The Determination of Distinctive Single Nucleotide Polymorphism Sets for the Diagnosis of Behçet's Disease

Yunus Emre Işık[1], Yasin Görmez[1], Zafer Aydın[2], and Burcu Bakir-Gungor[2]

Abstract— Behçet's Disease (BD) is a multi-system inflammatory disorder in which the etiology remains unclear. The most probable hypothesis is that genetic tendency and environmental factors play roles in the development of BD. In order to find the essential reasons, genetic changes on thousands of genes should be analyzed. Besides, there is a need for extra analysis to find out which genetic factor affects the disease. Machine learning approaches have high potential for extracting the knowledge from genomics and selecting the representative Single Nucleotide Polymorphisms (SNPs) as the most effective features for the clinical diagnosis process. In this study, we have attempted to identify representative SNPs using feature selection methods, incorporating biological information and aimed to develop a machine-learning model for diagnosing Behçet's disease. By combining biological information and machine learning classifiers, up to 99.64% accuracy of disease prediction is achieved using only 13,611 out of 311,459 SNPs. In addition, we revealed the SNPs that are most distinctive by performing repeated feature selection in cross-validation experiments.

Index Terms—Behçet's disease (BD), feature selection, machine learning, disease prediction, most informative SNPs

1 Introduction

Behçet's disease (BD), which is characterized by recurrent attacks, is one of the multi-system inflammatory diseases. BD is first described by Hulusi Behçet in 1937. It usually arises at ages between twenty to forty, and is rarely seen above the age of 50. Even though this disease appears worldwide, it is more prevalent throughout ancient Silk road, spanning from East Asia to the Middle East and the Mediterranean. Turkey has the highest prevalence rate of BD in the world with 420 cases per 100,000 persons. The prevalence rates of BD in Iran, Saudi Arabia, Iraq, Italy, Israel, China, Japan, and Egypt follow Turkey, respectively [1].

Due to the wide-range of symptoms, the confirmation of diagnosis can be difficult for BD. There is no gold-standard test to diagnose BD. Several tests such as blood and urine, skin biopsy, and pathergy may be neccessary to make a final decision. For these reasons, it is one of the diseases that are difficult to diagnose [2].

The symptoms of BD might vary from person to person, however, mouth, skin, genitals and eyes are commonly influenced by BD. It is reported that joints, tongue, vascular system, digestive system, and the brain can also be affected [3]. Like other auto-immune and auto inflammatory syndromes, the exact etiology of BD remains to be elucidated. However, the most probable hypothesis is that the viral and bacterial inflammatory reactions triggered by environmental effects and the genetic tendency plays important roles in the development of BD [4]. The most common genetic variations that may lead to disease are Single Nucleotide Polymorphisms (SNPs), which are single base-pair or nucleotide changes that occur almost once in every 1,000 nucleotides on average throughout the DNA sequence [5]. A SNP may not be harmful, but if it is located at the regulatory or coding region of a gene, they can change mRNA transcription stability and the type of amino acid during translation. Therefore, such genetic variations ultimately may change the function of the synthesized protein and can potentially cause disease.

Thanks to the technological developments in the field of genetics, researchers can perform case-control studies with manageable costs. Genome-wide association studies (GWAS) is a frequently used, powerful and efficient approach to reveal common genetic risk factors of diseases through measuring and comparing the frequencies of DNA sequence variations (SNPs) within the cases vs. controls. Using the information on genetic variations and risk factors, post analysis of GWAS enables the diagnosis of diseases and prediction of disease risk [6].

Several genome wide association studies on BD show that the SNPs located on HLA-B51 and HLA-B genes have the strongest association with BD [7]. Besides, recent GWA studies confirmed that STAT4, IL10, IL23R [8], CCL2 [9], NAALADL2, YIPF7 [10] genes include significant susceptibility factors for BD. Such post GWAS analysis methods help to integrate individual's genetic profiles into the models to predict the disease status in a robust manner [11].

In order to find the essential reasons of complex diseases such as BD where both genetic and environmental factors play a role and predict the disease state the genetic

[1]*Y.E. Işık and Y. Görmez are with the Department of Management Information Systems, Cumhuriyet University, Sivas 58140 Turkey. E-mail: {yeisik, yasingormez}@cumhuriyet.edu.tr.*
[2]*Z. Aydın and B.Bakir-Gungor are with the Department of Computer Engineering, Abdullah Gül University, Kayseri 38170 Turkey. E-mail: {burcu.gungor, zafer.aydin}@agu.edu.tr.*

changes on thousands of genes should be collectively analyzed. Machine-learning models can be useful in that respect allowing the automation of the diagnosis process. Recent developments in artificial intelligence and machine learning enable the implementation of computational models which use related inputs (SNPs) and outputs (phenotypes, affected or healthy) for an effective disease detection [12]. These machine learning models aim to maximize the prediction power at the level of individuals and try to provide individualized disease risk predictions based on genetic profiles of individuals [13]. Machine learning models can also help finding the most important genes that contribute to disease state. This can be achieved using feature selection algorithms, which search the space of features using combinatorial optimization algorithms to find the best feature subset.

Several studies are conducted on the disease detection using machine-learning models. Manor and Segal proposed a bootstrapping approach called BootRank with majority voting, which re-samples the data multiple times to improve the diagnosis of Type 1 Diabates. They tested this method on WTCCC dataset and obtained 0.90 AUC score by outperforming logistic regression (LR), support vector machines (SVM) and Random Forest (RF) [14]. Maciukiewicz et al. proposed an SVM model to predict response and remission to anti-depressant drugs prescribed for major depressive disorder (MDD) and forecasted treatment response with 52% accuracy by using SNPs obtained by GWAS analysis [15]. Recio and Forni compared Bayesian regression, boosting, RF to analyze discrete traits in a genome-wide prediction study on simulated animal SNPs dataset. In their study, RF gave the best performance of 0.67 AUROC score [16].

In these related studies, all of the SNPs that are genotyped in GWAS are used. However, it is known that only a subset of SNPs have an effect on the complex disease traits. Besides that, as the number of SNPs get high, the problem becomes computationally impractical. One approach to overcome these drawbacks is to find the most relevant SNPs using statistical and biological methods. Anekboon et. al. proposed an approach which utilize genetic algorithm to select a subset of relevant SNPs and BoostMode-SVM to improve their prediction accuracy. They tested their approach on Thalassemia and Crohn's disease, in which the numbers of SNPs are decreased from 853 to 6 and from 103 to 8 respectively. Then Boosted-SVM, Cart and Optimized Random Forest (ORF) algorithms were applied on the reduced data. Results shows that the proposed approach outperformed other methods with 71.57% accuracy for Thalassemia and 71.06% accuracy for Crohn's disease [17]. Wei et al. performed risk prediction for Crohn's Disease (CD) and Ulcerative Colitis (UC), which are subcategories of inflammatory bowel disease by using 17,379 CD cases, 13,453 UC cases and 22,442 healthy controls. GWAS analysis is performed and SNPs are filtered using p-values $< 10^{-4}$. Finally penalized logistic regression (LR with L1 norm regularization) is applied over SNPs. Results show that the LR model achieved 0.86 and 0.83 AUC scores for CD and UC, respectively [18]. Similarly, Kooperberg et al. compared linear regression, Lasso and Elastic Net methods on Crohn's disease and point out that Lasso resulted in best AUC score of 0.637 [19]. Lopez et al. built a decision-support system for type 2 diabetes risk prediction using SNPs and clinical information of patients. The number of SNPs is decreased to 96 using k-nearest neighbor (k-NN) as the classifier in a wrapper method for feature selection. Then Random Forest, Logistic Regression and SVM are employed on 3 different type of datasets: raw data, clinical data and feature selected data. As a result, RF yielded a 0.89 AUC score and is selected as the best classifier [20]. Montanez et al. applied different machine learning approaches to predict obesity using SNPs and clinic information of participant samples as features. They also decreased the number of SNPs to 13 using feature selection methods. Then various machine learning approaches are employed on the dataset and SVM obtained the best AUC score of 0.905 [21]. Shigemizu et al. investigated the effect of genetic and clinical factors to build a risk prediction model for type 2 diabetes for Japanese individuals. The most significant SNPs are detected using Cochran-Armitage trend test, asymptotic Bayes factor (ABF), and sure independence screening methods. Then ridge regression, elastic net, and lasso models were tested by adding clinical factors and 1-by-1 top-ranked SNPs. The best results were obtained with a 0.8057 AUC using the Lasso method, which included the top-9 SNPs and clinical factors as features [22].

