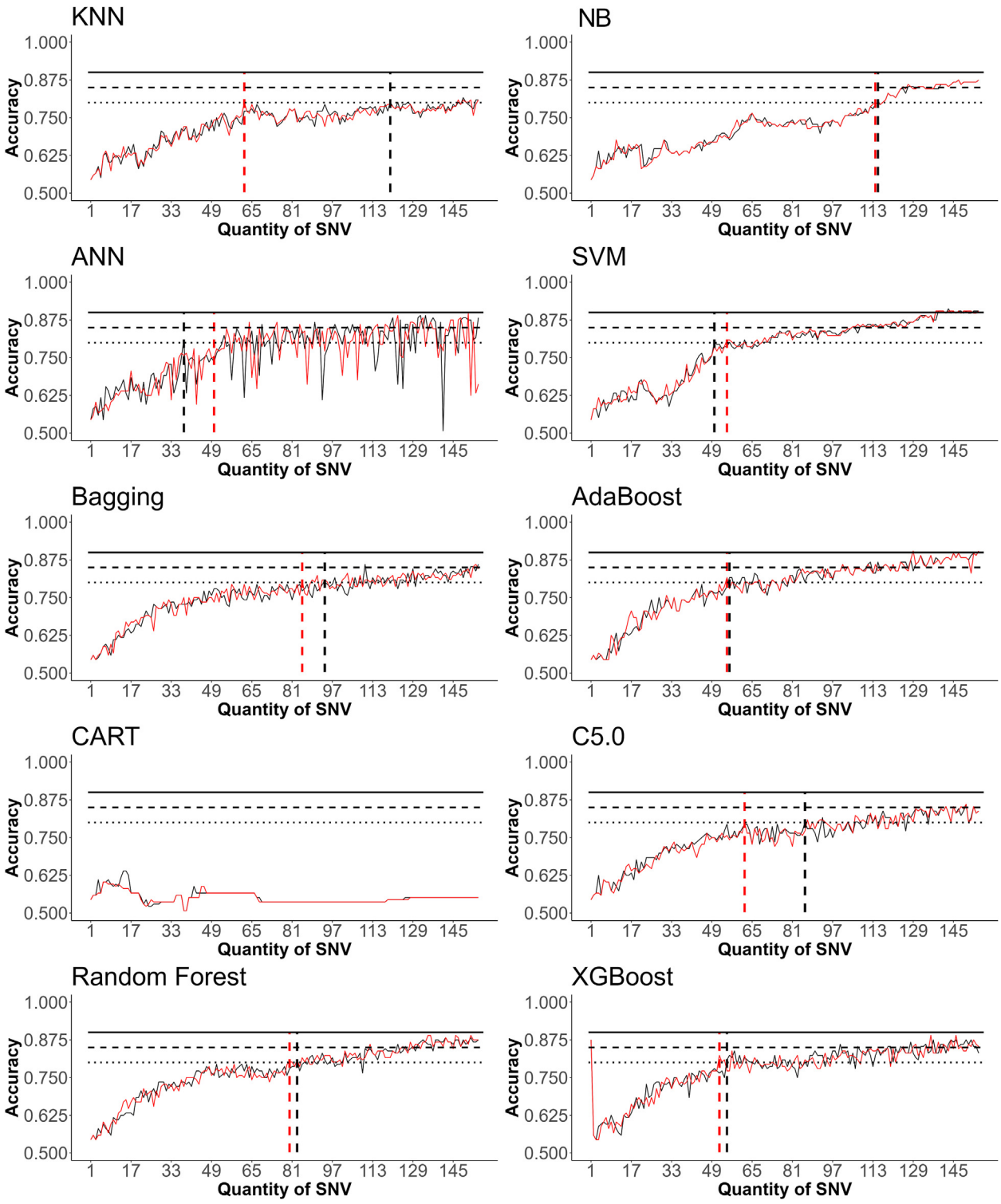

FIG 1. Cumulative sequential analysis, adding one feature at a time from highest to lowest importance score. Even for algorithms with great potential, dozens of markers are necessary for predictive accuracy to be powerful in asthma outcomes. Penalized and nonpenalized systems are similar, but in some cases, such as K-nearest neighbor and C5.0, prediction becomes satisfactory with fewer markers. The vertical lines in each graph represent the first appearance of 80% predictive accuracy (black lines = original data; red lines = penalized scores). Horizontal lines: dotted = 80%, dashed = 85%, filled = 90%.

highest occurrence of SNVs (16) selected for the predictive panel (see Fig E9, *A*, in this article's Online Repository at www.jaci-global.org).

