

Importance of Attribute Selection for Parkinson Disease

*¹Kemal Akyol, ² Şafak Bayır, ³ Baha Şen

¹ Computer Engineering Department, Kastamonu University, Turkey, kakyol@kastamonu.edu.tr 

² Educational Sciences Department, Karabük University, Turkey, safakbayir@karabuk.edu.tr 

³ Computer Engineering Department, Yıldırım Beyazıt University, Turkey, bsen@ybu.edu.tr 

Research Paper

Arrival Date: 18.03.2019

Accepted Date: 19.09.2019

Abstract

Parkinson disease is a neurological disorder occurring at older ages. It is one of the most painful, dangerous and untreated diseases. In this study, a new application based on assessing the importance of attributes using the ranking techniques was carried out for diagnosis of this disease. The effects of the attributes on the Parkinson disease are determined by utilizing Stability Selection method. The selected attributes dataset and all attributes dataset have been sent as input data to the Random Forest and Logistic Regression algorithms in order to investigate the best model which is to be effective in the diagnosis of this disease. This study including the model which presented the best performance might be a powerful tool for effective diagnosis of this disease.

Keywords: Parkinson disease, importance of attribute, stability selection, k-fold cross validation, machine learning.

1. INTRODUCTION

Parkinson Disease (PD) is a neurological disorder disease which effects a patient's life quality and progresses slowly. The prediction of this disease is quite difficult in an early age because of the symptoms of the disease arise in middle and later ages [1]. It occurs between 50 and 95 years old [2] and this disease is the second most familiar neurodegenerative disorder causing speech and speech disorders [3-5]. Furthermore, the result of the disruption of the dopamine-releasing cells and the deterioration of central nervous system, this disease affects the body movements in brain [4-6]. The main focus of this study is to detect important attributes for PD based on Stability Selection (SS) method. In this context, the machine learning-based models were designed, and deployed on the dataset, which consists of these attributes. In this context, the sub-datasets, which include the important attributes, were obtained from the raw datasets within the frame of the cross validation rules. Next, the successes of learning algorithms were evaluated on these sub-datasets. The rest of this paper was structured as follows; Section 2 presents the related studies. Section 3 introduces the information about the materials and methods. Section 4 presents experimental results in detail. Finally, Section 5 drawn the conclusion.

2. LITERATURE REVIEW

There are numerous computer-based models and

investigations on the disease. Speech disorder, which is very common in Parkinson patients, occurs in about 90% of patients [7-8].

Some of the works performed on this disorder and the PD are as follows: Das carried out a comprehensive study of Neural Networks, DMneural, Regression and Decision Trees in order to detect of this disease. The Neural Networks presented best results [5]. Umaphy et al. proposed a joint time-frequency approach in order to detect the pathological voices. Several features were extracted from these decomposed signals. Next, these features were analysed using statistical pattern identification techniques in their study [9]. Lorente et al. presented the system based on learning which allows the detection of voice disorders system [10]. Shrivastav carried out a study to identify the relationship between the acoustic spectrum and perceptual ratings of breathiness [11]. Sakar et al. built tele diagnosis and tele monitoring models using a wide variety of voice samples. They examined the discriminative information for PD using machine learning algorithms [12]. Harel et al. investigated a particular pattern of speech changes which show up in PD with idiopathic patients was investigated [13]. Shahbaba and Neal introduced a non-linear model based on Dirichlet mixtures in order to detect the folding class of protein sequences and PD [14]. Guo et al. proposed a hybrid model based on expectation maximization and a genetic algorithm for detecting of this disease and obtained 93.1% classification accuracy [15].

Daliri proposed a Chi-square distance kernel based Support Vector Machine (SVM) for the diagnosis of the PD. For this purpose, the measurements of gait signals, Chi-square distance were used and obtained 91.2% classification accuracy [16]. Li et al. introduced a fuzzy-based non-linear transformation approach in order to extract the optimal subset utilizing Principal Component Analysis. They obtained 93.47% classification accuracy by utilizing the SVM on these optimal features [17]. Sakar and Kursun proposed a method consists of hybrid mutual information on feature selection. A minimum subset of features with maximal joint relevance was selected. Next, they built a predictive model using SVM and achieved 92.75% classification accuracy [18].

