Bioengineering Research

Study of Genes Associated With Parkinson Disease Using Feature Selection

Hoda Rafieipour Azadeh [1], Abdollah Zadeh [2*], Atefeh Moradan [3], Zahra Salekshahrezaee [4]

[1] Department of Computer Science, Memorial University of Newfoundland, NF, Canada

[2] Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL, USA

[3] Department of Computer Science, Aarhus University, Aarhus, Denmark

[4] Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, FL, USA

Correspondence to: Abdollah Zadeh A. (E-mail: aabdollahzad2016@fau.edu)

Abstract

The second most prevalent age-related neurodegenerative disease is Parkinson's (PD) and Genes associated with human diseases like Parkinson are descriptive. Genome-wide association study (GWAS) is used to classify the genes associated with Parkinson's and other diseases. The information of identified genes empowers scientists to early diagnose, treat, and stop diseases. Due to the complexities of the illness, identifying such genes is a challenging task. In this article, we apply two methods of feature selection to choose a subset of genes that are used to predict PD with high precision in classification. The chromosome corresponding to selected features is analyzed by Perturbation-based Feature Selection (PFS) and Hilbert-Schmidt independence criterion (HSIC)-Lasso. These algorithms are used to identify how chromosomes play an important role with respect to PD. We used a dataset consist of 50 predominantly patients gene expression profiles with early-stage Parkinson's disease (PD) and 55 normal GEO samples. These methods provide a series of features involved in disease-specific processes that are applied to prioritize candidate genes in GWAS loci.

Keywords: Genome-wide association (GWAS), Single nucleotide polymorphisms, Perturbation- based feature selection

1. Introduction

Parkinson's disease as a neurological disease progresses during time. It causes moving disabilities, which starts with hand involuntary quivering movement and continues to movement difficulties and falling down because of balance loss.

Genome-wide association studies (GWAS) have linked thousands of single nucleotide polymorphisms (SNPs) to the risk of Parkinson's disease (PD), a neurodegenerative and age-related disorder, with a

total lifetime risk of about 10 [1]. People diversity is a result of genetic diversity on SNPs. Differences in a DNA is occurred by each SNP. Furthermore, in the prediction of person response to specific medication, SNPs are valuable. Tracking genes of heritage disease is applicable with the aim of SNPs.

Advances in identifying DNA sequencing have a significant impact on disease detection through sequencing a single suspect tissue. Parallel sequencing technologies are used to identifying sequence the normal and tumor skin cells genomic DNA [2] .

Genetic association research identifies specific chromosome regions containing only a small number of genes and it helps to diagnose a particular disease susceptibility gene. GWAS has several advantages over alternative methods. GWAS makes a complete genome sequence in an unbiased manner and has the ability to classify various risk factors while the candidate gene studies select genes for analysis based on known or suspected disease mechanisms.

GWAS requires the screening of thousands of samples using hundreds of thousands of PNS labels found in the human genome. These algorithms are used to compare the occurrence of any single PNS alleles or genotypes between disease and control cases. This analysis identifies regions (loci) with statistically meaningful variations in allele or genotype frequencies across cases and controls.

On a chromosome, there are specific points in which genetic markers or genes are located; these specific spots are called loci.

This study has been organized as follows. Section 2 reviews related works to GWAS and machine learning techniques. In section 3, we explain our utilized feature selection and classification algorithms. Section 4 describes selected genes in more details. Finally, in section 5, we conclude the study and propose and future works.

2. Related Work

Machine learning algorithms are widely used in various studies to identify the genes associated with various diseases. In [3] , a comparison between 13 popular open-source ML algorithms is done. They analyzed their performance over a set of 165 supervised classification problems in terms of best-balanced accuracy and then studied the impact of hyper-parameter tuning and model selection. In addition, they looked at how algorithms to cluster across the problems tested and performed a set of algorithms that maximize performance across datasets. Similar to this method, we implement a clustering algorithm in our previous work [4]

Genome-wide association studies (GWAS) are used in several experiments to explain various disorders' genetic nature, including Parkinson's disease. A meta-analysis of Parkinson's disease genome-wide PNS data is performed by [4] using a common set of 7,893,274 variants in 13,708 cases and 95,282 controls. They used a semi-customized genotyping array to replicate each locus in an independent sample series, used the 26 genome-wide important candidate loci involved in Parkinson's disease from the primary meta-analysis, and then investigated the association of six loci associated with the risk of Parkinson's disease. They tested whether there were multiple independent risk alleles in each of the 26 genome-wide; in the discovery process, also 22 independent risk loci for Parkinson's disease are found, and in the replication phase, two replicated loci confirmed in the replication process, and four loci found by a second risk allele.