The analysis of linkage disequilibrium between the panel SNVs showed an imbalance between only 2 markers: rs13153665 and rs6869545 (linkage disequilibrium $r^2 = 1.00$; Fig E9, *B*).

TABLE III. The final performance comparison of asthma predictive algorithms using the 155 SNVs selected by Boruta

		Actual, n (%)								
Prediction		**Health**		**Asthma**		**Accuracy test**	**Sensitivity**	**Specificity**	**Kappa**	**F1 score**
KNN	Health	46	(33.82)	8	(5.88)	0.79	0.88	0.69	0.57	0.81
	Asthma	21	(15.44)	61	(44.85)					
NB	Health	56	(41.18)	6	(4.41)	0.88	0.91	0.84	0.75	0.88
	Asthma	11	(8.09)	63	(46.32)					
ANN	Health	10	(7.35)	0	(0.00)	0.58	1.00	0.15	0.15	0.71
	Asthma	57	(41.91)	69	(50.74)					
SVM	Health	60	(44.12)	5	(3.68)	0.91	0.93	0.90	0.82	0.91
	Asthma	7	(5.15)	64	(47.06)					
Bagging	Health	53	(38.97)	8	(5.88)	0.84	0.88	0.79	0.68	0.85
	Asthma	14	(10.29)	61	(44.85)					
AdaBoost	Health	55	(40.44)	7	(5.15)	0.86	0.90	0.82	0.72	0.87
	Asthma	12	(8.82)	62	(45.59)					
CART	Health	36	(26.47)	30	(22.06)	0.55	0.57	0.54	0.10	0.56
	Asthma	31	(22.79)	39	(28.68)					
C5.0	Health	64	(47.06)	18	(13.24)	0.85	0.74	0.96	0.69	0.83
	Asthma	3	(2.21)	51	(37.50)					
RF	Health	57	(41.91)	9	(6.62)	0.86	0.87	0.85	0.72	0.86
	Asthma	10	(7.35)	60	(44.12)					
XGBoost	Health	60	(44.12)	9	(6.62)	0.88	0.87	0.90	0.76	0.88
	Asthma	7	(5.15)	60	(44.12)					

CART, Classification and regression tree; *KNN*, K-nearest neighbor.

When we removed SNV rs6869545 from the panel and reevaluated the predictive accuracy, it dropped from 91.18% to 89.71%. Therefore, we decided to maintain this marker.

Several distinct clusters were observed in STRING analysis (see Fig E10 in this article's Online Repository at www.jaci-global.org). The largest contained 17 genes involved in intracellular signaling processes such as the JAK-STAT pathway (*PRLR* and *ERBB4*), regulation of MAP kinase activity (*PDGFD*), and regulation of SMAD protein signal transduction. In addition, identified genes were involved in cell differentiation, with *SPEF2* contributing to respiratory system development and *JAG2* involved in T-cell differentiation. In this group, *ERBB4* and *DLG2* are also central genes. Finally, the STRING analysis highlighted a cluster of genes related to respiratory diseases, including *JAZF1*, *FTO*, *LINGO2*, and *WDR41* (see Fig E11 in this article's Online Repository at www.jaci-global.org).

DISCUSSION

Machine learning is a subdivision of artificial intelligence that can apply algorithms to make predictions.[13] In recent years, it has been widely applied to predict diagnosis and prognosis of many diseases.[10] In this study, we explore the use of machine learning to predict the genetic risk of asthma in a typically mixed-race Brazilian population. Here, we study the core of the disease, evaluating possible markers in common with the types of asthma or the most prevalent markers of the asthma subtype in our population. However, on the basis of the general characteristics of our asthmatic population, most of our population is, actually, T2 asthma.