Ozcift and Gulden classified the Parkinson, diabetes and heart diseases data by utilizing correlation-based feature selection algorithm and 30 machine-learning algorithms. They achieved 92.75% classification accuracy with Random Forest (RF) classifier [19]. Luukka proposed a new method based on feature selection utilizing fuzzy entropy measures to combine with similarity classifiers, and achieved 85.03% average accuracy [20]. Babu et al. introduced a novel approach that examines magnetic resonance images in order to detect critical brain regions responsible for PD. They achieved 82.3% classification accuracy [21]. Martinez-Murcia et al. presented a new computer aided diagnosis system including pre-processing of images, voxel selection, feature extraction and classification of the images provided by Parkinson Progression Markers Initiative. Next, selected N voxels were trained using a SVM with Radial Basis Function (RBF) kernel [22].

3. MATERIALS AND METHODS

3.1. Data Collection

The publicly available Parkinson dataset [23] consists of a 23-dimensional input vector; the target class of this vector is in the last column, and it has 22 attributes. Also, the dataset includes 195 instances.

These instances obtained from 31 people, 23 of whom are with PD and the attributes composed of a range of biomedical voice measurements.

3.2. Feature Selection and Machine Learning

Machine learning algorithms obtain meaningful information from actual data and try to determine which class of data that has never been shown before. All machine learning algorithms aim to find the role of the input variables for the outcome variable [24]. Selected learning algorithms are

Random Forest (RF) and Logistic Regression (LR), respectively in this study. In briefly, RF introduced by Breiman is an ensemble learning algorithm generated by random decision trees. It provides a successful model since it investigates for the best feature among the random subsets of features [25]. LR which is a regression-based algorithm is a special case of a generalized linear model used for binary classification. The output is considered as a range of [0,1] using a sigmoid function by this algorithm [26].

Besides, more effective learning can be realized by using feature selection algorithms. In this study, Stability Selection (SS) method was utilized for improving the performance of tree-based and regression-based machine learning algorithms. A small subset of features in a dataset is selected considering the combination of ‘The Least Absolute Shrinkage and Selection Operator (Lasso)’ in order to explain the output variable [27]. Randomized Lasso method [28] which extended of Lasso can consistently select variables even if the required constraints for consistency of the original Lasso method are violated [27].

In this study, Stability Selection (SS) method was utilized for improving the performance of tree-based and regression-based machine learning algorithms. A small subset of features in a dataset is selected considering the combination of ‘The Least Absolute Shrinkage and Selection Operator (Lasso)’ in order to explain the output variable [27]. The LASSO given in Equation 1 is a regularization technique which provides synchronous of prediction of the target class and attribute selection [28].

$$\min_w \frac{\lambda}{2} \|w\|_1 + \sum_{i=1}^N (y_i - w^T x^i)^2$$

where λ indicates the trade-off between fit and sparsity, or the ratio of removed features. The penalty term means that a solution w becomes sparser as λ increases. X , y , i and w^T indicate the all attributes, target variable, attribute number and transpose of w matrix, respectively. Therefore, a model is designed by utilizing a smaller set of features [27]. Randomized Lasso method [29] which extended of Lasso can consistently select variables even if the required constraints for consistency of the original Lasso method are violated [27].

4. EXPERIMENTAL RESULTS

Flowchart of the proposed study was introduced in Figure 1. This approach, which includes the hybrid combination of different methods, was carried out by using ‘scikit-learn’ as backend machine learning library in Python 2.7 programming language on Anaconda platform.