In [5] authors conducted a GWAS of around 7,607 PD-risk PNS with an additional 23,759 high relation disequilibrium-associated variants paired with eQTL gene expression. They examined different sets of genes associated with PD risk loci and related genes to nearby super-enhancers, which are frequently found in close proximity to major genes in the completely genome-wide screen. The speed of generating information outruns the speed of technology for designing capacious storage memory. For example, a clinical examination that involved recording the vital signals has to record the data, day and night. Moreover, this data will be stored for future assessment and checking the patient's treatment behavior in the long run. In [6], the authors presented a data compression method to improve the storage size and increase the transfer speed while the GWAS-like data are processing. Their method is applicable to the sparse signal in genome-wide association studies. There is some research for the optimal approach in the control area concerning the factorized systems and

data packet which can be used on the GWAS field [7] [8]. The authors in [8] claimed that the additional variable in the system state space is used to improve the optimal controller's performance in the presence of noise. The results of the simulations performed in the content software show the proposed method's efficiency compared to the conventional approaches [7].

In [9], a meta-analysis is performed between the PDWBS (Web-Based Parkinson's disease Study) GWAS and the findings for the top 10,000 PD meta-analysis models of more than 13,000 cases and 95,000 controls. The researchers used an inverse-variance weighted approach to combine association statistics.

In [10] a Weighted Protein-Protein Interaction Network Analysis (WPPINA) pipeline is used to define PD-specific impacted pathways and to stratify candidate genes within PD-GWAS loci. A hereditary type of PD is used to identify seed proteins and construct a protein network for Parkinson's genetics. They used 32 relevant SNPs, mapping them to the GWAS-loci and matching those encoding proteins to the PD-specific risk-processes highlighted by WPPINA to assist the gene rank within the PD-GWAS loci. The researchers have statistically confirmed their findings by generating 100,000 random sized gene-sets and measuring P-values.

A combination of multiple Microarray and RNA-seq platforms as a gene quantification technology is proposed in [11] to design a multiclass study to collect a higher number of samples and ensure the heterogeneity of their analysis. In the first step, they selected the possible Differentially Expressed Genes (DEGs) to recognize different types of Leukemia and then performed the minimum-Redundancy Maximum-Relevance (mRMR) feature selection algorithm to choose the most significant genes and evaluate the classifiers. Additionally, they performed different classification algorithms such as Support Vector Machines (SVM), Random Forest (RF), k-Nearest Neighbor (k-NN) and Naive Bayes (NB) and compared their performance with ANOVA test to decide if the classifiers have meaningful differences among them.

A study on DNA microarray datasets of existing feature selection methods is given in [12]. They study the characteristics of microarray data sets and the feature selection methods applied to the DNA micro array data analysis field, which due to its large number of features and the limited sample sizes is a difficult challenge for machine learning researchers. The selection of features has become an important step since the advent of microarray data classification to reduce the number of features (genes); the authors studied nine binary microarray datasets that suffer from several challenges such as class imbalance, overlap, or dataset shift. They divided datasets with Distribution optimally balanced stratified cross-validation and evaluated them using Support Vector Machine (SVM) and naive Bayes as classifiers, and used classification accuracy, and precision on the test partitions.

A genome-wide approach to RNAi screening, initially in Drosophila cells and confirmed in HeLa cells, is used [13] to classify 20 genes that have retained their role in promoting Parkinson translocation and mitophagy.

Evolutionary algorithms have been utilized for identifying disease-related genes as well. An overview of the analysis and monitoring of PD in humans is provided in [14]. The authors described computational approaches using evolutionary algorithms (EAs) that provide clinically relevant objective measures to recognize and to quantify PD, both in humans and animal models. They used EAs to provide robust classifiers for discrimination between disease and controls, and between disease types.