All these studies show that using patient's SNP information to predict and diagnose disease risk can play a significant role in precision medicine or to develop a diagnostic tool. However, to the best of our knowledge there is no study that develops a decision system for automatic diagnosis of BD using SNP information as inputs and machine learning methods as the classifiers. In this study, we have attempted to identify representative SNP subset and to develop a decision support model for detecting Behçet's disease. To obtain the most useful genetic identifiers, we determined distinctive SNPs using feature selection methods. Then, we employed these features (SNPs) as input variables in machine learning models, which are trained for deciding whether the samples are healthy or unhealthy. Furthermore, we proposed a method called Domain Knowledge based Subset Selection (DKSS), which utilizes information from the identified disease associated gene sub-networks to select representative SNPs. Comparing various feature selection methods, we showed that finding the right subset of SNPs plays an important role to predict the disease state accurately.

2 Materials and Methods

2.1 Problem Definition

Given a set (S) of n SNPs genotyped in m samples (case/control subjects) $S = \{s_1, s_2, \dots, s_n\}$, where n equals to the number of SNPs genotyped in a GWAS study of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2021.3053429, IEEE/ACM Transactions on Computational Biology and Bioinformatics

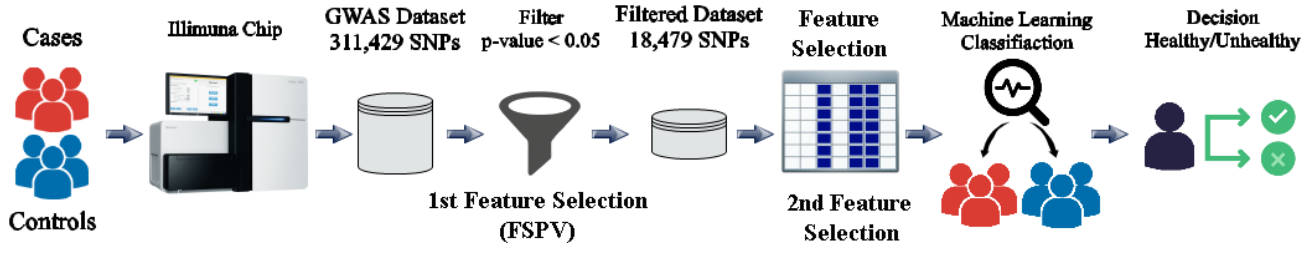


Fig. 1. Our flowchart to predict Behçet's Disease

Behçet's Disease, the goal is to find a subset of S, which includes the most relevant SNPs and to predict whether a sample has disease or not using the selected SNPs. The simplest way to solve this problem is to follow a brute force approach on all SNPs and evaluate all possible SNP subsets. Unfortunately, the number of SNPs in a GWAS is around a million or on the order of a few hundred thousands, which makes this approach computationally expensive. One solution to overcome this problem is to carry out feature selection methods over all available SNPs and to obtain a near-optimal subset in a reasonable running time.

Our main goal is to decrease the number of SNPs as much as possible, while keeping the accuracy of Behçet's disease prediction on an acceptable scale. Therefore, following the feature selection process, we need an algorithm, which takes the optimal SNP subset as the input and makes a decision on whether the tested person is affected from the disease or not. In this regard, we utilized several state-of-the-art machine learning algorithms. The flow of our disease prediction system is summarized in Figure 1.

2.2 Dataset

The GWAS dataset of Behçet's Disease consists of 1215 affected and 1278 unaffected (control) samples from Turkish population [7]. Human CNV370-Duo v1.0 and Human CNV370-Quad v3.0 chips had been used to type DNA samples. Thereafter, following strict quality control standards, SNPs were refiltered using the call rate (>95%), minor allele frequency (>1%) and Hardy-Weinberg equilibrium (> 0.00001) criterias, and finally 311,459 SNPs were obtained. For each SNP, the dataset included a genotypic p-value, which indicates the significance of a SNP for the disease. These genotypic p-values are calculated via comparing the genotypic frequencies of SNPs among cases and controls. A chi-square test was performed to obtain the p-values in GWAS analysis of BD.

2.3 Data Preprocessing

Data preprocessing is one of the most important steps of knowledge extraction [23]. In the raw form of our data set, each SNP reading can take one of the following four values, i.e., '*A_A*', '*B_B*', '*A_B*', or '*?_?*'. These represent the type of variant, i.e., homozygous reference, homozygous variant, heterezygous or unknown, respectively. In addition, due to bit read errors, some SNPs may not be matched to any zygosity. These unread SNPs are also specified as '*?_?*', denoting unknown (i.e. missing values).

In the data set that is obtained after p-value filtering (i.e. by FSPV method explained in Section 2.4.1), the number of samples is 2,493, the number of SNPs is 18,479, number of *A_A* values is 14,082,566 (30.569%), number of *B_B*'s is 15,924,781 (34.568%), number of *A_B*'s is 16,016,797 (34.768%), and number of *?_?*'s is 44,003 (0.0955%). Note that the ratio of *?_?*'s is quite low. In the data set obtained after Info Gain feature selection method, there are 19,864,000 SNPs (2493 samples and 8,000 SNPs) and 21,119 of them are missing. This gives a ratio of 0.1% for the missing values, which is also low.

In this work, class values of case and control are converted to 1 and 0, respectively. Each of the non-numeric SNP readings are converted to numeric values such that '*A_A*' is mapped to 0, '*A_B*' to 1, '*B_B*' to 2, and '*?_?*' to 3. The mapping of '*?_?*' to 3 does not cause any problems for prediction models due to the scarcity and randomness of these missing values. To demonstrate this, we include Supplementary Figures E-F. Supplementary Figure E contains a bar chart about the number of samples that contain the number of missing SNPs in a given range. Although there are a few samples that have more than 200 missing SNPs, the majority of samples have lower than 25 missing SNPs. In addition to the ratio of SNPs with missing values we also analyzed how they are distributed. Supplementary Figure F shows a heatmap of samples versus SNPs (for the set of 18,479 SNPs), in which the yellow dots represent missing values. Similarly, Supplementary Figure H contains a heatmap for Info Gain feature selection method. Based on these two figures, except for a handful of SNPs and data samples, the distribution of missing values is random (i.e. there is no preference for particular SNPs for containing missing values). Supplementary Figure G illustrates the number of samples with missing values with respect to individual SNPs. In this figure, the number of SNPs that do not contain any missing values is 8,666. Based on these observations, the distribution of missing values is random and the missing values do not have a noticeable effect on prediction results.

2.4 Feature Selection (FS) Methods

Feature selection (FS) methods are widely used to remove irrelevant and redundant features. Hence they help to reduce the number of dimensions and may improve the accuracy of prediction. In this paper, we apply a two-step feature selection strategy. In the first step, we select features by filtering SNPs with respect to their p-values applying a p-value threshold (i.e. eliminating those SNPs that have p-values higher than the threshold). In the second step, we attempt to reduce the feature set further

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2021.3053429, IEEE/ACM Transactions on Computational Biology and Bioinformatics

using a second feature selection method, which is either a machine learning based approach that performs combinatorial optimization or a biologically driven method that uses sub-network information.

2.4.1 Feature Selection by p-value information of SNPs (FSPV)

Genotypic p-values (GWAS p-value) represents the significance of odds ratio about putative disease associated variant (a measure to assess whether it could happen due to random chance) [24]. In GWA studies, although the traditional stringent p-value cutoff is $5 * 10^{-8}$, it is reported in the literature that the p-value lower than 0.05 indicates a mild relation between a SNP and the disease [25]. Consequently, the genotypic p-value threshold was selected as 0.05 and the number of SNPs decreased to 18,479. As a result of this filtering procedure, the dataset included 2493 samples and 18,479 features, which is used to test our models. The distribution of the genotypic p-values of SNPs genotyped in Behçet's Disease GWAS study is shown in Table 1, in which the sum of the second column gives 18,479.