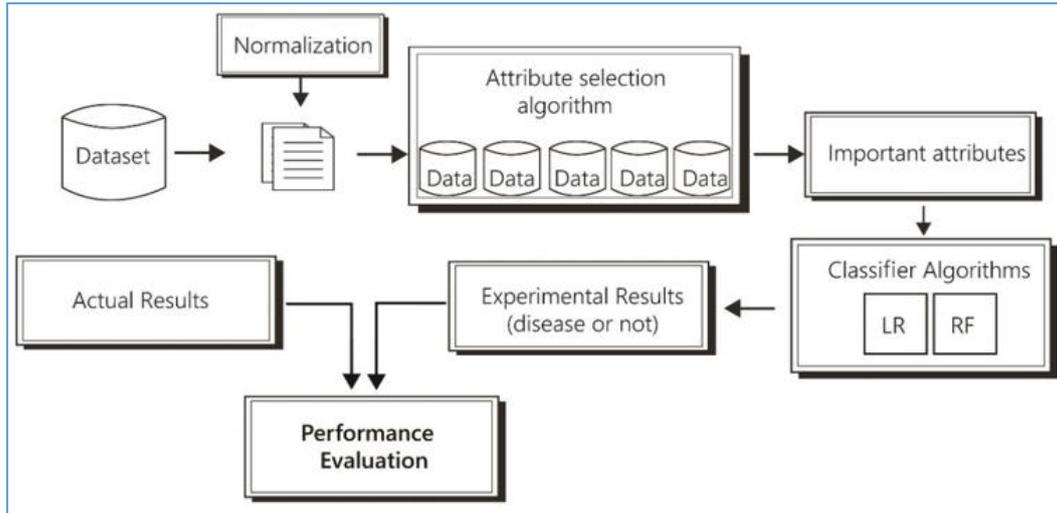


Figure 1. A flowchart of the proposed approach.

Firstly, the min-max normalization approach [30] is applied to dataset. Therefore, the data was normalized into the range from 0 to 1 values. Then, the SS attribute selection method was used in order to detection important attributes for target variable. SS method gives best attributes considering a threshold value which is taken 0.25 as default. Attributes with the importance value greater than this threshold value were accepted as important. Therefore, the sub-datasets which contain best attributes were obtained respectively. Also, the significance levels of the important attributes obtained by this algorithm were presented in Figure 2.

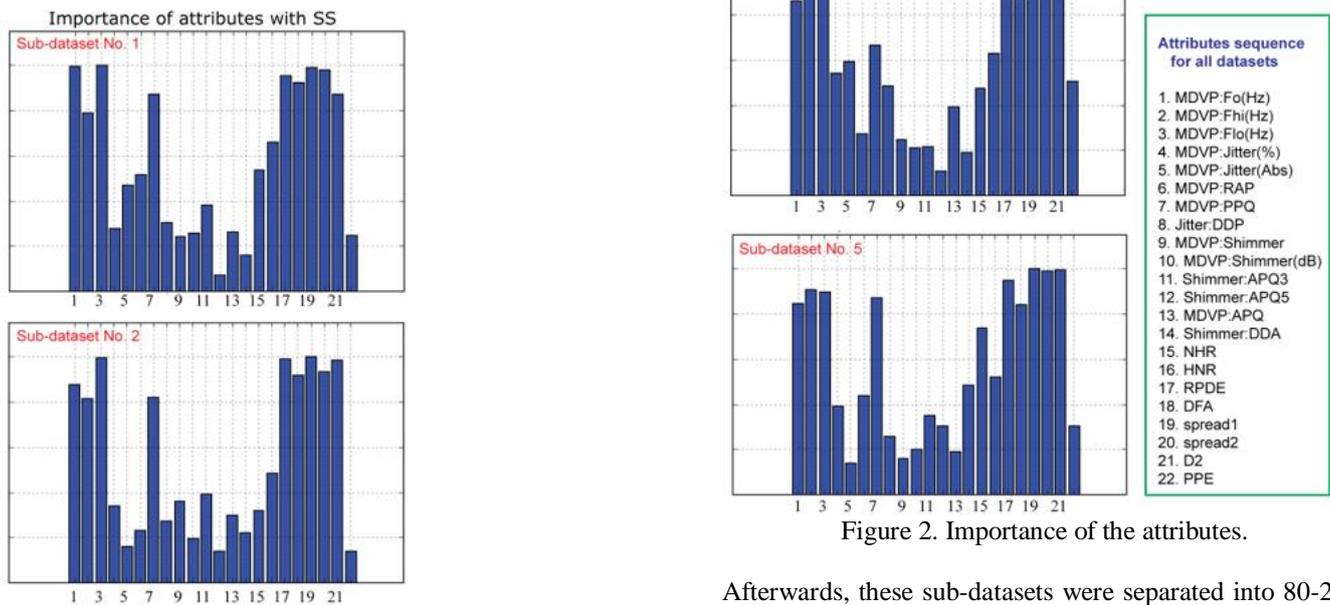


Figure 2. Importance of the attributes.