3. Feature selection

3.1. Feature Selection Methods

In machine learning and statistics, feature selection is the method of selecting a subset of relevant features, while we have needless or unrequired features, without incurring much loss of information. Feature selection methods are often used in domains where there are many features and relatively few samples. Feature selection can be most beneficial in reducing the dimension of the data to be processed by the classifier, reducing execution time and improving predictive accuracy [15].

Many different feature selection methods are widely used for micro array analysis. Aforementioned methods attempt to eliminate irrelevant and useless features so that the classification of new cases will be

more accurate. The popular micro-array data analysis methods are available in [16].

Feature selection process has been divided into four principal steps: feature subset generation, evaluation of that subset, ending criterion and validation of result. Generating the feature is done by e heuristic search method, which uses searching approaches such as sequential, complete, and random search to build features subsets.

Then in the second step, we check the produced subset whether it is superior to the prior or not and as a result, we return the greatest subset will be returned.

The two procedures are repeated up to reaching the stopping criterion. The ultimate greatest feature subset is validated by past knowledge or by applying different tests. Figure 1 shows the feature selection process [17].

Algorithms of feature selection are categorized into three types [18]:

- The filters: They select features from the data without any learning required.
- The wrappers: They employ learning techniques for selecting useful features.
- The embedded: combine the feature selection step and the classifier, structure [16].

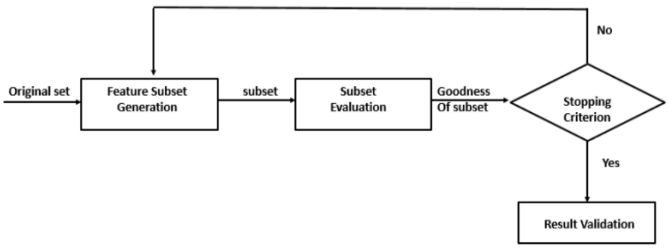

Figure 1. *Feature Selection Process*

The best feature subset is selected by applying statistical standards. Filter-based feature selection techniques are categorized to two major types such as feature ranking and subset selection [19].

Simple feature ranking methods include the use of statistical metrics, like the correlation coefficient. The most common subset selection approaches are wrapper-based approaches [20].

We explain applied feature selection methods in section 3.1.1 and 3.1.2.

3.1.1 Perturbation-based feature selection

Data perturbation is the procedure of eliminating samples from the original dataset and forming couple of shortened datasets [21].

Dataset perturbation is used by researchers to inspect their results by implementing feature selection methods on the improved datasets. Sometimes the original dataset and provide a ranked list for each of the datasets and measure the stability between the ranked lists [22]. Perturbation-based feature selection method (PFS) selects a smaller number of features while achieving the classification accuracy of other methods. We employ PFS to choose the most affected genes to discriminate between PD patients and normal patients. Due to its inherent structure of the algorithm, PFS removes noisy and irrelevant features and then selects a small subset of features that are not correlated with each other.

3.1.2 Hilbert-Schmidt Independence Criterion Lasso based feature selection

Feature selection is also known as variable selection in statistics. Least Absolute Shrinkage and Selection Operator (LASSO) and Least Angle Regression (LARS) are the most important methods of variable selection. LASSO is a subset selection based on least square regression [23] and LARS is a forward stage wise feature selector [24]. It is an efficient way of solving the same problem as LASSO. [25] Introduce a non-linear feature selection method, which is useful for high-dimensional and small sample data and the instances numbers, called Hilbert-Schmidt Independence Criterion Lasso (HSIC Lasso). HSIC Lasso is another feature selection method that we employ in our project to select non-redundant features related to PD using a set of kernel functions.

It should be pointed out that having HSIC-Lasso gives us superior accuracy of classification however; lesser feature subset is selected versus both LARS and Lasso methods.

3.1.3 Stratified K-fold cross-validation

Cross-validation is another alternative method to examine the durability of feature selection methods.

By applying cross validation procedure, the original data is divided into n folds. Training and testing data are provided by choosing n-1 folds as training and the

last part as testing data. To have an improved results, the procedures are executed n times [26].

In case of a dichotomous classification, this means that each fold contains roughly the same proportions of the two types of class labels. In other words, stratification is the process of rearranging the data to ensure each fold is a valid delegate of the whole. For example, in a binary classification problem where each class comprises 50% of the data, it is best to arrange the data such that in every fold, each class includes around half the instances. Stratification is generally a better scheme, in terms of both bias and variance, when compared to regular cross-validation [27]. In this project, we use stratified 4-fold cross-validation where proportion of each label class is preserved in each fold.

