PLOS ONE

RESEARCH ARTICLE

Machine learning approach to single nucleotide polymorphism-based asthma prediction

Joverlyn Gaudillo[1,5☯], Jae Joseph Russell Rodriguez[2☯], Allen Nazareno[1☯], Lei Rigi Baltazar[1,5☯], Julianne Vilela[3☯], Rommel Bulalacao[4☯], Mario Domingo[4☯], Jason Albia[1,5☯*]

1 Institute of Mathematical Sciences and Physics, University of the Philippines Los Baños, Philippines, **2** Genetics and Molecular Biology Division, Institute of Biological Sciences, University of the Philippines Los Baños, Philippines, **3** Philippine Genome Center Program for Agriculture, Office of the Vice Chancellor for Research and Extension, University of the Philippines Los Baños, Philippines, **4** Domingo Artificial Intelligence Research Center, Los Baños, Philippines, **5** Computational Interdisciplinary Research Laboratories (CINTERLabs), University of the Philippines Los Baños, Philippines

☯ These authors contributed equally to this work.
* jralbia@up.edu.ph

Citation: Gaudillo J, Rodriguez JJR, Nazareno A, Baltazar LR, Vilela J, Bulalacao R, et al. (2019) Machine learning approach to single nucleotide polymorphism-based asthma prediction. PLoS ONE 14(12): e0225574. https://doi.org/10.1371/journal.pone.0225574

Editor: Enrique Hernandez-Lemus, Instituto Nacional de Medicina Genomica, MEXICO

Received: December 4, 2018

Accepted: November 7, 2019

Published: December 4, 2019

Data Availability Statement: The data underlying the results presented in the study are available from: https://github.com/jdgaudillo/SNP-ML.

Funding: We gratefully acknowledge the support of the Department of Science and Technology - Philippine Council for Health Research and Development (DOST-PCHRD) through its CANDLE Program.

Competing interests: The authors have declared that no competing interest exist.

Abstract

Machine learning (ML) is poised as a transformational approach uniquely positioned to discover the hidden biological interactions for better prediction and diagnosis of complex diseases. In this work, we integrated ML-based models for feature selection and classification to quantify the risk of individual susceptibility to asthma using single nucleotide polymorphism (SNP). Random forest (RF) and recursive feature elimination (RFE) algorithm were implemented to identify the SNPs with high implication to asthma. K-nearest neighbor (kNN) and support vector machine (SVM) algorithms were trained to classify the identified SNPs whether associated with non-asthmatic or asthmatic samples. Feature selection step showed that RF outperformed RFE and the feature importance score derived from RF was consistently high for a subset of SNPs, indicating the robustness of RF in selecting relevant features associated with asthma. Model comparison showed that the integration of RF-SVM obtained the highest model performance with an accuracy, precision, and sensitivity of 62.5%, 65.3%, and 69%, respectively, when compared to the baseline, RF-kNN, and an external MeanDiff-kNN models. Furthermore, results show that the occurrence of asthma can be predicted with an Area under the Curve (AUC) of 0.62 and 0.64 for RF-SVM and RF-kNN models, respectively. This study demonstrates the integration of ML models to augment traditional methods in predicting genetic predisposition to multifactorial diseases such as asthma.

Introduction

Risk prediction of complex diseases such as asthma proves to be challenging due to the combinatorial effects of environmental and genetic factors. The major challenges are: 1) obtaining a

sufficient sample size to develop univariate or multivariate predictor and, 2) equipping the predictor with biological variables associated with complex disease. In general, biological modelling studies for disease risk assessment using genetic information have been supplemented by several technologies and study designs. For example, genome-wide association studies (GWAS) use single-nucleotide polymorphism (SNP), a single-base pair change in the DNA sequence, as biomarkers to locate genetic regions associated with a particular trait. Recent GWAS [1–5] identified SNPs linked to candidate genes implicated to asthma. However, GWAS only accounts for a small fraction of individual traits, while heritability related to complex diseases remains unresolved, thus, the emergence of "mystery of missing heritability" phenomenon. Based on [6, 7], the estimated missing heritability of asthma from twin studies ranges from 35% to 70%.

One possible explanation for the "mystery of missing heritability" is the compounding biological interactions inherent to a complex disease [8]. Most of the biological modeling studies [9–11] which established SNP association with asthma still adopt the traditional statistical tests on individual SNP subject to a certain threshold, consequently neglecting the SNP-SNP interactions. Machine learning has the potential to unravel complex genetic and environment interactions constituting a disease by considering SNP-SNP interactions in model development. In genomic analysis, machine learning is employed to identify genetic variants and predict individual's susceptibility to disease based on the identified biomarkers [12]. Despite the limited attempts, recent works have shown that machine learning-based approach augments the analysis on asthma at the genomic level, i.e., from genotype-phenotype association to phenotype prediction. For instance, Xu et al. [12] utilized random forest (RF) algorithm to identify relevant SNPs and to classify asthma case and control samples based on clinical data and SNPs.