TABLE 1
DISTRIBUTION OF THE P-VALUES OF SNPS IN BD GWAS

Range of P-value	Number of SNPs
$10^{-46} - 10^{-43}$	2
$10^{-43} - 10^{-40}$	2
$10^{-40} - 10^{-25}$	2
$10^{-25} - 10^{-20}$	5
$10^{-25} - 10^{-15}$	3
$10^{-15} - 10^{-10}$	34
$10^{-10} - 10^{-5}$	170
$10^{-5} - 10^{-4}$	142
$10^{-4} - 10^{-3}$	506
$10^{-3} - 10^{-2}$	3526
$10^{-2} - 10^{-1}$	14087

2.4.2 Feature Selection by Machine Learning and Biological Sub-Network Information

This section explains the feature selection methods that are employed in the second feature selection step. Machine learning based FS methods can be grouped into three main categories: filtering, wrapping and embedding. Filtering approaches assign a score to each feature and rank them to find the optimal feature subset by evaluating each feature or feature subset using different measures such as information, similarity, or correlation. Then, these ranked features with scores lower than the threshold are eliminated. In our experiments, we employed CFS [26], ReliefF [27], Fisher score [28], trace ratio [29], f-score [30], t-score [31], gini index [32], information gain [33], gain ratio [34], robust feature selection (RFS) [35] and, chi-square [36] filtering methods.

Wrapper methods employ a learning algorithm and build classifiers each time when a feature subset has to be evaluated. Then, the features are selected depending on the prediction performance. For this reason, wrapper based feature selection methods are classifier-dependent. In this paper, we used a wrapper feature selection method that uses logistic regression as the classifier.

Embedding FS methods perform variable selection in the process of training and they are usually specific to the learning algorithm. Tree-based and Lasso models are most known embedding methods. Tree methods compute impurity-based feature importance to discard irrevelant features. On the other hand, Lasso uses penalty with the L1 norm to find features with non-zero coefficients. Among Embedding FS methods, we used Lasso (with logistic regression as the classifier) and Extra Tree classifier based feature selection method [37].

Note that our original feature values are categories which are converted to integer values. Gain ratio, information gain, chi-square, and gini index methods require the features to be discrete. Therefore, these methods treat the features as discrete-valued (i.e. integer values are treated as categories by these algorithms). Among the remaining methods, Fisher score, RFS, Fscore, ReliefF, t-score, trace ratio, CFS, and embedding methods can handle numeric valued features. Therefore, these methods treat integer values as real valued numbers.

FS methods are usually accompanied by a search algorithm which is used to sample the features subsets. Best First Search [38], Genetic Algorithm [39], and Greedy [40] are the search methods that are employed in this work for CFS attribute subset evaluator. For the remaining filtering methods, the features are ranked by the evaluator and then top 8000 features are selected because the DKSS method (explained below) on average selected 8000 features. The search algorithms used in wrapper method are BestFirst and Greedy. The embeding methods implemented in this study do not use a search algorithm but apply a threshold to feature importance weights (i.e. select features with weight greater than the threshold).

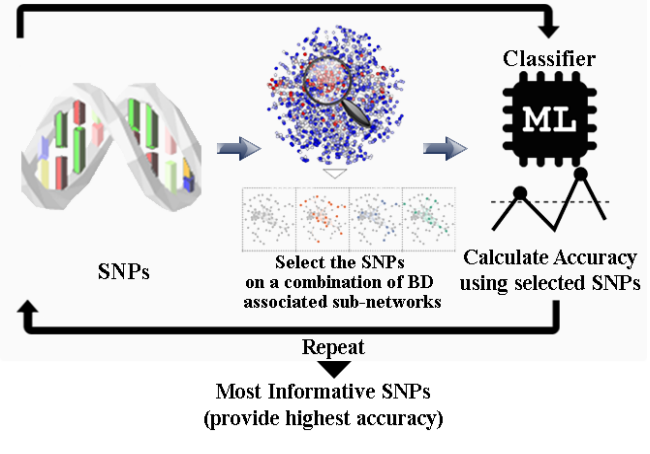


Fig. 2. Steps of Domain Knowledge based Subset Selection

Besides these well-known methods, we propose another method called Domain Knowledge based Subset Selection (DKSS), which utilizes information from disease associated sub-networks to select SNPs. Biological net-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2021.3053429, IEEE/ACM Transactions on Computational Biology and Bioinformatics

work is a graphical representation of genes, proteins or other biological molecules that include physical / functional interactions and help to understand cellular processes and disease mechanisms [41]. An active subnetwork is the connected subgraph of a biological network that has high total significance of genotypic p-values computed from disease-predisposing SNPs. The genes which are located in an active sub-network and their interactions can help to understand the mechanisms of disease development. Thus, active sub-networks might be utilized to predict disease. The idea behind DKSS is that a SNP is selected if the gene, which is related to that SNP, is included in active sub-network of BD. The first step of DKSS is to define the minimum and the maximum number of sub-networks that are going to be selected. Then, the active sub-networks are randomly picked from the sub-network pool and the SNPs related to the genes that are located on the picked sub-network are selected. Using a base algorithm, the classification score of samples with selected SNPs is calculated. This process is repeated 100 times, where the number of trials can be specified as a parameter, and the SNPs that give the highest classification score are held as the final set of SNPs as seen in Figure 2. This way, the biological information contained in active subnetworks of BD, their associated proteins, genes and SNPs is integrated with the statistical methods. In this work, we used the active sub-networks that are identified in [42] for Behçet's Disease.

2.5 Classification Methods

Classification methods can employ a set of SNPs as input features and disease phenotype(s) as output to train models that predict disease status given inputs [43]. In this work, we employed Logistic Regression, Support Vector Machine, Random Forest, k Nearest Neighbors [44], Voting Ensemble [45] and XGBoost [46] as the classification methods where random forest, voting ensemble and XGBoost are the ensemble methods, which aggregate multiple learners to obtain a combined model that may outperform its base learners [47].

3 Experiments and Results

3.1 Software

WEKA software is used [48] for implementing CFS, gain ratio, and information gain as attribute evaluators and best-first, genetic, greedy as search methods. The wrapper feature selection method is also implemented by WEKA's WrapperSubsetEvaluator class. The embedding methods (i.e. Lasso and Extra Trees classifier) are implemented using the SelectFromModel class of Python's scikit-learn library. For the remaining feature selection methods mentioned in Section 2.4.2, Python's Scikit-Feature library is employed [49]. Implementation of classification models are performed using Python's scikit-learn library [50].

3.2 Cross Validation and Repeated Hold-out

Training and testing a model on the same data causes bias in which the model learns distinctive parameters from the repeating samples and obtain high classification score. However, when models are used to predict yet-unseen data they may fail to estimate the output class accurately, which is known as "overfitting". To avoid this situation and to obtain a meaningful estimate of the prediction accuracy, k fold cross-validation can be used. Cross validation is a simple and efficient approach widely used by machine learning community and provides unbiased performance evaluation [51].

In our experiments, first, a 10-fold cross validation is performed. For this purpose, our dataset is shuffled randomly and then split into 10 different subsets. One fold is saved as test set for the final models, and the rest is used as the train set. The steps of the cross-validation experiment are summarized in Figure 3. The FSPV method (i.e. first feature selection step) is applied to features directly before cross-validation (i.e. before data set is split into train and test sets). Feature selection methods in second feature selection step are applied separately for each train set of the cross-validation experiment. Then from each train set, 20% of the samples are chosen randomly in order to generate a validation set and 40% are saved as training set for optimizing the hyper-parameters of the models.