3.2 Classification Methods

3.2.1 Support Vector Machine (SVM)

Support Vector Machines introduced by [28] is considered one of our applied classification approaches, which allows a high accuracy in comparison with other classifiers such as decision trees and logistic regression. This work solves linear and non-linear problems and it is known for its kernel trick to handle nonlinear problems. Given a set of training samples as relating to one or the other of two classes, the algorithm creates a line or a hyperplane, which separates the data into classes. In SVM model the samples are represents as points in the space and mapped them into separated classes, which are divided by a clear gap that is as wide as possible. New samples are then mapped into the same space and go to a predicted class based on which side of the gap they are fall.

3.2.2. Random Forest

Applying Random Forest for SNP discovery is related to human disease and it has been grown in recent years. We examine the use of random forest, which is a supervised classification algorithm that uses a set of classification trees. It was developed by [29]. It is an improved of previous work on classification and regression trees or (CART) [30] and bootstrap aggregating [31]. CART is a useful tool for developing a classifier and it shows a binary tree. Bagging is a technique for sampling data in which sampled data is accompanied with replacement and the classifier is developed by the bootstrap sample. After several repetitions, results are aggregated over all trees to form a less variable classifier with a lower prediction error in comparison with the original classifier. In Bagging, the variance reduction is restricted by the correlation between trees; while correlation is increased or maximized, the potential for reduction is decreased.

In the RF algorithm, CART trees are bagged and to drop the association between trees, instead of searching over all p variables at each node the bagging process is done. The method divides the data continuously up to no more splits are possible or there are no more variables. Since the bagging process is part of RF algorithm, RF leaves are unpruned and bagging helps to minimize the variance of lacking pruning. In other hand, CART helps to the stability by using pruning the trees.

3.2.3 Adaboost

Another classification algorithm is AdaBoost, which is a machine learning meta-algorithm developed by [32] .In this project, it is utilized in conjunction with the SVM algorithm to improve accuracy and performance. In some machine learning algorithms, each sample consists of a huge number of features, (for instance, in this study, there are more than 20,000 features for each sample). Consequently, evaluating each feature reduces the speed of classifier training and power of prediction. The Adaboost training method selects only essential features and does not need to process irrelevant features to improve predictive power, reduce dimensionality, and potentially improve execution time.

3.3 Neural Network

3.3.1 Auto encoder

Auto encoder (AE) is a type of neural networks that aims to copy input values to the output values. They compress and reduce the input into a latent-space form, and then build the output from this form. The network is comprised of two parts so that the first part compresses and reduces the input into a latent-space form and can be an encoding function $h = f(x)$; the second part reconstructs the input from the latent-space form and can be a decoding function $r = g(h)$. In our project, the input of the auto encoder is more

than 22000 genomes, which is compressed into 100 genomes, then we apply SVM algorithm on the compressed genomes that are extracted from auto encoder algorithms.

4. Experiments

The data set we used in this study consist of 50 predominantly patients gene expression profiles with early stage Parkinson's disease (PD) and 55 normal GEO samples under accession number GSE6613 [33]. Expression levels of 22,283 genes were calculated by the means of the human genome array Affymetrix. The aim of the project is to find a subset of genes that are more useful in predicting or diagnosing Parkinson's disease and can be used to create a model for the detection of a new patient with a disease.

We are finding two main obstacles in dealing with our Genomic Dataset. First, it has a large scale that make it difficult to provide accurate evaluation of the data. There are a number of genes with negligible impact in, which are often considered irrelevant or noisy genes. The identification and elimination of irrelevant genes is an essential phase in our applied algorithms. Second obstacle is that, there is a high correlation between some of the existing genes. In the method, feature selection algorithms identify main features while eliminating redundant features. We apply two methodologies for the collection of functions, namely PFS and HSIC-Lasso [25]. When we have a specified subset of genes using PFS or HSIC-Lasso, we use Support Vector Machine (SVM), Random Forest, Adaboost and Auto Encoder as classifiers to have a model based on the training dataset and selected subset of genes, and then the model is tested and validated on the test set. PFS and HSIC-Lasso select gene features with a prediction accuracy of 86 and 94 respectively to PD and normal patients.