Due to the well-known curse of dimensionality inherent to the SNP data, i.e., higher feature size compared to sample size, constructing accurate machine learning-based predictive models for complex disease presents a challenge [12]. One way to circumvent the curse of dimensionality problem is to integrate various machine learning algorithms. The general approach is to employ feature selection method on the SNP data before classifying samples using machine learning-based classifiers. For example, Mieth et al. [13] used support vector machine (SVM) to select relevant SNPs associated with breast cancer. SVM along with Naive Bayes and decision trees have also been used to identify breast cancer cases using SNPs selected via information gain [14]. Furthermore, mean difference calculation and k-nearest neighbor (kNN) have also been employed to quantify SNP relevance and to perform classification task on the breast cancer dataset [15], respectively.

In this work, we developed a novel machine learning-based approach to quantify the individual's susceptibility to asthma based on SNP data. We implemented a two-step approach which involves feature selection and classification steps. The RF and RFE were utilized to select SNPs highly associated to asthma, while SVM and kNN were employed to quantify the propensity of an individual to develop asthma. The predictive models were evaluated using the following metrics: accuracy, sensitivity, precision, and area under the curve (AUC). In the next section, we describe the procedure of data collection, preprocessing, SNP selection, and classification. Finally, the results of integrating feature selection and classification methods will be compared with baseline and external models.

Materials and methods

Data collection and preprocessing

The SNP data was obtained from openSNP database [16], a public repository from 23andMe, deCODEme, and FamilyTreeDNA. The raw dataset is composed of an MS Excel file

containing phenotypes information of users, and text files containing SNP data encoded in Variant Call Format (VCF). Each text file contains meta-information lines and a header line with the following field names: SNP ID, chromosome number, position, and genotype. Phenotypes irrelevant to the study were eliminated and users with incomplete report on asthma were discarded. The cleaned data is composed of 143 samples which represents the users with complete asthma report, 1,088,991 SNPs as features of each sample, and class column that labels each sample based on their phenotype (non-asthmatic or asthmatic).

We adopted similar quality control procedure from [15] which involves missing call filtering and deviation from Hardy-Weinberg equilibrium to select the SNPs. For the missing call filtering, we formulated criteria based on the quality and characteristic of the collected data. The first criterion excludes SNPs whose percentage of present samples is $\leq$ 98%. For the second criterion, the samples were ranked based on the percentage of present SNP in which 90% of the samples were retained. The Hardy-Weinberg equilibrium principle states that in an infinitely sized population with no selection, migration and mutation, random mating will bring about constant allele and genotype frequencies across generations. Genotypes at a bi-allelic locus can then be predicted to have the following distribution of frequencies:

$$\mathrm{p}^2 + 2\mathrm{pq} + \mathrm{q}^2 = 1 \tag{1}$$

where p^2 is the frequency of homozygotes for allele p, $2pq$ is the frequency of heterozygotes for alleles p and q, and q^2 is the frequency of homozygotes for allele q. SNPs that deviated from Hardy-Weinberg equilibrium due to selection, population admixture, cryptic relatedness, genotyping error and genuine genetic association [17] were excluded in the dataset. Chi-square test was implemented using Hardy-Weinberg package in R [18] on allele frequencies of each SNP from the control samples, wherein SNPs with $p - value < 0.001$ were discarded. After data preprocessing and quality control, 128 samples (57 samples with asthma condition during childhood and adulthood and 71 control samples are unaffected with asthma) were retained. A total of 176,288 SNPs was retained with genotypes comprised of minor and major alleles. Finally, genotypic encoding was implemented wherein homozygous, heterozygous, and variant homozygous SNPs were represented by 1, 2, 3, respectively.

Feature selection

Disease detection driven by high-dimensional data entails the use of feature selection method to identify the features highly associated to the phenotype [19]. In this study, RF algorithm which involves building multiple decision trees based on relevant SNPs was used. RF outperforms other feature selection methods in terms of robustness as a result of its ensemble trait of measuring the predictive importance of each feature to build efficient decision trees [19]. The RF algorithm was trained using the following input matrix,

$$\mathrm{D_N} = \{ (X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N) \} \tag{2}$$

where N is the number of samples, and X_i and Y_i are the SNP values and the output class of i^{th} sample, respectively. For each data point, an out-of-bag error (OOBE) and the mean over the forest were calculated. After training, the j^{th} SNP was permuted among the training data and the difference the OOBE before and after permutation was calculated. To obtain the importance score of the j^{th} SNP, the mean of the differences were calculated and normalized.

Model construction

To predict the individual's susceptibility to asthma, two different integrated machine learning models were constructed. These models are RF-kNN and RF-SVM which both employed RF

as the feature selection algorithm to identify the relevant SNPs. Subsequently, the selected SNPs were used as inputs to kNN and SVM classifiers.

k-Nearest Neighbor (*kNN*). kNN algorithm is a nonparametric lazy learner due to the absence of a training phase when performing classification tasks. In this work, the kNN algorithm was implemented on the SNP data to construct a predictive model that classifies non-asthmatic and asthmatic samples. A particular set of SNPs associated to a sample is represented as a continuous variable that can be plotted into the feature space. Hence, given a sample with SNP as features and an unknown label, the kNN algorithm gives prediction by performing a majority voting on the output class of the sample's k nearest neighbors. For further discussion of kNN using SNP data, refer to [15].