There are several ways to choose genetic markers to establish a panel, including using a GWAS.[14] Boruta's algorithm is based on an RF and already widely used in the function of selecting variables.[15] We evaluated 3 different selection methods. As demonstrated, these averages of prediction accuracies were equal between the 200 SNVs selected by GWASs and those chosen by Boruta's algorithm. However, the formation and constitution of the panels were quite different. Because of better performance and less variation between test accuracies, we selected Boruta as the best variable selection algorithm.

Regarding panel construction, we chose a novel method, comparing 10 different algorithms (K-nearest neighbor, NB, ANN, SVM, classification and regression tree, C5.0, Bagging, AdaBoost, RF, and XGBoost).[16-18] The sequence of markers was established using 10 different algorithms, as each gives its own judgment level (or score) for each variable. As such, these importance scores can vary dramatically from one algorithm to another. This would result in a different predictive genetic panel for each of the algorithms, which was not our goal. Instead, we added the scores to reflect the joint judgment of each feature by many algorithms.

However, it was still necessary to give more weight to the algorithms that managed to understand the subject most. With penalized scores, we were able to obtain minimum accuracy provisions of 80% with inclusion of fewer variants, indicating that smaller SNV panels could potentially be used. Nevertheless, we decided to keep all 155 SNVs, so that we could evaluate the importance of each SNV and also retain the highest test accuracy possible.

Overall, the SVM was proven to be the most effective algorithm. Other studies had obtained similar results comparing different methods and found superior outcomes with SVM; However, the accuracy of a previous work was 62.5% in predicting asthma using SNVs.[15] We observed that adding important features to the model, such as spirometry, can lead to better predictive accuracies using this algorithm. In this regard, it is important to emphasize that the high accuracy found in predicting asthma does not directly imply its heritability. The classification model can predict asthma according to the variables imputed in the construction of the model, and certainly unable to detect all possible disease determinants of a patient at once, considering asthma a complex disease where the environment, along with genetics, plays a role.