Afterwards, these sub-datasets were separated into 80-20% training and test data respectively within the framework of 5-fold cross validation to obtain a high level of efficiency for the proposed method. Namely, the sub-datasets were selected randomly to avoid favoritism. Hence, there are 156 training and 39 test records. The sub-datasets sent to as input data machine learning models to predict a person with the disease or not. Thus, successes of the models were evaluated

on each sub-dataset comparatively. The experimental metrics, Accuracy (Acc), Sensitivity (Sen), Specificity (Spe) [31] and F-measure [32], and the performance evaluations were presented in confusion matrix structure, Table 1.

Table 1. Confusion matrix [34].

		Actual values	
		No	Yes
Predicted values	No	TN	FN
	Yes	FP	TP

where TP is the number of patients correctly classified as having PD, TN is the number of patients correctly classified as not having PD, FP is the number of patients incorrectly classified as having PD and FN is the number of patients

incorrectly classified as not having PD. These metrics were given below.

$$Acc = (TP + TN)/(TP + FP + TN + FN) \quad (1)$$

$$Sen = TP/(TP+FN) \quad (2)$$

$$Spe = TN/(TN+FP) \quad (3)$$

$$F\text{-measure} = 2TP/(FP+FN+2TP) \quad (4)$$

Prediction results were given in Table 2. Besides, the average results, which were obtained by each model, were calculated. According to the proposed approach and results obtained, for the best important attributes determined by applying SS method, 94.36%, 97.49%, 84.7%, 0.96 values of Acc, Sen, Spe and F-measure, respectively were achieved with the RF classifier algorithm. Therefore, it can be clearly seen that RF algorithm outperforms the LR considering all sub-datasets.

Table 2. The results of k=5 fold cross-validation method for all models.

	LR			RF		
		No	Yes		No	Yes
Sub-dataset no. 1	No	7	6	No	13	0
	Yes	2	24	Yes	0	26
	Acc: 79.49%, Sen: 92.31%, Spe: 53.85%, F-measure: 0.86			Acc: 100.0%, Sen: 100.0%, Spe: 100%, F-measure: 1.0		
Sub-dataset no. 2	No	2	8	No	6	4
	Yes	1	28	Yes	1	28
	Acc: 76.92%, Sen: 96.55%, Spe: 20%, F-measure: 0.86			Acc: 87.18%, Sen: 96.55%, Spe: 60%, F-measure: 0.92		
Sub-dataset no. 3	No	3	4	No	6	1
	Yes	1	31	Yes	0	32
	Acc: 87.18%, Sen: 96.88%, Spe: 42.86%, F-measure: 0.93			Acc: 97.44%, Sen: 100.0%, Spe: 85.71%, F-measure: 0.98		
Sub-dataset no. 4	No	4	5	No	7	2
	Yes	2	28	Yes	0	30
	Acc: 82.05%, Sen: 93.33%, Spe: 44.44%, F-measure: 0.89			Acc: 94.87%, Sen: 100.0%, Spe: 77.78%, F-measure: 0.97		
Sub-dataset no. 5	No	6	0	No	6	0
	Yes	4	29	Yes	3	30
	Acc: 89.74%, Sen: 87.88%, Spe: 100%, F-measure: 0.94			Acc: 92.31%, Sen: 90.91%, Spe: 100%, F-measure: 0.95		
Average Acc, Sen, Spe, F-measure	83.08%, 93.39%, 52.23%, 0.9			94.36%, 97.49%, 84.7%, 0.96		

The performance of this study is compared with existing studies as shown in Table 3. Results of previous studies summarized in this table show that previous prediction methods provided good results with accuracy levels ranging between 92.75% and 93.47%. The proposed approach

produced a similar prediction performance with 94.36% overall accuracy. In addition, the Receiver Operating Characteristic (ROC) results for all models were given in Figure 3.