To implement kNN, we define X_i be the p-dimensional input vector of the i^{th} sample, and is defined as,

$$\mathrm{X_i} = (SNP_{i1}, SNP_{i2}, \ldots, SNP_{ip}), i = 1, 2, \ldots, p \tag{3}$$

where p is the total number of SNPs. The corresponding output vector Y_i is given by,

$$\mathrm{Y_i} \in \{0, 1\}, i = 1, 2, \ldots, N \tag{4}$$

where N is the total number of samples. The dataset are plotted on a multi-dimensional feature space with their corresponding class label. To determine the class of an unseen sample U, the k nearest neighbor is determined by calculating the Euclidean distance $d(U, X_i)$ relative to the other labeled samples. The Euclidean distance between U and X_i is given by,

$$\mathrm{d(U, X_i)} = \sqrt{\sum_{j=1}^{p} (U_j - X_{ij})^2} \tag{5}$$

where j represents a certain SNP. We note that our selection of Euclidean distance as a similarity measure is motivated by numerous literatures [14, 15, 20] which extensively used this metric to analyze SNPs. Furthermore, to make our work more comprehensive, we tested several distance metrics and compared their performance with the Euclidean distance (see S2 Table).

Support vector machine (*SVM*). SVM is a learning algorithm used in tasks such as nonlinear classification [11, 13, 14], function estimation, and density estimation. The learning algorithm constructs a hyperplane that provides maximum separation or margin, between classes. For detailed discussion of SVM implementation on SNP data, refer to [13, 14].

The construction of the classification model can be viewed as an optimization problem. To implement SVM, the training vector representing SNP values and its corresponding label is given by,

$$\mathrm{D_i} = \{(X_i, Y_i)\}_{i=1}^{N} \tag{6}$$

where N is the total number of samples, and X_i and Y_i represent the SNP values and class label of the i^{th} sample, respectively. The algorithm minimizes,

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \varepsilon_i \tag{7}$$

subject to the following condition,

$$Y_i(w^T X_i + b) \geq 1 - \varepsilon_i, \varepsilon_i \geq 0, \forall_i \tag{8}$$

where w is learning weight vector, b is the bias, ε is the error term, and C is the tradeoff

parameter between the error and margin. After finding the optimal solution, the function,

$$f(z) = sgn(w^T z + b) \tag{9}$$

is used to predict the class of a unseen sample. The sign of the function is given by,

$$f(z) \in \{ -1, +1\} \tag{10}$$

wherein positive (+) and (-) values indicate non-asthmatic and asthmatic, respectively.

Hyperparameter selection and performance evaluation

One crucial step in constructing a machine learning predictive model is the selection of hyperparameters to achieve the highest model performance. Shown in Table 1 are the tested hyperparameter and their corresponding test values to train and optimize the machine learning algorithms.

Cross-validation techniques were employed to estimate the model performance on the unseen data. The RF classifier was evaluated using stratified k-fold cross validation ($k = 10$) which ensures that each fold contains roughly the same number of samples per class. RF-SVM and RF-kNN were evaluated using leave-one-out cross-validation (LOOCV), which discards a single sample to be used as a test set before training the model on the remaining samples; hence, k is equal to the number of samples ($k = 128$). The performance metrics used to evaluate the models were accuracy (ratio of correctly predicted observations to the total observations), precision (ratio of correctly predicted positive observations to the total predicted positive observations), and sensitivity (ratio of correctly predicted positive observations to the actual positive observations). The AUC—receiver operating characteristic (ROC) curve is another measure of performance of a machine learning classifier model. The ROC is a probability curve while the AUC represents the capability of the model to distinguish among classes. The ROC curve is constructed by plotting the true positive rate versus false positive rate.

To further evaluate the results, we compare the integrated models with baseline model and an external model [15] implemented for breast cancer prediction. Scripts were developed using Python programming language and simulations were executed in a machine with 2.3 GHz Intel Core i5 processor, 8Gb random access memory (RAM), and 2 cores.

Table 1. List of hyperparameter values.

Method	Hyperparameter	Range of values
Random Forest	optimum number of trees in the forest, n_estimators	{20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200}
	maximum number of features considered for splitting a node to achieve least uncertainty when creating a tree, max_features	{400, 500, 600}
SVM	kernel	{linear, sigmoid, rbf, poly}
	regularization parameter, C	{1, 10, 100, 1000}
	tolerance, ε	{0.002, 0.1, 0.001, 0.0001}
kNN	number of neighbors, k	1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35

https://doi.org/10.1371/journal.pone.0225574.t001

Results and discussion

Feature selection

We optimized the hyperparameters to obtain the best RF architecture. Results show that the highest performance having accuracy, precision, and sensitivity of 68.75%, 83.3%, and 68%, respectively, is obtained when *n_estimators = 40* and *max_features = 400*. This optimum RF architecture is then used to select the SNPs which will serve as input to train the SVM and kNN classifiers. Training phase revealed that the highest accuracy for RF-SVM and RF-kNN were obtained when the optimum number of features are 310 and 400, respectively, see Fig 1.