In this project, we have experimented a dataset containing the gene expression profiles of 50 patients predominantly with early-stage Parkinson's disease (PD) and 55 normal samples from GEO under accession number GSE6613 [33]. The expression levels of 22,283 genes were measured using an Affymetrix Human Genome Array.

The first stage of our project is the preprocessing phase that we checked whether there are Nan values alternatively, not and then normalized the genes before the training phase.

We applied PFS to the dataset GSE6613 where a subset of features is chosen by PFS to differentiate between normal and patient sample. We employed different classification algorithms such as SVM, Random Forest and AdaBoost and reported the results in the table 1. You can see the highest obtained accuracy is 86.03 and standard deviation of 4.8113 with SVM classification algorithm for 10 times run.

Then we applied HSIC-Lasso to dataset GSE6613 in two different approaches. In the first approach, HSIC-Lasso has been implemented on the whole dataset, and 88 features are selected then we employed different classification algorithms such as SVM, Random Forest and AdaBoost to train the model and evaluate it as well and summarized the results in table 1. The highest obtained accuracy is 94.9 with SVM classification algorithm for 10 times run. In the second approach, the feature selection mechanism only applies to the training dataset. The achieved accuracy in the latter approach is more accurate because feature selection is applied separately on the train dataset and test dataset. Therefore, it is possible to identify whether the significant features are approximately the same in both the test dataset and the training dataset or not.

In table 1, we have summarized the results of applying PFS and HSIC-Lasso with different classification methods and summarized the average classification accuracy and the average number of features over ten runs. We have also run PFS 100 times; the average classification accuracy is 83.56 with the standard deviation of 6.4165. Then we extract the chromosome of selected features and report on the concentrated loci of the selected features. In figure 2, we have plotted the distribution of all the chromosomes (in the whole dataset) and the distribution of chromosomes containing selected features using PFS (in the reduced dataset). We can see that PFS has chosen features from all the chromosomes, where in eight chromosomes, namely 3, 6, 7, 10, 13, 17, 19, and x the frequency is higher than the complete dataset. Next, we investigate the distribution of chromosomes corresponding to selected features by HSIC-Lasso.

Table 1. Accuracy of selected features by the means of PFS and HSIC-Lasso with different classifiers

Method	Classifier	Feature Selections	CA %
PFS	SVM	65.2	86.03%
	Random Forest	63.1	86.0%
	AdaBoost	52.2	82.0%
HSIC-Lasso	SVM	63.3	94.9%
	Random Forest	63.3	74.0%
	AdaBoost	63.3	84.0%

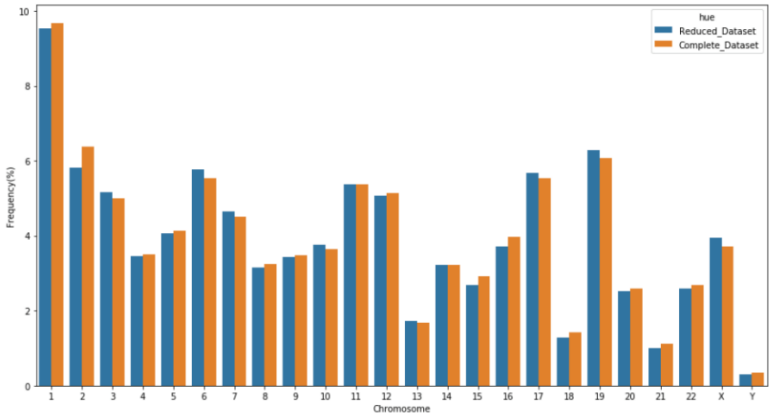

Figure 2. Frequency of chromosomes appearing in the whole dataset compared to Frequency of chromosomes containing reduced dataset using PFS

Figure 2 shows the frequency of chromosomes containing in the whole dataset compare to the frequency of chromosomes appearing in selected features using HSIC-Lasso over twenty runs. We can see the chromosomes 1, 4, 6, 7, 8, 14, 15, 20, 21, 22, and X are over-expressed. It is the reason that to examine whether the over-expressed chromosomes perform a crucial role concerning Parkinson diseases.