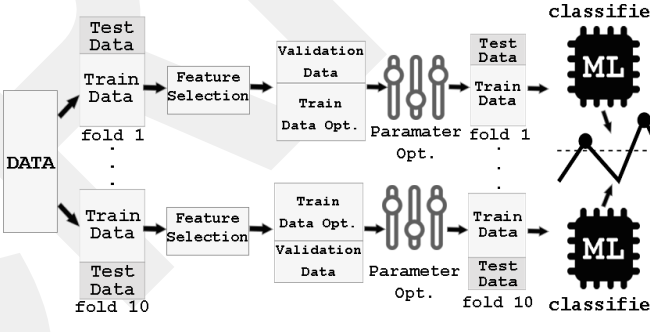


Fig. 3. Cross Validation and Optimization Steps

As a result, four different sets are obtained for each fold: train set, test set, optimization train set and optimization test set (i.e. validation set) as seen in Figure 3. Selecting a subset of train set is done to speed up the hyper-parameter optimization step. Once the best hyper-parameters are found, models are trained using the optimums on train set and evaluated on test set for each fold of cross-validation.

In addition to cross-validation, we also performed a repeated hold-out experiment by randomly selecting 20% of the dataset as test set and saving the remaining data as train set. This process is repeated a total of six times and average as well as standard deviation of the accuracy values are obtained for various feature selection and classification methods. Similar to cross-validation, feature selection is performed on train sets. The hyper-parameters of the models are optimized by performing a 10-fold cross-validation on train sets. The results of repeated hold-out experiments can be found in Supplementary Table C. These results are similar to those obtained in cross-validaton experiments reported in Tables 3-9.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2021.3053429, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE 2
AVERAGE NUMBER OF SELECTED FEATURES OF 10-FOLD CROSS-VALIDATION

Others	LR Lasso	Extra DT	CFS Bestfirst	CFS Greedy	CFS Genetic	Wrapper Greedy	DKSS
8000	1983.6	7760.1	199	201.4	3926.7	27	7999

3.3 Hyper-Parameter Optimization

Obtaining high prediction performance from machine learning methods depends crucially on proper tuning of the hyper-parameters. In this work a separate optimization is performed for each fold of of cross-validation using the optimization train set and the validation set. For this purpose, a given prediction model is trained on optimization train set for each hyper-parameter configuration and predictions are computed on validation set. Finally, the particular hyper-parameter setting that gives the best accuracy on validation set is selected as the optimum. This approach is similar to nested cross-validation and reduces bias as much as possible because the optimizations are performed within training sets without using any samples from test sets.

There are different approaches for sampling the parameter configurations during hyper-parameter optimization. In this study, we used Bayesian optimization as the search algorithm. Bayesian optimization is a type of non-linear optimization procedure where the parameter value to explore in each step is decided based on the analysis of a distribution over a function such as a Gaussian process or other surrogate models [52]. Unlike the grid search, it uses hyper-parameters in the previous step to find the hyper-paremater set in the next step.

Compared to other techniques, it typically requires lower number of iterations to find the optimum parameters. The regularization parameter (i.e. C) of SVM and LR is optimized by sampling values from $2^{(-25)}$ to 2^{25}, the number of trees parameter of random forest is optimized by sampling values from 10 to 800, the number of neighbors parameter of k-NN is optimized by using the same parameter interval as random forest. Parameters of XGBoost are also tuned using the Bayesian optimization technique. In this work, learning rate (eta), depth of tree (max_depth), number of estimators and gamma regularization parameter of XGBoost are tuned. Furthermore, the number of iterations and early stopping values are set to 3000 and 50, respectively. Due to the high number of features, linear models are more suitable for our data. Therefore, we preferred linear booster for XGBoost.

3.4 Number of Features Selected

Correlation-based Feature Subset Selection (CFS) method selects features automatically using correlation information and therefore does not allow specifying the number of features that are going to be selected as a hyper-parameter. For the rest of the methods, this parameter can be set as input. The proposed DKSS method chooses the SNPs from the selected sub-networks and on average 8,000 SNPs are selected across the 10 folds of cross-

TABLE 3
LOGISTIC REGRESSION RESULTS

LR	ACC	AUC	AUPRC	Train Time	Test Time
CFS BestFirst	64.58%	70.83%	68.47%	64.306	0.090
CFS Genetic	96.87%	99.19%	98.94%	186.925	1.305
CFS Greedy	65.10%	70.75%	68.19%	73.354	0.091
Chi Square	88.41%	95.83%	95.78%	236.401	2.406
DKSS	96.91%	99.30%	99.10%	279.953	2.689
Extra DT	97.43%	99.46%	99.39%	274.086	2.540
F Score	76.66%	82.30%	80.74%	145.748	1.092
Fisher Score	76.41%	83.19%	81.67%	149.273	1.034
FSPV	**99.56%**	**99.99%**	**99.98%**	554.026	5.770
Gain Ratio	97.67%	99.51%	99.46%	361.083	2.752
Gini Index	84.72%	91.75%	91.06%	327.397	2.802
Information Gain	98.47%	99.63%	99.62%	396.885	4.123
LR Lasso	76.54%	81.67%	79.90%	103.278	0.445
ReliefF	97.27%	99.43%	99.38%	236.141	1.739
RFS	93.82%	97.43%	97.25%	137.530	0.865
T Score	76.90%	82.62%	81.26%	195.870	1.234
Trace Ratio	76.34%	83.02%	80.90%	154.848	1.053
Wrapper Greedy	64.58%	69.27%	67.41%	66.283	0.056

TABLE 4
SUPPORT VECTOR MACHINES RESULTS

SVM	ACC	AUC	AUPRC	Train Time	Test Time
CFS BestFirst	64.62%	71.48%	69.62%	83.238	0.949
CFS Genetic	96.83%	99.59%	99.55%	187.877	3.786
CFS Greedy	65.58%	71.80%	70.90%	84.269	1.114
Chi Square	88.77%	95.81%	95.73%	173.679	4.820
DKSS	96.83%	99.54%	99.50%	316.216	7.614
Extra DT	97.39%	99.71%	99.69%	305.412	9.777
F Score	76.74%	84.81%	84.12%	328.877	11.771
Fisher Score	76.62%	84.33%	83.57%	293.585	11.333
FSPV	**99.56%**	**99.99%**	**99.99%**	556.862	21.810
Gain Ratio	97.95%	99.75%	99.73%	357.367	12.082
Gini Index	85.24%	92.21%	92.06%	264.927	8.238
Information Gain	97.87%	99.73%	99.70%	293.447	9.244
LR Lasso	76.25%	84.52%	83.82%	209.474	7.839
ReliefF	97.35%	99.63%	99.59%	316.088	10.145
RFS	93.70%	99.06%	99.04%	295.368	10.834
T Score	77.06%	84.62%	83.80%	289.787	10.110
Trace Ratio	76.50%	84.60%	83.74%	331.891	11.504
Wrapper Greedy	64.78%	69.51%	67.57%	64.931	0.314

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2021.3053429, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE 5
RANDOM FOREST RESULTS

RF	ACC	AUC	AUPRC	Train Time	Test Time
CFS BestFirst	66.19%	72.18%	69.87%	242.851	2.054
CFS Genetic	74.21%	82.99%	81.86%	997.010	12.234
CFS Greedy	66.87%	72.59%	70.18%	258.676	2.144
Chi Square	69.60%	76.93%	75.76%	1443.386	16.277
DKSS	73.69%	82.20%	81.96%	1407.542	19.635
Extra DT	71.36%	80.56%	79.46%	1432.645	16.974
F Score	67.55%	73.72%	72.90%	1491.833	15.910
Fisher Score	66.95%	73.37%	71.35%	1308.551	15.021
FSPV	74.85%	**84.49%**	**83.90%**	2266.199	30.690
Gain Ratio	73.32%	82.55%	81.60%	1556.403	21.644
Gini Index	68.71%	76.33%	74.52%	1546.853	21.355
Information Gain	**75.61%**	82.35%	81.39%	2193.064	32.227
LR Lasso	68.55%	74.51%	73.06%	622.732	6.572
ReliefF	71.12%	80.12%	79.43%	1574.588	15.852
RFS	73.16%	80.84%	80.94%	1609.311	23.488
T Score	67.87%	73.36%	71.61%	1460.893	16.150
Trace Ratio	67.23%	73.23%	71.20%	1580.726	20.729
Wrapper Greedy	62.66%	65.70%	63.43%	205.979	1.070