To further evaluate the performance of RF in selecting highly associated SNPs, we implemented recursive feature elimination (RFE) in combination with SVM and kNN. As a feature selection method, RFE initially builds a model on the entire dataset and computes feature importance score for each predictor. For the next iteration, the predictor with the lowest feature importance score is omitted from the dataset that will be used to rebuild the model. Result shows (see Table 2) that RF when integrated to SVM and kNN classifiers has a better performance as a feature selection method compared with RFE.

Shown in Fig 2 is the feature importance scores of top 20 SNPs. Table 3 shows the characteristics of the top five SNPs that were selected by the RF based on feature importance score. Based on the feature importance score, the top five SNPs were consistently selected by RF in every fold during the cross-validation procedure. This result is indicative that integration of

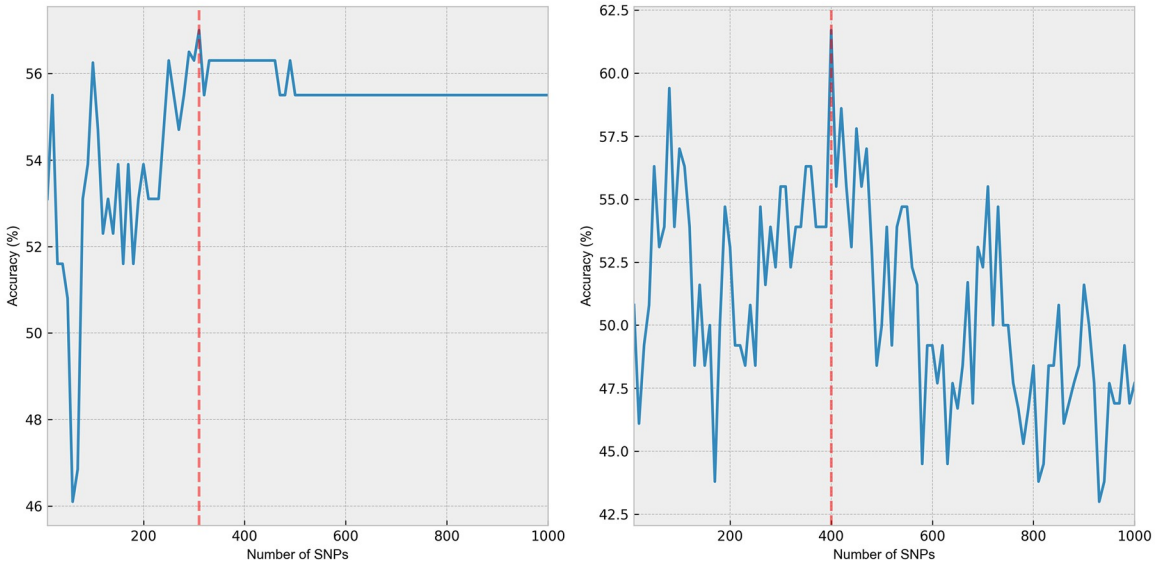

Fig 1. Accuracy vs. number of SNP for SVM (left) and kNN (right).

https://doi.org/10.1371/journal.pone.0225574.g001

Table 2. Comparison of feature selection methods.