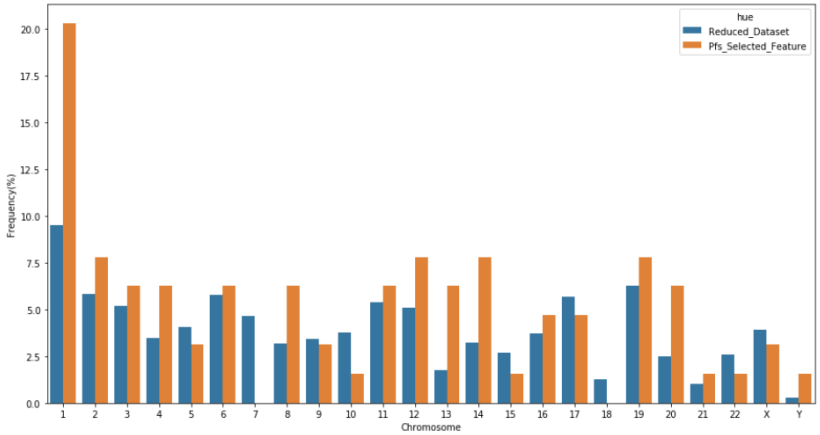

Figure 3. The frequency of chromosomes appearing in the whole dataset compared to Frequency of chromosomes containing selected features by HSIC-Lasso

We also use Jaccard similarity to measure the similarity between chromosome sets and selected features using PFS and HSIC-Lasso and report the result in the table 2. We can see we have a low similarity between selected features of PFS and HSIC-lasso.

Table 2. Jaccard Similarity between chromosomes and selected features with PFS and HSIC-Lasso

Jaccard Similarity between PFS and HSIC-Lasso	Value
Chromosomes	0.875
Selected features	0.0059

We also show the association between selected genes using PFS and HSIC-Lasso in figure 4 and 5. The heat map in figure 4 indicates there is very little association between selected genes. We can also see in figure 5 that selected genes by HSIC-Lasso have a very little correlation. We need a small correlation within selected features but the high correlation with the classification outcome.

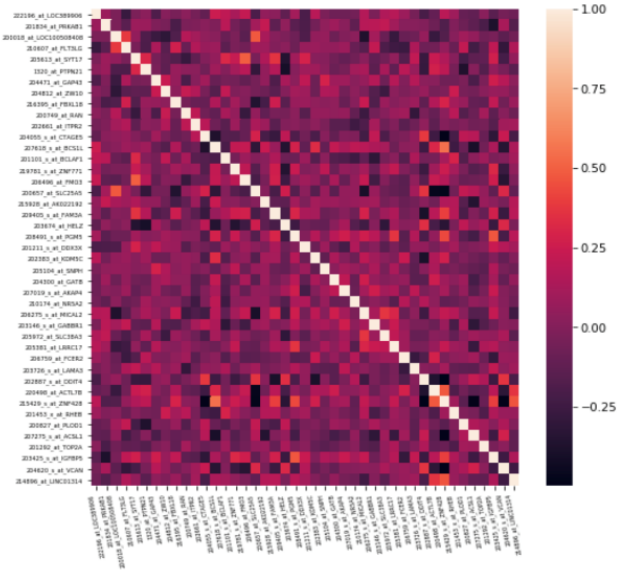

Figure 4. Heat map generated using the top-ranked genes selected by PFS

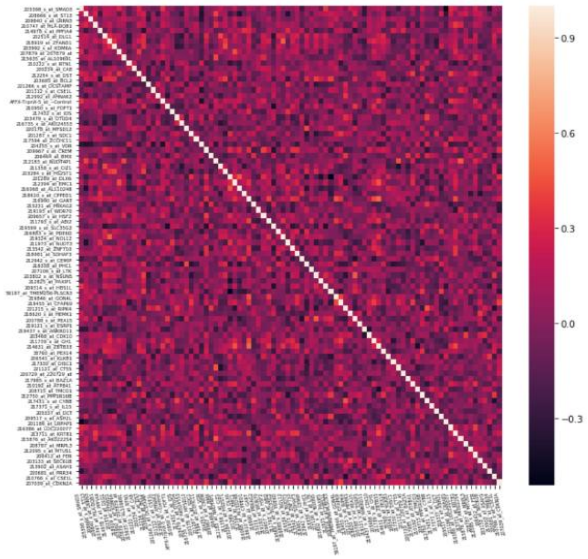

Figure 5. Heat map generated using the top-ranked genes selected by HSIC-Lasso

In table 3, we summarized the correlations between selected genes using PFS and the 54 genes reported by [4], [9], [13] and [10]; Pearson Correlation is used for calculating the correlation between genes. We have reported only correlations more than 0.7. We note that the highest correlation is SNCA and GSPT1. SNCA has a high effect on Parkinson Disease (related genes to Parkinson Disease based on related works); The result shows this gene has a high correlation with selected genes by PFS algorithm.