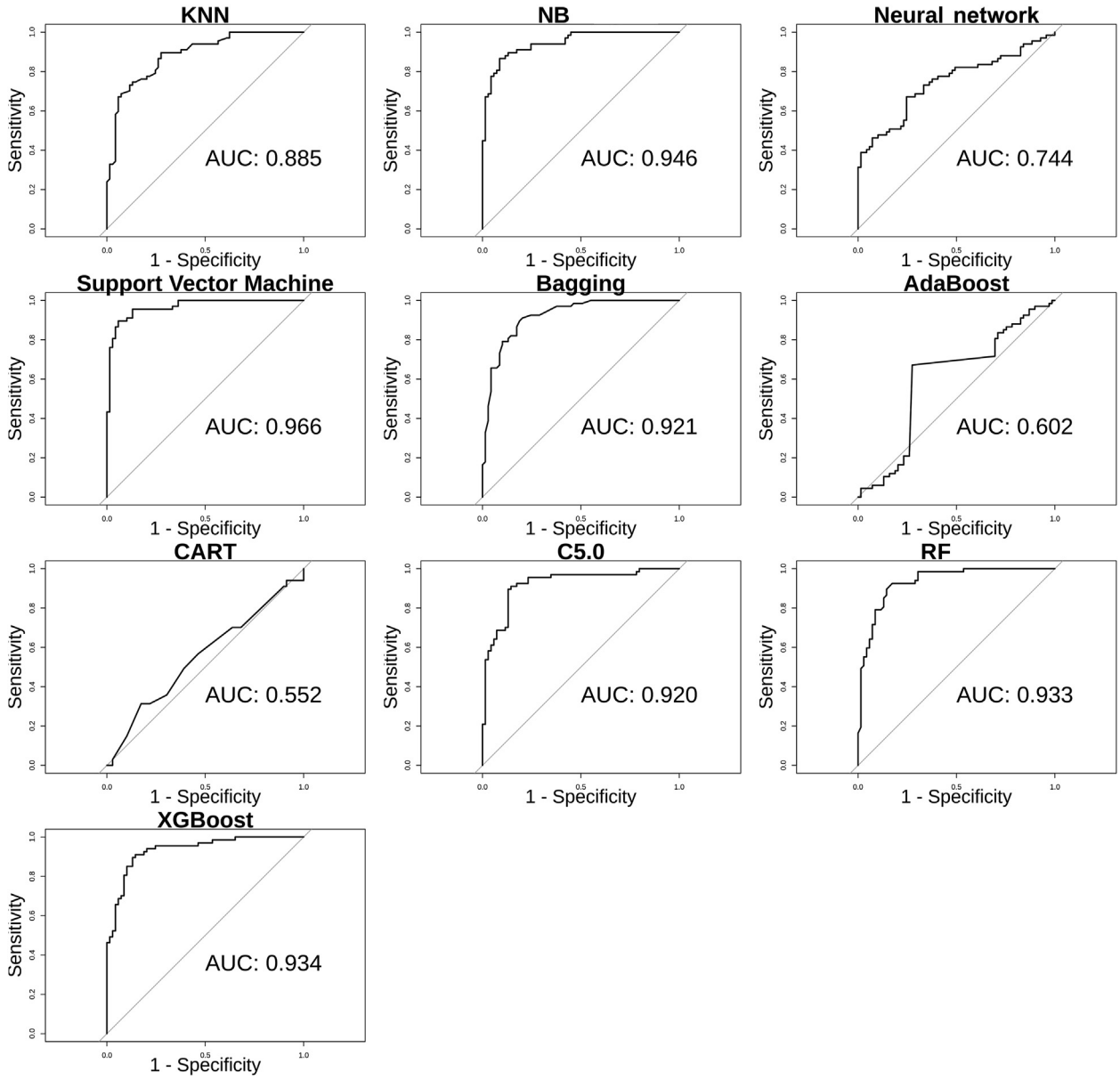

FIG 2. ROC curves of the genetic panel predictions. Compared with other algorithms, the ROC curve and AUC generated by the SVM model were superior. *AUC*, Area under the curve; *KNN*, K-nearest neighbor; *ROC*, receiver-operating characteristic.

The search for new predictive methodologies is not recent. The use of machine learning to assess SNVs associated with asthma was previously explored by Tomita et al[19] in 2004 and 24 polymorphisms in genes of interest in the Japanese population, achieving a model evaluation accuracy of 74.4%.

Recent studies, such as Gaudillo et al,[15] have demonstrated how the sequential addition of genetic markers behaves in predicting a complex trait. In this work, it was observed that there is a plateau of optimal accuracy when selecting up to 310 SNVs for SVM models and up to 400 SNVs for RF models. In our study, we were able to reduce and optimize these numbers to 155 through prior selection via Boruta.

In fact, the search for genetic associations with diseases using machine learning has grown exponentially and is not limited to asthma research.[9,10] Lim et al[20] also tested different algorithms and achieved predictions above 90% for rheumatoid arthritis. The selection method they used was the recursive feature elimination with cross-validation algorithm implemented with an RF (indicating similarities with our methodology), which selected 9 of 76,713 SNVs. These diverse approaches help us to explore a universe of possibilities, moving beyond conventional regression methods to uncover refined associations that may have gone unnoticed before.

In our panel, more than half the SNVs were intronic variants, and a considerable part were intergenic variants. The 7 missense variants belonged to the *NOD1*, *DSG1*, *LOC728743*, *LOC105370777/LOC124903467*, *IGH*, *DST*, and *UFL1-AS1* genes.