Model	Accuracy (%)
RF-SVM	62.17
RF-kNN	61.70
RFE-SVM	50.90
RFE-kNN	49.68

https://doi.org/10.1371/journal.pone.0225574.t002

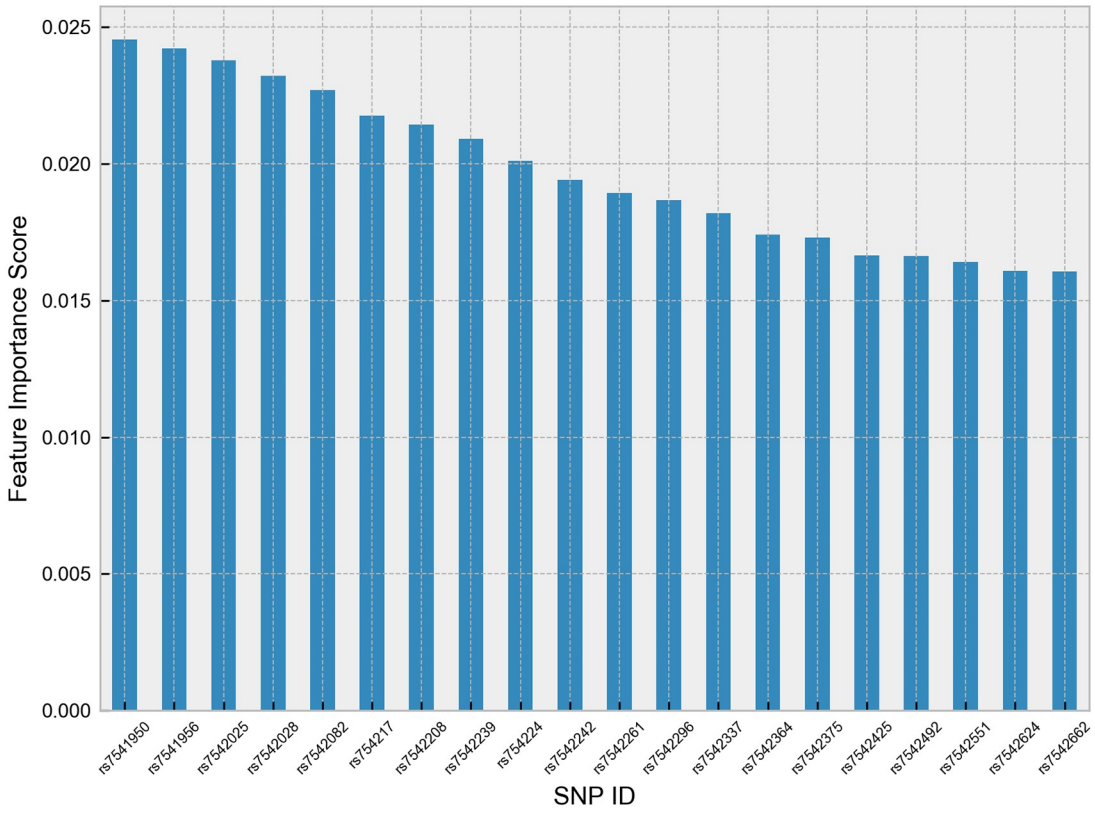

Fig 2. Feature importance score of top 20 SNPs.

https://doi.org/10.1371/journal.pone.0225574.g002

RF with classifiers is robust in selecting important SNPs. Favoring a robust feature selection algorithm apart from classification performance is highly desirable to further gain insight on the SNP data. Moreover, the robustness of the RF can be attributed to its stability in selecting important features, i.e., RF is insensitive to perturbations in the dataset [19].

While the asthma-associated SNPs selected by RF has low feature importance scores, it is interesting to note that prior studies have not identified these SNPs to be associated to asthma, nor included in the 19 asthma-associated SNPs [22] present in the dataset. This observation is attributed to: 1) various sequencing platforms were used which may lead to the exclusion of

Table 3. Characteristics of top five SNPs selected by RF based on importance score [21].

SNP ID	Chromosome Location	Gene	Functional Consequence	GeneCard Summary
rs7541950	1:147903855	*GJA8*	intron variant	GJA8 (Gap Junction Protein Alpha 8) is a Protein Coding gene. Diseases associated with GJA8 include Cataract 1, Multiple Types and Cataract Microcornea Syndrome. Among its related pathways are Development Slit-Robo signaling and Vesicle-mediated transport. Gene Ontology (GO) annotations related to this gene include channel activity and gap junction channel activity
rs7541956	1:111366426	*LOC105378904*	intron variant	RNA Gene, and is affiliated with the ncRNA class
rs7542025	1:40643890	*RIMS3*	intron variant	RIMS3 (Regulating Synaptic Membrane Exocytosis 3) is a Protein Coding gene. Gene Ontology (GO) annotations related to this gene include ion channel binding
rs7542028	1:168718584	*DPT*	intron variant	DPT (Dermatopontin) is a Protein Coding gene. Diseases associated with DPT include Commensal Bacterial Infectious Disease and Anisometropia
rs7542082	1:118617643	NA	NA	NA

https://doi.org/10.1371/journal.pone.0225574.t003

Table 4. Comparison of performance between RF-SVM and RF-kNN to the baseline SVM and kNN models. The selected SNPs for Baseline SVM and kNN refers to the 19 asthma-associated SNPs.

Models	Accuracy (%)	
	Entire Dataset	Selected SNPs
Baseline SVM	52.83	51.81
Baseline kNN	49.69	49.09
RF-SVM	56.30	62.17
Baseline kNN	54.70	61.70

https://doi.org/10.1371/journal.pone.0225574.t004

lung-function associated SNPs or affect the phenotypic variance within the population; 2) as demonstrated in [23–26], RF takes into account SNP-SNP interactions while the traditional approach only performs direct genotype-phenotype association testing. Considering the result of our feature selection scheme, in which the selected SNPs are currently excluded in the list of known SNPs associated to asthma [22], we posit that the integration method can be utilized as a preliminary method to discover new biomarkers associated with complex diseases.

Model comparison

The SVM and kNN were implemented on both entire and selected dataset to establish baseline models. Furthermore, a more stringent test on the performance of SVM and kNN classifiers were implemented on the 19 asthma-associated SNPs. Considering the higher accuracy of RF as a feature selection method when integrated to SVM and kNN as discussed above, we compare the performance of RF-SVM and RF-kNN with the baseline model. Table 4 shows that the integrated models obtained higher classification performance than the baseline models.

To further evaluate the performance of the integrated models, MeanDiff-kNN derived from [15] was implemented on the dataset. Fig 3 shows the comparison of the three integrated models RF-SVM, RF-kNN, MeanDiff-kNN. The RF-SVM attained the highest performance among the models with an accuracy, precision, and sensitivity of 62.17%, 65.3%, and 69%, respectively. We note that the baseline model [15] previously proposed for breast cancer detection when applied to the asthma dataset achieved 55% accuracy. Similar performance metrics percentage of 61.7% and 55% were attained for RF-kNN and MeanDiff-kNN, respectively. The SVM outperformed other classifiers in predicting asthma susceptibility by considering the interactions among SNPs to construct a more suitable predictive model tailored to the input data [13, 14, 27]. Compared to other machine learning methods, SVM is very powerful in detecting intricate trends and patterns in complex genomic dataset [20]. While SVM has better performance in high-dimensional data with few samples, kNN works well with data points in low-dimensional space. Shown in Table 5 is the optimal hyperparameters for each model.