Table 3. The correlation between related genes and selected genes using PFS

PFS	Related Genes to Disease	CORRELATION
GSPT1	SNCA	0.90
MKRN1	SNCA	0.89
FECH	SNCA	0.86
SACMIL	VPS13C	0.76
EFR3A	VPS13C	0.76
PEX11A	FZD5	0.71
AK024527	DDRGK1	0.70

We compared the correlations between selected features by HSIC-Lasso and the reported genes as well. We have just reported the correlations of more than 0.6.

Table 4. Correlation between related genes and selected genes using HSIC-Lasso

HSIC-Lasso	Related Genes to Disease	CORRELATION
NUDT4P1	ZDHHC8P1	0.69
AL109691	SPPL2B	0.66
CREM	FZD5	0.63
SEC61B	MMP16	0.62

We have shown in table 4 that unlike PFS, there are no significant correlations in this case. It could be because the selected features contain new and not previously investigated genes related to PD.

5. Conclusion

Through this study, we point out machine learning methods, which are applied to defined disease-specific biological processes. These methods provide a series of features involved in disease-specific processes that are applied to prioritize candidate genes in GWAS loci. In this work, we implemented two different feature selection methods that select a subset of genes that can be used to discriminate PD patients from normal samples. However, an accurate splitting of dataset is adopted for training and testing data, we observe that all applied feature selection methods work well. We have also analyzed the chromosome corresponding to selected features by PFS and HSIC-Lasso to identify how the chromosomes play an important role with respect to PD.

Conflict of interest

The authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest, or non-financial interest in the subject matter or materials discussed in this manuscript.

Acknowledgments

No applicable.

References

[1] Driver, J. A., Logroscino, G., Gaziano, J. M., & Kurth, T., "Incidence and remaining lifetime risk of Parkinson disease in advanced age," *Neurology,* vol. 72, no. 5, pp. 432-438, 2009.

[2] Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., McLellan, M. D., Chen, K., ... & Cook, L. , "DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome," *Nature,* vol. 456, no. 7218, pp. 66-72, 2008.

[3] Olson, R. S., La Cava, W., Mustahsan, Z., Varik, A., & Moore, J. H., "Data-driven advice for applying machine learning to bioinformatics problems," *arXiv preprint arXiv,* 2017.

[4] Rafieipour, H., Zadeh, A. A., & Mirzaei, M., "Distributed Frequent Itemset Mining with Bitwise Method and Using the Gossip-Based Protocol," *Journal of Soft Computing and Decision Support Systems,* vol. 7, no. 3, pp. 32-39, 2020.

[5] Nalls, M. A., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., Saad, M., ... & Schulte, C., "Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease," *Nature genetics,* vol. 46, no. 9, pp. 989-993, 2014.

[6] Pierce, S., & Coetzee, G. A., "Parkinson's disease-associated genetic variation is linked to quantitative expression of inflammatory genes," *PLoS One,* vol. 12, no. 4, 2017.

[7] V. Izadi, H. Ahani and P. K. Shahri, A compressed-sensing-based compressor for ECG, Biomedical engineering letters, 2020.

[8] S. R. Surakanti, S. A. Khoshnevis, H. Ahani and V. Izadi, "Efficient Recovery of Structrual Health Monitoring Signal based on Kronecker Compressive Sensing," *International Journal of Applied Engineering Research,* vol. 14, pp. 4256--4261, 2019.

[9] H. Ahani, M. Familian and R. Ashtari, "Optimum Design of a Dynamic Positioning Controller for an Offshore Vessel," *Journal of Soft Computing and Decision Support Systems,* vol. 7, pp. 13--18, 2020.