TABLE 6
K-NN RESULTS

KNN	ACC	AUC	AUPRC	Train Time	Test Time
CFS BestFirst	65.95%	71.56%	69.73%	72.971	0.257
CFS Genetic	85.68%	94.43%	93.52%	546.888	5.003
CFS Greedy	65.22%	71.22%	69.16%	80.582	0.259
Chi Square	76.33%	86.78%	84.89%	1147.815	10.976
DKSS	86.88%	96.77%	96.56%	988.006	9.567
Extra DT	87.44%	95.17%	94.69%	991.586	9.487
F Score	66.26%	76.52%	74.53%	1114.953	10.767
Fisher Score	68.03%	78.45%	75.99%	1015.376	10.037
FSPV	**93.26%**	**99.14%**	**99.06%**	2189.197	21.522
Gain Ratio	82.91%	92.91%	90.65%	1161.505	11.394
Gini Index	72.00%	83.32%	80.78%	1111.711	10.712
Information Gain	93.22%	93.42%	91.88%	2013.611	20.061
LR Lasso	74.69%	82.22%	80.45%	268.013	2.295
ReliefF	88.49%	95.17%	94.60%	1105.847	11.182
RFS	88.81%	97.02%	96.82%	1082.181	11.125
T Score	66.26%	77.05%	74.70%	1033.935	10.274
Trace Ratio	68.15%	78.90%	77.52%	1134.333	11.150
Wrapper Greedy	63.58%	67.13%	64.72%	68.499	0.092

TABLE 7
XGBOOST RESULTS

XGB	ACC	AUC	AUPRC	Train Time	Test Time
CFS BestFirst	63.95%	70.89%	69.26%	126.594	0.431
CFS Genetic	95.44%	98.37%	98.18%	185.490	0.463
CFS Greedy	64.82%	70.52%	68.87%	455.752	2.208
Chi Square	92.31%	97.00%	97.05%	876.370	7.832
DKSS	96.15%	98.81%	98.79%	1487.282	13.540
Extra DT	96.32%	98.86%	98.84%	759.222	8.781
F Score	74.30%	81.67%	80.83%	941.374	10.675
Fisher Score	73.87%	79.86%	78.76%	949.713	13.009
FSPV	**99.12%**	**99.92%**	**99.92%**	1716.624	13.000
Gain Ratio	96.64%	99.24%	99.10%	700.043	3.513
Gini Index	82.28%	88.92%	88.08%	828.508	11.133
Information Gain	96.32%	98.78%	98.68%	821.819	10.293
LR Lasso	74.37%	83.24%	83.36%	317.972	2.031
ReliefF	95.66%	99.09%	99.14%	849.870	7.929
RFS	91.40%	97.11%	97.00%	944.679	6.570
T Score	74.96%	81.12%	80.61%	738.165	13.171
Trace Ratio	75.60%	81.02%	80.24%	1097.323	9.731
Wrapper Greedy	64.66%	69.32%	67.34%	215.749	0.144

validation (feature selection methods are applied separately on each training set of cross-validation experiment as explained in Section 3.2). To be able to make a fair comparison between DKSS and the FS methods that allow entering the number of features as input (i.e. those excluding the CFS method), the first 8,000 features are selected as the best representative subset of SNPs. After feature selection step, data is processed by machine learning algorithms. Table 2 lists the averages of the number of features selected on train sets of 10-fold cross-validation for each feature selection method. In this table, the "Other" category includes FS methods that rank the features according to a metric. These include Fisher score, Fscore, gini index, information gain, chi-square, relief, T-score, RFS, and trace ratio. Detailed numbers of features selected by each method is given in Supplementary A for each fold of cross-validation.

3.5 Diagnostic Prediction Accuracy of Classifiers

The following performance metrics are used to evaluate the success of the classificaton methods: the overall accuracy (Acc), area under the ROC curve (AUC), and area under the precision and recall curve (AUPRC). For these metrics, the averages across the 10 folds of cross-validation are computed. Besides, training and test times of final models are computed by neglecting the data read times.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2021.3053429, IEEE/ACM Transactions on Computational Biology and Bioinformatics

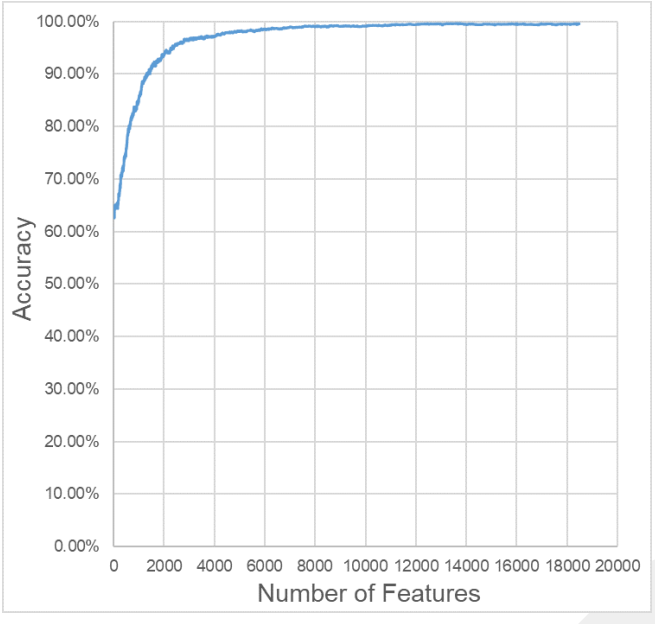


Fig. 4. Accuracy of logistic regression with respect to number of features selected by p-value based feature selection method

Tables 3-8 show the diagnostic prediction accuracies of logistic regression, support vector machines, random forest, k-Nearest neighbor, XGBoost, and voting ensemble algorithms, respectively as well as the train and test times of the models reported in seconds. Default values are used for the parameters of the algorithms except for the parameters mentioned in section 3.3, which are optimized. Linear kernel is employed for SVM. *"FSPV"* row of tables represents dataset that has 18,479 features when features are selected using p-value information only at the first feature selection step (without applying a second feature selection method).

When results are compared, it's seen that more than 99% of the samples are predicted correctly, i.e. as healthy or affected by LR, SVM, voting ensemble and XGBoost. Voting ensemble method has two different approaches for combining the predictions made by the base learners: hard voting and soft voting. Hard voting decides the class by taking the majority of the classes predicted by the base learners. On the other hand, soft voting computes the average of prediction probability scores for each class obtained by the base learners. To reduce the misclassification rate of the ensemble, we used the soft voting approach and combined LR and SVM methods only (as base learners) because RF and k-NN methods obtained lower accuracy scores, which could have decreased the overall accuracy of voting. Despite obtaining close results, voting ensemble approach did not perform better than LR and SVM methods alone. The other ensemble method, XGBoost, obtained comparable scores among all the methods compared.