Fig 4 shows the ROC curves of RF-SVM, RF-kNN, and MeanDiff-kNN. Among the integrated models, RF-kNN attained the highest AUC of 0.64. The AUC of the RF-SVM and RF-kNN were comparable to the results in [12] which reported an AUC of 0.66 using 160 SNPs and clinical data to predict severe asthma exacerbations. Despite the absence of clinical data in the predictor variables, our integrated models obtained good performance in distinguishing case from control samples. Moreover, the integrated models achieved good prediction results, notwithstanding the limited sample size. This further indicate that the combination of machine learning algorithms used in the proposed models are able to predict asthma susceptibility with satisfactory performance despite the limited biological variables used in the model construction.

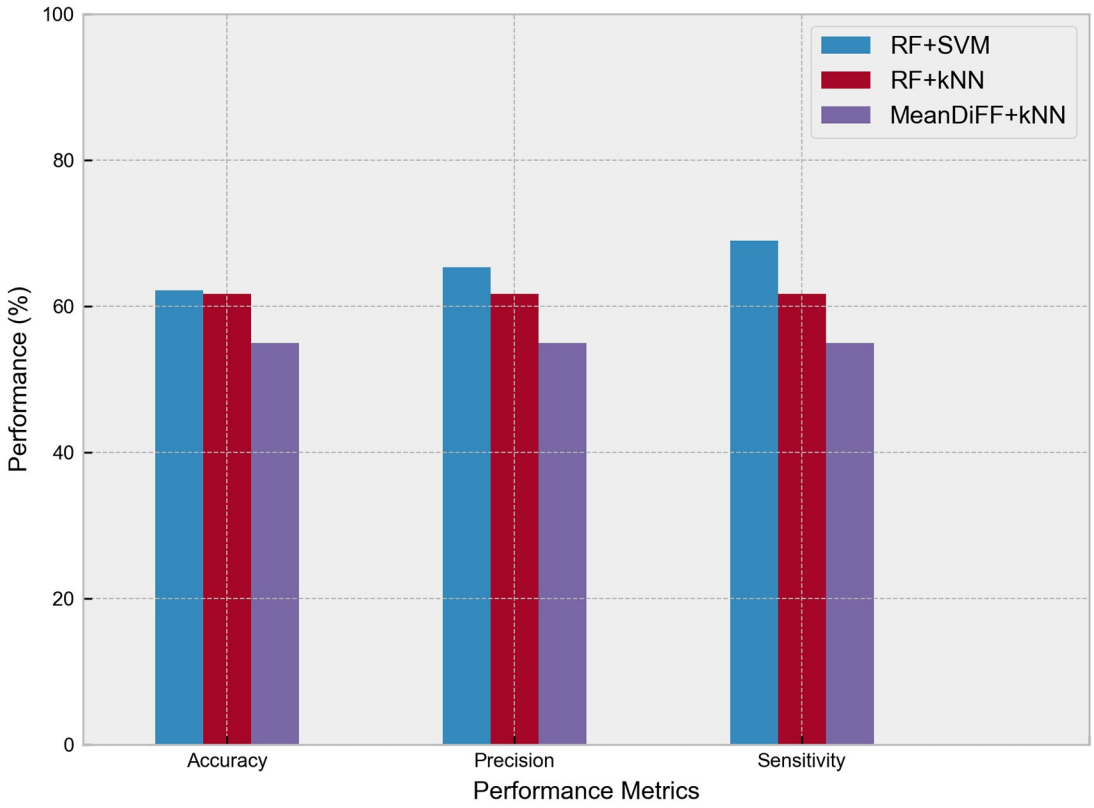

Fig 3. Performance of the three ML models (RF-SVM, RF-kNN, MeanDiff-kNN).

https://doi.org/10.1371/journal.pone.0225574.g003

Table 5. Optimal hyperparameters determined for the models.

Method	Hyperparameter values
SVM	kernel = rbf
	regularization parameter, C = 100
	tolerance, ε = 0.001
	number of selected SNP = 310
kNN	number of nearest neighbors, k = 7
	number of selected SNP = 400

https://doi.org/10.1371/journal.pone.0225574.t005

Conclusion

We developed machine learning-based predictive models which utilized SNP data to quantify the individual susceptibility to asthma. More precisely, we employed an integrated approach consisting of a feature selection step to identify the SNPs highly associated to asthma, and a classification step to predict asthma susceptibility. Results show that the integrated RF-SVM model achieves the highest accuracy, precision, sensitivity, and AUC compared to RF-kNN and baseline models. Furthermore, this study demonstrated that the integration of various machine learning methods is a well-suited approach to investigate high dimensional SNP data for genotype-phenotype association and phenotype prediction. To a large extent, we have shown how machine learning approach can augment the insights derived from GWAS in analyzing massive and complex biological data obtained from next generation sequencing (NGS)