NOD1 is well reported in the literature regarding its association with asthma. This gene encodes a member of the nucleotide-binding oligomerization domain (NOD)-like receptors, which is important in the host response to infection and may be involved in T_H2 immune responses and could be critical in T_H2-related conditions such as asthma and atopy.[21] The *DSG1* gene encodes the desmoglein 1 protein, a transmembrane glycoprotein that forms desmosomes. In addition to being associated with autoimmune diseases, it is associated with skin diseases and asthma.[22] The expression of *LOC728743* (a pseudogene) is positively associated with the number of B cells, $CD4^+$ T cells, macrophages, and neutrophils in patients with esophageal carcinoma.[23]

The SNV rs115699578 is an undescribed missense marker in the *IGH* gene, a region that encodes the heavy chain of antibodies.[24] rs79225819, a missense SNV in the *DST* gene,[25] has not been specifically tied to asthma, although *DST* has.[26] *DST* encodes dystonin, a cytoskeleton-binding and cell junction protein whose overexpression is a marker of severity in lung diseases such as pulmonary fibrosis and emphysema.[26] The missense SNV rs4590278 occurs in the long noncoding RNA gene *UFL1-AS1*, which is related to transcriptional regulation but has no known association with asthma.[27] Many other genetic markers in the panel are undescribed, and thorough investigation of these will provide an interesting avenue for follow-up studies.

Our panel also showed protein-protein clusters based on STRING analysis, including some suggesting association with respiratory diseases and regulation of intracellular signal transduction. Thus, interference in the physiological functioning of these proteins by SNVs located in promoter and intronic regions may be the key to understanding the susceptibility to possibly undiscovered asthma endotypes. Therefore, our panel is not restricted to prediction of asthma susceptibility but could contribute to the identification of new candidate genes on asthma, including new pathways never explored before linked to the pathophysiology of this disease.

We describe a predictive genetic panel for asthma built using a hybrid machine learning method. SVM, the best algorithm, was able to classify asthma and nonasthma with high accuracy (91%), sensitivity (93%), and specificity (90%) in our population. In a near future, this panel (or a panel like this) could be useful for detecting an individual's genetic susceptibility to asthma, as some of the markers studied here confirm old associations and shed light on others. The primary function of our machine learning model is to classify the core aspects of the asthmatic disease, without evaluating its phenotypes at this time. This represents the initial step toward contributing to the early diagnosis of the disease.

DISCLOSURE STATEMENT

This study was supported by Brazilian funding agencies (FAPESB/CNPq 009/2014, process no. 8305/2014; FAPESB/CNPq-08/2014, process no. 8665/2014; MCTI/CNPq 14/2023 - 442337/2023-0; CAPES/PRINT, process no. 88887.911599/2023-00; and ERC/CONFAP/CNPQ–INT0002/2023, process no. 1125/2023), CNPq 306705/2022.

Disclosure of potential conflict of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Clinical implications: New insights are possible when new technologies are implemented. Machine learning methods demonstrate new perspectives on genetic relationships, which reflect on the diagnosis and appropriate response to asthma treatment.