Among feature selection methods FSPV alone performed the best reaching 99.56% accuracy and 99.99%

TABLE 8
ENSEMBLE VOTING RESULTS

VOTING	Acc	AUC	AUPRC	Train Time	Test Time
CFS BestFirst	64.70%	71.34%	69.36%	NA	NA
CFS Genetic	96.83%	99.52%	99.37%	NA	NA
CFS Greedy	65.50%	71.59%	70.23%	NA	NA
Chi Square	88.45%	95.88%	95.81%	NA	NA
DKSS	96.87%	99.47%	99.37%	NA	NA
Extra DT	97.59%	99.63%	99.60%	NA	NA
F Score	76.29%	83.96%	82.86%	NA	NA
Fisher Score	76.53%	84.25%	83.45%	NA	NA
FSPV	**99.56%**	**99.99%**	**99.99%**	NA	NA
Gain Ratio	97.83%	99.67%	99.63%	NA	NA
Gini Index	84.52%	92.24%	91.85%	NA	NA
Information Gain	97.95%	99.74%	99.72%	NA	NA
LR Lasso	76.69%	83.16%	81.49%	NA	NA
ReliefF	97.27%	99.63%	99.61%	NA	NA
RFS	94.34%	98.70%	98.68%	NA	NA
T Score	76.94%	83.93%	82.77%	NA	NA
Trace Ratio	76.49%	84.41%	83.65%	NA	NA
Wrapper Greedy	64.74%	69.41%	67.52%	NA	NA

TABLE 9
STANDARD DEVIATIONS OF 10-FOLD ACCURACIES

Method	LR	SVM	KNN	RF	XGB
CFS BestFirst	2.430%	2.802%	2.532%	2.07%	2.007%
CFS Genetic	1.044%	1.222%	1.697%	2.53%	2.491%
CFS Greedy	2.515%	1.964%	2.911%	2.50%	1.755%
Chi Square	1.255%	1.401%	2.854%	2.55%	15.444%
DKSS	0.896%	0.702%	2.084%	1.864%	0.821%
Extra DT	0.988%	1.155%	1.951%	2.63%	1.238%
F Score	1.753%	1.633%	2.675%	2.01%	2.281%
Fisher Score	1.423%	1.013%	2.397%	2.53%	3.925%
FSPV	0.333%	0.334%	1.765%	1.74%	0.561%
Gain Ratio	0.913%	0.773%	3.165%	2.33%	0.989%
Gini Index	2.031%	1.845%	2.443%	2.92%	2.173%
Information Gain	0.950%	0.826%	1.910%	2.50%	1.338%
LR Lasso	2.315%	1.617%	2.496%	2.23%	1.913%
ReliefF	1.055%	1.135%	1.588%	2.38%	1.164%
RFS	1.142%	1.392%	2.282%	2.02%	1.992%
T Score	1.445%	1.395%	1.732%	2.37%	2.136%
Trace Ratio	1.691%	1.695%	1.950%	2.12%	2.448%
Wrapper Greedy	1.312%	1.460%	2.028%	2.73%	1.447%

AUC and AUPRC scores. When the results of feature selection methods at the second feature selection step are compared, decision tree based methods (Extra DT, Infor-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2021.3053429, IEEE/ACM Transactions on Computational Biology and Bioinformatics

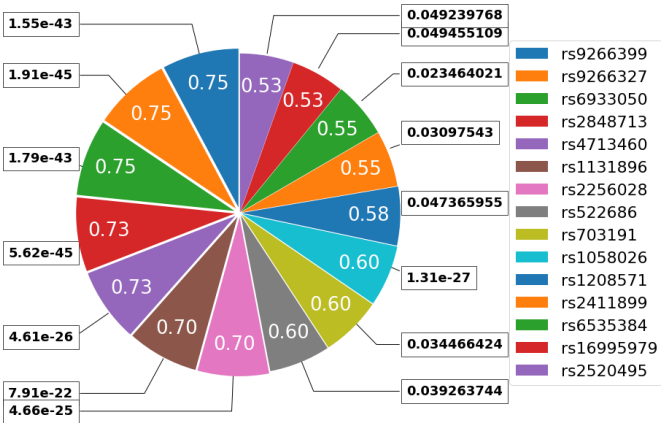


Fig. 5. Most representative SNPs that are identified by feature selection methods. Their genotypic p-values are shown in boxes. Numbers on the slices of the pie chart represent occurrence rates.

mation gain, Gain Ratio) take the lead. They are followed by DKSS, Genetic search based CFS and RFS methods respectively. As a consequence of selecting less number of SNPs, methods that use Best First/Greedy Search based CFS and Wrapper obtained the lowest prediction accuracies.

The fact that the best prediction accuracy is obtained by FSPV as compared to FSPV followed by a second feature selection method may indicate that all of the 18,479 SNPs have a contribution to disease prediction. To test this hypothesis, we performed the following experiment. We ranked features by p-value and performed a forward feature selection strategy (by increasing the number of features by 1 at each step) in the first feature selection step and did not perform the second feature selection step. We evaluated the accuracy of these feature subsets using logistic regression. Figure 4 shows the disease detection accuracy (obtained as the average of the accuracies in 10-fold cross-validation) with respect to the number of features selected. Based on this figure, the maximum accuracy of 99.64% is obtained when the number of features is equal to 13,611. This accuracy is also obtained for some of the higher number of features. This shows that it is not mandatory to use all of the 18,479 features to get the best classification accuracy and FS methods that employ CFS as the subset evaluator and embedding methods are conservative in terms of the number of features selected. For the ranker-based methods, the number of features was set to 8000 to perform a fair comparison with the DKSS method, which uses biological sub-network information. If the number of features selected by ranker-based methods are allowed to be any other value it may be possible to find a feature subset that contains a lower number of features. We leave exploring the feature space further as future work. In this paper, the FS methods used in the second feature selection step (though not giving the best accuracies) allowed us to find which SNPs are frequently selected by performing cross-validation experiments (i.e. which SNPs are important).

Table 9 shows the standart deviations of 10-fold test accuracies. It can be stated that a 2% variance is observed in test accuracies as a result of deviations in train sets of cross-validation.

3.6 Which SNPs are more Important

Besides information about prediction score, we analyzed the most represented SNPs. The occurrence rate of a given SNP is computed as follows. Firstly, the top 25 SNPs that ranked as highly related by each of the best four feature selection methods are determined for each fold. This produces 40 lists in total, where each list includes 25 SNPs. To find the most represented SNPs, we used extra decision tree, gain ratio, information gain and reliefF as the best feature selection methods. Secondly, the number of times that the given SNP appears in top 25 lists of 10-fold cross-validation and for the best four feature selection methods is identified. The occurrence rate is computed as the number of times the SNP appears in top 25 lists divided by 40. For instance, if a given SNP appears in all of the top 25 lists, then the occurrence rate becomes 1 and if it never appears in those lists then the rate becomes 0. Figure 5 shows the occurrence rate and the genotypic p-values of the top 15 SNPs selected by the most successful four feature selection methods. Numbers on the slices of the pie chart represent occurrence rates. Detailed information about top SNPs is given in supplementary B.

We applied further analysis over these top 15 most represented SNPs. As a result, we have observed that 7 of the top 15 SNPs are associated with 6 different genes using SPOT tool, which is also used in [42] to map the SNPs to genes and then to find associated active subnetworks [53]. These genes are HLA-B (rs1058026), HCP5 (rs1131896, rs2848713), KIRREL3 (rs522686), LAMP5-AS1 (rs16995979), MICA (rs2256028) and SCD5 (rs6535384). Note that, HLA-B genes were also found to have a strong association with BD in literature [7]. We also checked for the associated pathways and GO terms with these genes. They are not associated with KEGG biological pathways, but associated with antigen processing and presentation, defense response, regulation of immune response GO Biological Process terms; integral component of membrane, cell surface GO Cellular Component terms and antigen binding GO Molecular Function terms.. 7 variants (rs1058026, rs522686, rs6535384, rs1131896, rs2256028, rs9266399, rs2848713) are located in intronic regions. rs9266399, rs6933050, rs4713460, are located in 5′ upstream regions of DHFRP2, FGFR3P1, ZDHHC20P2 genes respectively, according to UCSC annotation [54].

4 Discussions and Conclusions

Precision medicine is a rapidly advancing field that provides personalized treatments and preventive interventions to the patients. Especially it may be very helpful for patients who are affected from auto-immune diseases. In order to realize that, each patient should be handled and analyzed individually. In this regard, using patients' SNPs to predict disease risk could be essential. But, scanning all SNPs manually to diagnose disease is very time consuming and is not practical. Furthermore, the diagnosis of some diseases such as Behçet's disease could be difficult, since their clinical signs or symptoms can also be

seen in other diseases. Therefore, there is a need for an advanced system that facilitates the diagnosis process.

In this study, we attempt to generate a classification model for Behçet's disease that predicts the disease status of a given sample using the genotyped SNP information. In our experiments, the models that use features selected by p-value information only classified almost all affected samples correctly with 99% AUC score and it can be stated clearly that genotypic p-value information has high capacity to indicate disease-causing SNPs.