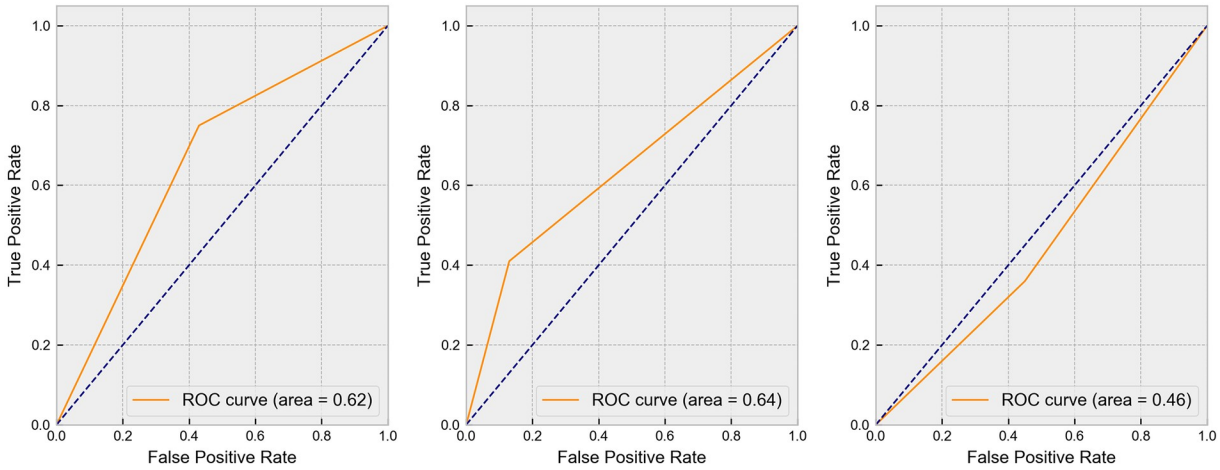

Fig 4. ROC Curves of RF-SVM (a), RF-kNN (b), and MeanDiff-kNN (c).

https://doi.org/10.1371/journal.pone.0225574.g004

platforms. The integrated approach can be naturally extended on a wide variety of biological data such genetic mutations, copy number variations as well as clinical data which may take into account environmental factors to come up with a more holistic understanding of genetic etiology of complex diseases.

Source code

The Python codes as well as the SNP data used in this study is available in this github repository.

Supporting information

S1 Table. Feature importance table.
(PDF)

S2 Table. Distance metrics performance table.
(PDF)

Author Contributions

Conceptualization: Joverlyn Gaudillo, Jae Joseph Russell Rodriguez, Allen Nazareno, Lei Rigi Baltazar, Julianne Vilela, Jason Albia.

Data curation: Joverlyn Gaudillo, Rommel Bulalacao, Jason Albia.

Formal analysis: Joverlyn Gaudillo, Jae Joseph Russell Rodriguez, Allen Nazareno, Lei Rigi Baltazar, Julianne Vilela, Jason Albia.

Funding acquisition: Mario Domingo.

Investigation: Joverlyn Gaudillo, Jae Joseph Russell Rodriguez, Jason Albia.

Methodology: Joverlyn Gaudillo, Jae Joseph Russell Rodriguez, Lei Rigi Baltazar, Julianne Vilela, Jason Albia.

Project administration: Mario Domingo, Jason Albia.

Resources: Jason Albia.

Software: Rommel Bulalacao, Jason Albia.

Supervision: Mario Domingo, Jason Albia.

Validation: Joverlyn Gaudillo, Jason Albia.

Visualization: Jason Albia.

Writing – original draft: Joverlyn Gaudillo, Jae Joseph Russell Rodriguez, Allen Nazareno, Julianne Vilela, Jason Albia.

Writing – review & editing: Joverlyn Gaudillo, Jae Joseph Russell Rodriguez, Allen Nazareno, Julianne Vilela, Jason Albia.