REFERENCES

1. Global Initiative for Asthma (GINA). Global Strategy for Asthma Management and Prevention. Available at: http://www.ginasthma.org. Accessed August 14, 2023.
2. Carr TF, Zeki AA, Kraft M. Eosinophilic and noneosinophilic asthma. Am J Respir Crit Care Med 2018;197:22-37.
3. Schoettler N, Strek ME. Recent advances in severe asthma: from phenotypes to personalized medicine. Chest 2020;157:516-28.
4. Augustine T, Al-Aghbar MA, Al-Kowari M, Espino-Guarch M, van Panhuys N. Asthma and the missing heritability problem: necessity for multiomics approaches in determining accurate risk profiles. Front Immunol 2022;13:822324.
5. Kuruvilla ME, Lee FEH, Lee GB. Understanding asthma phenotypes, endotypes, and mechanisms of disease. Clin Rev Allergy Immunol 2019;56:219-33.
6. Ntontsi P, Photiades A, Zervas E, Xanthou G, Samitas K. Genetics and epigenetics in asthma. Int J Mol Sci 2021;22:1-14.
7. Figueiredo RG, Costa RS, Figueiredo CA, Cruz AA. Genetic determinants of poor response to treatment in severe asthma. Int J Mol Sci 2021;22:4251.
8. Li B, Zhang N, Wang YG, George AW, Reverter A, Li Y. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. Front Genet 2018;9:237.
9. Lam S, Arif M, Song X, Uhlén M, Mardinoglu A. Machine learning analysis reveals biomarkers for the detection of neurological diseases. Front Mol Neurosci 2022;15:889728.
10. Nicholls HL, John CR, Watson DS, Munroe PB, Barnes MR, Cabrera CP. Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci. Front Genet 2020;11:350.
11. Khotimah BK, Miswanto S, Suprajitno H. Modeling naïve bayes imputation classification for missing data. IOP Conf Ser Earth Environ Sci 2019;243:012111.
12. Daya M, Rafaels N, Brunetti TM, Chavan S, Levin AM, Shetty A, et al. Association study in African-admixed populations across the Americas recapitulates asthma risk loci in non-African populations. Nat Commun 2019;10:880.
13. Lantz B. Machine learning with R: learn techniques for building and improving machine learning models, from data preparation to model tuning, evaluation, and working with big data. 4th ed. Birmingham: PACKT Publishing; 2023.
14. Sordillo JE, Lutz SM, Jorgenson E, Iribarren C, McGeachie M, Dahlin A, et al. A polygenic risk score for asthma in a large racially diverse population. Clin Exp Allergy 2021;51:1410-20.
15. Gaudillo J, Rodriguez JJR, Nazareno A, Baltazar LR, Vilela J, Bulalacao R, et al. Machine learning approach to single nucleotide polymorphism-based asthma prediction. PLoS One 2019;14:1-12.
16. Andrew SA, Gui J, Sanderson AC, Mason RA, Morlock EV, Schned AR, et al. Bladder cancer SNP panel predicts susceptibility and survival. Hum Genet 2009;125:527-39.
17. Grandell I, Samara R, Tillmar AO. A SNP panel for identity and kinship testing using massive parallel sequencing. Int J Legal Med 2016;130:905-14.
18. Gu JQ, Zhao H, Guo X-Y, Sun H-Y, Xu JY, Wei Y-L. A high-performance SNP panel developed by machine-learning approaches for characterizing genetic differences of Southern and Northern Han Chinese, Korean, and Japanese individuals. Electrophoresis 2022;43:1183-92.
19. Tomita Y, Tomida S, Hasegawa Y, Suzuki Y, Shirakawa T, Kobayashi T, et al. Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma. BMC Bioinformatics 2004;5:120.
20. Lim AJW, Tyniana CT, Lim LJ, Tan JWL, Koh ET. TTSH Rheumatoid Arthritis Study Group, et al. Robust SNP-based prediction of rheumatoid arthritis through machine-learning-optimized polygenic risk score. J Transl Med 2023;21:92.
21. Trindade BC, Chen GY. NOD1 and NOD2 in inflammatory and infectious diseases. Immunol Rev 2020;297:139-61.
22. Bao K, Yuan W, Zhou Y, Chen Y, Yu X, Wang X, et al. A Chinese prescription Yu-Ping-Feng-San administered in remission restores bronchial epithelial barrier to inhibit house dust mite-induced asthma recurrence. Front Pharmacol 2020;10:1698.
23. Liu XS, Liu JM, Chen YJ, Li FY, Wu RM, Tan F, et al. Comprehensive analysis of hexokinase 2 immune infiltrates and m6A related genes in human esophageal carcinoma. Front Cell Dev Biol 2021;9:715883.

24. Collins AM, Yaari G, Shepherd AJ, Lees W, Watson CT. Germline immunoglobulin genes: disease susceptibility genes hidden in plain sight? Curr Opin Syst Biol 2020;24:100-8.
25. National Center for Biotechnology Information (NCBI): dbSNP rs79225819. Available at: https://www.ncbi.nlm.nih.gov/snp/rs79225819. Accessed November 8, 2022.
26. Wang AL, Lahousse L, Dahlin A, Edris A, McGeachie M, Lutz SM, et al. Novel genetic variants associated with inhaled corticosteroid treatment response in older adults with asthma. Thorax 2023;78:432-41.
27. Piao HY, Guo S, Jin H, Wang Y, Zhang J. LINC00184 involved in the regulatory network of ANGPT2 via ceRNA mediated miR-145 inhibition in gastric cancer. J Cancer 2021;12:2336-50.