In addition to extracting the SNPs that are selected by p-values and those that are found by the second feature selection step, we also analyzed the coefficients of SNPs for logistic regression with L2 norm penalty and number of SNP features that are used by random forest model. In logistic regression all SNPs had non-zero ceofficients. The number of SNP features employed in random forest models are given in Supplementary Table D for each feature selection method and for each fold of cross-validation. Based on this analysis the logistic regression and random forest models use all the features available (i.e. those obtained after feature selection step). This shows all SNPs that are input to prediction models contribute to the decision for diagnosing the BD.

In our study, dimension reduction methods such as PCA or Autoencoder are not experimented to decrease the number of SNPs. Because, these methods use linear or non-linear combination of features to reduce the dimensionality of the feature space. Therefore, the features that appear in new-low dimensional space do not refer to SNPs, but indefinite combination of SNPs. However, the information of which SNPs are effective or explanatory to diagnose the disease is lost. Additionally, performing dimension reduction needs all features to be processed. However, one of the goals of our study is to keep accuracy as sufficiently high, while obtaining efficiency in terms of computational cost by processing much fewer number of SNPs. For all these reasons, dimension reduction methods are not included in this study.

As a future work, we are planning to apply the methods developed in this work to other disease detection problems that use GWAS or gene expression data.

ACKNOWLEDGMENT

The numerical calculations reported in this paper were partially performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

REFERENCES

[1] F. Davatchi *et al.*, "Behcet's disease: epidemiology, clinical manifestations, and diagnosis," *Expert Rev Clin Immunol*, vol. 13, no. 1, Art. no. 1, Jan. 2017, doi: 10.1080/1744666X.2016.1205486.

[2] "Behçet's disease," *nhs.uk*, Oct. 20, 2017. https://www.nhs.uk/conditions/behcets-disease/ (accessed Sep. 14, 2020).

[3] Y. Görmez, Y. E. Işık, and B. Bakır-Güngör, "The Identification of Discriminative Single Nucleotide Polymorphism Sets for the Classification of Behçet's Disease," in *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, 2018, pp. 443–447.

[4] E. Alpsoy, "Behçet Hastalığı: Etyopatogenezde Güncel Bilgiler.," *Turkish Journal of Dermatology*, vol. 7, no. 1, Art. no. 1, 2013.

[5] G. H. Reference, "What are single nucleotide polymorphisms (SNPs)?," *Genetics Home Reference*, Jan. 26, 2019. https://ghr.nlm.nih.gov/primer/genomicresearch/snp (accessed Jan. 27, 2019).

[6] D. Shriner and C. N. Rotimi, "Whole-Genome-Sequence-Based Haplotypes Reveal Single Origin of the Sickle Allele during the Holocene Wet Phase," *The American Journal of Human Genetics*, vol. 102, no. 4, Art. no. 4, Apr. 2018, doi: 10.1016/j.ajhg.2018.02.003.

[7] E. F. Remmers *et al.*, "Genome-wide association study identifies variants in the MHC class I, IL10, and IL23R-IL12RB2 regions associated with Behcet's disease," *Nature genetics*, vol. 42, no. 8, Art. no. 8, 2010.

[8] S. Hou *et al.*, "Identification of a susceptibility locus in STAT4 for Behçet's disease in Han Chinese in a genome-wide association study," *Arthritis & Rheumatism*, vol. 64, no. 12, Art. no. 12, 2012, doi: 10.1002/art.37708.

[9] Y. Huang *et al.*, "The Association of Chemokine Gene Polymorphisms with VKH and Behcet's Disease in a Chinese Han Population," *BioMed Research International*, 2017. https://www.hindawi.com/journals/bmri/2017/1274960/ (accessed Aug. 08, 2019).

[10] S. W. Kim *et al.*, "Identification of genetic susceptibility loci for intestinal Behçet's disease," *Sci Rep*, vol. 7, Jan. 2017, doi: 10.1038/srep39850.

[11] C. A. C. Montañez, P. Fergus, A. C. Montañez, and C. Chalmers, "Deep Learning Classification of Polygenic Obesity using Genome Wide Association Study SNPs," *arXiv:1804.03198 [cs, q-bio]*, Apr. 2018, Accessed: Aug. 09, 2019. [Online]. Available: http://arxiv.org/abs/1804.03198.

[12] P. Bellot, G. de los Campos, and M. Pérez-Enciso, "Can Deep Learning Improve Genomic Prediction of Complex Human Traits?," *Genetics*, vol. 210, no. 3, Art. no. 3, Nov. 2018, doi: 10.1534/genetics.118.301298.

[13] S. Okser, T. Pahikkala, A. Airola, T. Salakoski, S. Ripatti, and T. Aittokallio, "Regularized machine learning in the genetic prediction of complex traits," *PLoS genetics*, vol. 10, no. 11, Art. no. 11, 2014.

[14] O. Manor and E. Segal, "Predicting Disease Risk Using Bootstrap Ranking and Classification Algorithms," *PLOS Computational Biology*, vol. 9, no. 8, Art. no. 8, Ağu 2013, doi: 10.1371/journal.pcbi.1003200.

[15] M. Maciukiewicz *et al.*, "GWAS-based machine learning approach to predict duloxetine response in major depressive disorder," *J Psychiatr Res*, vol. 99, pp. 62–68, Apr. 2018, doi: 10.1016/j.jpsychires.2017.12.009.

[16] O. González-Recio and S. Forni, "Genome-wide prediction of discrete traits using bayesian regressions and machine learning," *Genetics Selection Evolution*, vol. 43, no. 1, Art. no. 1, Feb. 2011, doi: 10.1186/1297-9686-43-7.

[17] K. Anekboon, S. Phimoltares, C. Lursinsap, and S. Tongsima, and S.

Fucharoen, "Searching single nucleotide polymorphism markers to complex diseases using genetic algorithm framework and a BoostMode support vector machine," in *2010 4th International Conference on Bioinformatics and Biomedical Engineering*, 2010, pp. 1–4.

[18] Z. Wei *et al.*, "Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease," *Am. J. Hum. Genet.*, vol. 92, no. 6, Art. no. 6, Jun. 2013, doi: 10.1016/j.ajhg.2013.05.002.

[19] C. Kooperberg, M. LeBlanc, and V. Obenchain, "Risk Prediction using Genome-Wide Association Studies," *Genet Epidemiol*, vol. 34, no. 7, Art. no. 7, Nov. 2010, doi: 10.1002/gepi.20509.

[20] B. López, F. Torrent-Fontbona, R. Viñas, and J. M. Fernández-Real, "Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction," *Artificial Intelligence in Medicine*, vol. 85, pp. 43–49, Apr. 2018, doi: 10.1016/j.artmed.2017.09.005.

[21] C. A. C. Montañez *et al.*, "Machine learning approaches for the prediction of obesity using publicly available genetic profiles," in *2017 International Joint Conference on Neural Networks (IJCNN)*, May 2017, pp. 2743–2750, doi: 10.1109/IJCNN.2017.7966194.

[22] D. Shigemizu *et al.*, "The Construction of Risk Prediction Models Using GWAS Data and Its Application to a Type 2 Diabetes Prospective Cohort," *PLoS One*, vol. 9, no. 3, Art. no. 3, Mar. 2014, doi: 10.1371/journal.pone.0092549.

[23] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, and A. Zanasi, *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc., 1998.

[24] "GWAS 2: P-values in GWAS." https://www.mv.helsinki.fi/home/mjxpirin/GWAS_course/material/GWAS2.html (accessed Sep. 24, 2020).

[25] Y. Zhang, "On The Use of P-Values in Genome Wide Disease Association Mapping," *Journal of Biometrics & Biostatistics*, vol. 7, no. 3, Art. no. 3, 2016, doi: 10.4172/2155-6180.1000297.

[26] M. A. Hall, "Correlation-based Feature Selection for Machine Learning," 1999.

[27] M. Robnik-Šikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1, Art. no. 1, Oct. 2003, doi: 10.1023/A:1025667309714.