References

1. Hancock DB, Romieu I, Shi M, Sienra-Monge JJ, Wu H, Chiu GY, et al. Genome-wide association study implicates chromosome 9q21.31 as a susceptibility locus for asthma in Mexican children. PLoS Genet. 2009; 5 (8): e1000623. https://doi.org/10.1371/journal.pgen.1000623 PMID: 19714205
2. Himes BE, Hunninghake GM, Baurley JW, Rafaels NM, Sleiman P, Strachan DP, et al. Genome-wide association analysis identifies PDE4D as an asthma-susceptibility gene Genet. 2009; 84(5): 581–593.
3. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, Depner M, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature. 2007; 448(7152): 470–473. https://doi.org/10.1038/nature06014 PMID: 17611496
4. Li X, Howard TD, Zheng SL, Haselkorn T, Peters SP, Meyers DA, et al. Genome-wide association study of asthma identifies RAD50-IL13 and HLA-DR/DQ regions. J Allergy Clin Immunol. 2010; 125 (2): 328–335.e311. https://doi.org/10.1016/j.jaci.2009.11.018 PMID: 20159242
5. Sleiman PM, Flory J, Imielinski M, Bradfield JP, Annaiah K, Willis-Owen SA, et al. Variants of DENND1B associated with asthma in children. N Engl J Med. 2010; 362 (1): 36–44. https://doi.org/10.1056/NEJMoa0901867 PMID: 20032318
6. Duffy DL, Martin NG, Battistutta D, Hopper JL, Mathews JD. Genetics of asthma and hay fever in Australian twins. Am Rev Respir Dis 1990; 142:1351–1358. https://doi.org/10.1164/ajrccm/142.6_Pt_1.1351 PMID: 2252253
7. Nieminen MM, Kaprio J, Koskenvuo M. A population-based study of bronchialasthma in adult twin pairs. Chest 1991; 100:70–75. https://doi.org/10.1378/chest.100.1.70 PMID: 2060393
8. König IR, Auerbach J, Gola D, Held E, Holzinger ER, Legault MA, et al. Machine learning and data mining in complex genomic data—a review on the lessons learned in Genetic Analysis Workshop 19. InBMC genetics 2016 Dec (Vol. 17, No. 2, p. S1). BioMed Central.
9. Savenije OE, John JM, Granell R, Kerkhof M, Dijk FN, de Jongste JC, et al. Association of IL33–IL-1 receptor—like 1 (IL1RL1) pathway polymorphisms with wheezing phenotypes and asthma in childhood. Journal of Allergy and Clinical Immunology. 2014 Jul 1; 134(1):170–7. https://doi.org/10.1016/j.jaci.2013.12.1080 PMID: 24568840
10. Forno E, Celedón JC. Predicting asthma exacerbations in children. Current opinion in pulmonary medicine. 2012 Jan; 18(1):63. https://doi.org/10.1097/MCP.0b013e32834db288 PMID: 22081091
11. Spycher BD, Henderson J, Granell R, Evans DM, Smith GD, Timpson NJ, et al. Genome-wide prediction of childhood asthma and related phenotypes in a longitudinal birth cohort. Journal of allergy and clinical immunology. 2012 Aug 1; 130(2):503–9. https://doi.org/10.1016/j.jaci.2012.06.002 PMID: 22846752
12. Xu M, Tantisira KG, Wu A, Litonjua AA, Chu JH, Himes BE, et al. Genome Wide Association Study to predict severe asthma exacerbations in children using random forests classifiers. BMC medical genetics. 2011 Dec; 12(1):90. https://doi.org/10.1186/1471-2350-12-90 PMID: 21718536
13. Mieth B, Kloft M, Rodríguez JA, Sonnenburg S, Vobruba R, Morcillo-Suárez C, et al. Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. Scientific reports. 2016 Nov 28; 6:36671. https://doi.org/10.1038/srep36671 PMID: 27892471
14. Listgarten J, Damaraju S, Poulin B, Cook L, Dufour J, Driga A, et al. Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. Clinical cancer research. 2004 Apr 15; 10 (8):2725–37. https://doi.org/10.1158/1078-0432.ccr-1115-03 PMID: 15102677
15. Hajiloo M, Damavandi B, HooshSadat M, Sangi F, Mackey JR, Cass CE, et al. Breast cancer prediction using genome wide single nucleotide polymorphism data. BMC bioinformatics. 2013 Oct; 14(13):S3. https://doi.org/10.1186/1471-2105-14-S13-S3 PMID: 24266904

16. Opensnp.org. (2018) openSNP. [online] Available at: https://opensnp.org/ [Accessed Mar. 2018].
17. Zeng P, Zhao Y, Qian C, Zhang L, Zhang R, Gou J, et al. Statistical analysis for genome-wide association study. Journal of biomedical research. 2015 Jul; 29(4):285. https://doi.org/10.7555/JBR.29.20140007 PMID: 26243515
18. Graffelman J, Camarena JM. Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. Human heredity. 2008; 65(2):77–84. https://doi.org/10.1159/000108939 PMID: 17898538
19. Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. InJoint European Conference on Machine Learning and Knowledge Discovery in Databases 2008 Sep 15 (pp. 313–325). Springer, Berlin, Heidelberg.
20. Batnyam N, Gantulga A, Oh S. An efficient classification for single nucleotide polymorphism (SNP) dataset. InComputer and Information Science 2013 (pp. 171–185). Springer, Heidelberg.
21. Genecards.org. (2018). [online] Available at: http://www.genecards.org/ [Accessed Oct. 2018].
22. Snpedia.com. (2018). SNPedia. [online] Available at: https://www.snpedia.com/ [Accessed Oct. 2018].
23. Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics. 2012 Jun 1; 99(6):323–9. https://doi.org/10.1016/j.ygeno.2012.04.003 PMID: 22546560
24. Heidema AG, Boer JM, Nagelkerke N, Mariman EC, Feskens EJ. The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. BMC genetics. 2006 Dec; 7(1):23. https://doi.org/10.1186/1471-2156-7-23 PMID: 16630340
25. Schwender H, Zucknick M, Ickstadt K, Bolt HM, GENICA network. A pilot study on the application of statistical classification procedures to molecular epidemiological data. Toxicology letters. 2004 Jun 15; 151 (1):291–9. https://doi.org/10.1016/j.toxlet.2004.02.021 PMID: 15177665
26. Lunetta K. L., Hayward L. B., Segal J., Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. Toxicology letters. 2004 Jun 15; 151(1):291–9.
27. Huang S, Cai N, Pacheco PP, Narandes S, Wang Y, Xu W. Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics-Proteomics Cancer Genomics-Proteomics. 2018 Jan 1; 15(1):41–51. https://doi.org/10.21873/cgp.20063 PMID: 29275361