[10] Chang, D., Nalls, M. A., Hallgrímsdóttir, I. B., Hunkapiller, J., Van Der Brug, M., Cai, F., ... & Hinds, "A meta-analysis of genome-

wide association studies identifies 17 new Parkinson's disease risk loci," *Nature genetics*, vol. 49, no. 10, p. 1511, 2017.

[11] Ferrari, R., Kia, D. A., Tomkins, J. E., Hardy, J., Wood, N. W., Lovering, R. C., ... & Manzoni, C. , "Stratification of candidate genes for Parkinson's disease using weighted protein-protein interaction network analysis," *BMC genomics*, vol. 19, no. 1, pp. 1-8, 2018.

[12] Castillo, D., Galvez, J. M., Herrera, L. J., Rojas, F., Valenzuela, O., Caba, O., ... & Rojas, I. , "Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level," *PloS one*, vol. 14, no. 2, p. e0212127, 2019.

[13] Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A., Benítez, J. M., & Herrera, F., "A review of microarray datasets and applied feature selection methods," *Information Sciences*, vol. 282, pp. 111-135, 2014.

[14] Ivatt, R. M., Sanchez-Martinez, A., Godena, V. K., Brown, S., Ziviani, E., & Whitworth, A. J., "Genome-wide RNAi screen identifies the Parkinson disease GWAS risk locus SREBF1 as a regulator of mitophagy," in *Proceedings of the National Academy of Sciences*, 2014.

[15] Smith, S. L., Lones, M. A., Bedder, M., Alty, J. E., Cosgrove, J., Maguire, R. J., ... & Elliott, C. J., "Computational approaches for understanding the diagnosis and treatment of Parkinson's disease," in *IET systems biology*, 2015.

[16] Devi, S. N., & Rajagopalan, S. P., "A study on feature selection techniques in bio-informatics," *International Journal of Advanced Computer Science and Applications*, 2011.

[17] Hira, Z. M., & Gillies, D. F., " A review of feature selection and feature extraction methods applied on microarray data," *Advances in bioinformatics*, 2015.

[18] Sutha, K., & Tamilselvi, J. J., "A review of feature selection algorithms for data mining techniques," *International Journal on Computer Science and Engineering*, vol. 7, no. 6, p. 63, 2015.

[19] Mwangi, B., Tian, T. S., & Soares, J. C., "A review of feature reduction techniques in neuroimaging," *Neuroinformatics*, vol. 12, no. 2, pp. 229-244, 2014.

[20] Hall, M. A., & Holmes, G. , "Benchmarking attribute selection techniques for discrete class data mining," *IEEE Transactions on Knowledge and Data engineering*, vol. 15, no. 6, pp. 1437-1447, 2003.

[21] Kohavi, R., & John, G. H., "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273-324, 1997.

[22] Varshavsky, R., Gottlieb, A., Horn, D., & Linial, M., "Unsupervised feature selection under perturbations: meeting the challenges of biological data," *Bioinformatics*, vol. 23, no. 24, pp. 3343-3349, 2007.

[23] Boulesteix, A. L., & Slawski, M. , "Stability and aggregation of ranked gene lists," *Briefings in bioinformatics*, vol. 10, no. 5, pp. 556-568, 2009.

[24] Tibshirani, R. , "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267-288, 1996.

[25] Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R., "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407-499, 2004.

[26] Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., & Sugiyama, M. , "High-dimensional feature selection by feature-wise kernelized

lasso," *Neural computation*, vol. 26, no. 1, pp. 185-207, 2014.

[27] Jung, Y., & Hu, J., "AK-fold averaging cross-validation procedure," *Journal of nonparametric statistics*, vol. 72, no. 2, pp. 167-179, 2015.

[28] Kohavi, R., "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International Joint Conference on AI(IJCAI-95)*, 1995.

[29] Cortes, C., & Vapnik, V. , "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273-297, 1995.

[30] Breiman, L. , "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.

[31] Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A., Classification and regression trees, CRC press, 1984.

[32] Breiman, L., "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123-140, 1996.

[33] Freund, Y., & Schapire, R. E., "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119-139, 1997.

[34] Scherzer, C. R., Grass, J. A., Liao, Z., Pepivani, I., Zheng, B., Eklund, A. C., ... & Bresnick, E. H., "GATA transcription factors directly regulate the Parkinson's disease-linked gene α-synuclein," in *Proceedings of the National Academy of Sciences*, 2008.