[28] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition)*. New York, NY, USA: Wiley-Interscience, 2000.

[29] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, "Trace Ratio Criterion for Feature Selection," in *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, Chicago, Illinois, 2008, pp. 671–676, Accessed: Jul. 14, 2019. [Online]. Available: http://dl.acm.org/citation.cfm?id=1620163.1620176.

[30] S. Wright, "The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating," *Evolution*, vol. 19, no. 3, Art. no. 3, 1965, doi: 10.2307/2406450.

[31] J. C. Davis, *Statistics and Data Analysis in Geology*, 2nd ed. New York, NY, USA: John Wiley & Sons, Inc., 1990.

[32] "Gini coefficient," *Wikipedia*. Aug. 23, 2020, Accessed: Sep. 14, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Gini_coefficient&oldid=974584849.

[33] C. Lee and G. G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization," *Information processing & management*, vol. 42, no. 1, pp. 155–165, 2006.

[34] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection," *International Journal of Information Technology and Knowledge Management*, vol. 2, no. 2, pp. 271–277, 2010.

[35] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and Robust Feature Selection via Joint \mathscrl2,1-Norms Minimization," in *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 1813–1821.

[36] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," in *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*, 1995, pp. 388–391.

[37] "1.13. Feature selection — scikit-learn 0.23.2 documentation." https://scikit-learn.org/stable/modules/feature_selection.html (accessed Sep. 14, 2020).

[38] N. Kwak and Chong-Ho Choi, "Input feature selection for classification problems," *IEEE Transactions on Neural Networks*, vol. 13, no. 1, Art. no. 1, Jan. 2002, doi: 10.1109/72.977291.

[39] W. Siedlecki and J. Sklansky, "A note on genetic algorithms for large-scale feature selection," *Pattern Recognition Letters*, vol. 10, no. 5, Art. no. 5, Nov. 1989, doi: 10.1016/0167-8655(89)90037-8.

[40] "GreedyStepwise." https://weka.sourceforge.io/doc.dev/weka/attributeSelection/GreedyStepwise.html (accessed Sep. 23, 2020).

[41] H. Nguyen, S. Shrestha, D. Tran, A. Shafi, S. Draghici, and T. Nguyen, "A Comprehensive Survey of Tools and Software for Active Subnetwork Identification," *Front. Genet.*, vol. 10, 2019, doi: 10.3389/fgene.2019.00155.

[42] B. Bakir-Gungor *et al.*, "Identification of possible pathogenic pathways in Behçet's disease using genome-wide association study data from two different populations," *Eur. J. Hum. Genet.*, vol. 23, no. 5, pp. 678–687, May 2015, doi: 10.1038/ejhg.2014.158.

[43] D. S. W. Ho, W. Schierding, M. Wake, R. Saffery, and J. O'Sullivan, "Machine Learning SNP Based Prediction for Precision Medicine," *Front. Genet.*, vol. 10, 2019, doi: 10.3389/fgene.2019.00267.

[44] C. Crisci, B. Ghattas, and G. Perera, "A review of supervised machine learning algorithms and their applications to ecological data," *Ecological Modelling*, vol. 240, pp. 113–122, Aug. 2012, doi: 10.1016/j.ecolmodel.2012.03.001.

[45] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1–2, Art. no. 1–2, 2010.

[46] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[47] Y. Ren, L. Zhang, and P. N. Suganthan, "Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article]," *IEEE Computational Intelligence Magazine*, vol. 11, no. 1, Art. no. 1, Feb. 2016, doi: 10.1109/MCI.2015.2471235.

[48] E. Frank *et al.*, "Weka-a machine learning workbench for data mining," in *Data mining and knowledge discovery handbook*, Springer, 2009, pp. 1269–1277.

[49] J. Li *et al.*, "Feature Selection: A Data Perspective," *ACM Comput. Surv.*, vol. 50, no. 6, Art. no. 6, Dec. 2017, doi: 10.1145/3136625.

[50] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, p. 2825−2830, Oct. 2011.

[51] G. C. Cawley and N. L. C. Talbot, "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, Aug. 2010.

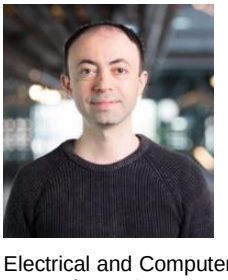
[52] R. Martinez-Cantin, "Bayesopt: A bayesian optimization library for nonlinear optimization, experimental design and bandits," *The Journal of Machine Learning Research*, vol. 15, no. 1, Art. no. 1, 2014.

[53] S. F. Saccone *et al.*, "SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study," *Nucleic Acids Res*, vol. 38, no. Web Server issue, pp. W201–W209, Jul. 2010, doi: 10.1093/nar/gkq513.

[54] W. J. Kent *et al.*, "The Human Genome Browser at UCSC," *Genome Res.*, vol. 12, no. 6, pp. 996–1006, Jun. 2002, doi: 10.1101/gr.229102.

Yunus Emre IŞIK Mr. Işık received the baclehor and master degree in Management Information Systems from the Mehmet Akif Ersoy University and Cumhuriyet university respectively. Currently a Ph.D. student in Electrical and Computer Engineering at the Abdullah Gul University and a research assistant in Department of Management Information Systems, Cumhuriyet University, Sivas, Turkey. He studies the machine learning implementation for detecting genetic and infectuous diseases.

Yasin GÖRMEZ Mr. Gormez graduated from Computer Engineering Department of Meliksah University and he received his Master of Science (M. Sc.) degrees with high honor from the Electrical and Computer Engineering Department of Abdullah Gul University in 2015 and 2017, respectively. Currently, he continues his PhD in same department and he is a research assistant in Management Information Sytems of Cumhuriyet University, Sivas, Turkey



Zafer AYDIN Dr. Aydin received his Bachelor of Science (B.Sc.) and Master of Science (M.Sc.) degrees with high honor from the Electrical and Electronics Engineering Department of Bilkent University in 1999 and 2001, respectively. He then enrolled in the PhD program of the same department and worked as a teaching assistant for one year. Starting from 2002, he worked as a Graduate Research Assistant in School of Electrical and Computer Engineering at Georgia Institute of Technology, Atlanta GA USA and received the PhD degree in 2008. As a result of maintaining an interest in bioinformatics research, he worked as a post-doctoral fellow for three years in Noble Research Lab, which is part of the Genome Sciences Department at University of Washington, Seattle, WA USA. From September 2011 to Februray 2014, he worked as an Assistant Professor in Electrical and Electronics Engineering Department of Bahcesehir University, Istanbul, Turkey. Currently he is an Assistant Professor in Computer Enginering Department of Abdullah Gul University, Kayseri, Turkey.

Burcu BAKIR-GUNGOR Burcu Bakir-Gungor received her B.Sc. degree in Biological Sciences and Bioengineering from Sabanci University; her M.Sc. degree in Bioinformatics from Georgia Institute of Technology; and her PhD degree from Georgia Institute of Technology/Sabanci University. She worked at the Bioinformatics Research Center, Medical College of Wisconsin, from 2007-2009. From 2009 to 2011, she worked at the Department of Computer Engineering, Bahcesehir University. Then, she worked as an Assistant Professor at the Department of Genetics and Bioinformatics, at the same university. From 2012 to 2013, she was part of the Advanced Genomics and Bioinformatics Research Center, UEKAE, BILGEM, TUBITAK. Now, she works as an Assistant Professor at the Department of Computer Engineering at Abdullah Gul University. She is the recipient of "Best Paper" award at the 4th EvoBIO Conference. She acted as a member of the bioinformatics advisory board of the Turkish Genome Project. She is the reviewer of several prestigious international journals including Bioinformatics, Machine Learning, Journal of Computational Biology; she is a Technical Program Committee member of SIU, UBMK and HIBIT conferences; and she is an Editorial Board member of PeerJ journal. Her research interests include bioinformatics, computational genomics, network and pathway oriented analysis of genome-wide association studies and next-generation sequencing datasets; applications of machine learning and data mining in bioinformatics