Table 3. The comparison of the studies for diagnosis of PD.

Study	Method	Acc (%)
[5]	Neural Networks	92.9%
[15]	Expectation Maximization and a Genetic Algorithm	93.1%
[17]	Fuzzy-based non-linear transformation	93.47%
[18]	Hybrid mutual information-based on feature selection	92.75%
[19]	Correlation-based feature selection	92.75%
Proposed approach	The combination of SS and RF	94.36 %

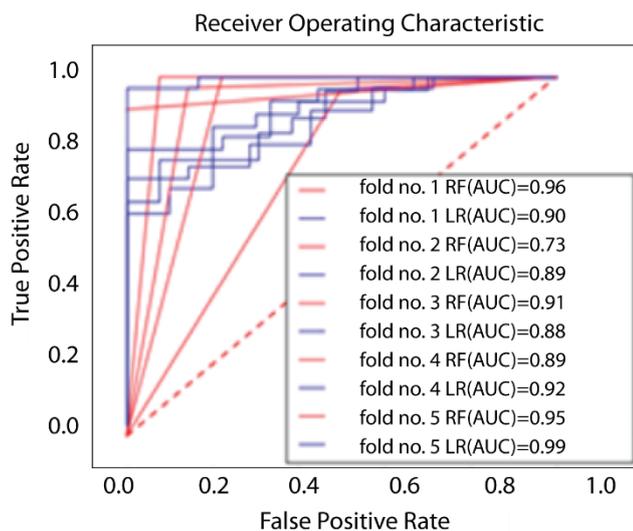


Figure 3. ROC results.

5. CONCLUSIONS

PD is a common type of neurological disorder that causes disturbances in speech and accurate voice. It has social and economic negative effects on patients' daily life significantly. The main aim of this study is to diagnose the PD according to the analysis of meaningful data based on prediction models as the most accurate way. Attribute selection is a very useful method in the processes of obtaining meaningful information from data, and machine learning. An ideal attribute set was obtained from raw dataset by eliminating less relevant attributes with the attribute selection methods. And then, the models are designed and developed during the classification stage. The learning process of this application consists of the data pre-processing, and extraction and evaluation of attributes phases. The main novelty of the proposed study uses of a hybrid methodology herein referred to as SS attribute selection method, 5-fold cross validation technique and classification algorithms. Experimental results showed that the proposed approach is very promising and comparable to other studies in the relevant literature. This approach can be used as computer-aided diagnosis system, and it is thought that this system will be able to shed a light to the future studies. In addition, it is aimed to work with a more

comprehensive attribute set and different attribute selection methods in the future. Also, additional attributes such as socio-demographic and medical diagnostic characteristics may have significant impacts on accurate diagnosis of PD. The more successful models can be achieved by applying this method on more comprehensive datasets with multidisciplinary approach.

ACKNOWLEDGEMENTS

The authors are thankful to publicly available Parkinson dataset which was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado.

REFERENCES

- [1] G. Yadav, Y. Kumar, G. Sahoo, "Predication of Parkinson's disease using data mining methods: a comparative analysis of tree, statistical, and support vector machine classifiers", *Indian J Med Sci* vol. 65, no. 6, pp. 231-242, 2011.
- [2] K. Al-Tawil, A. Akrami, H. Youssef, "A new authentication protocol for GSM networks", In: *Proceedings of the 23rd Annual Conference on Local Computer Networks*, 11-14 Oct 1998, Lowell, MA, USA, 1998.
- [3] R. Subrata and A. Zomaya, "Artificial Life Techniques for Reporting Cell Planning in Mobile Computing", In: *Proceedings of the International Parallel and Distributed Processing Symposium* vol. 14, pp. 169-187, 2003.
- [4] M.A. Little, P.E. McSharry, E.J. Hunter, J. Spielman, L.O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease", *IEEE T Bio-Med Eng* vol. 56, no. 4, pp. 1015-1022, 2009.
- [5] R. Das, "A comparison of multiple classification methods for diagnosis of Parkinson disease", *Expert Syst Appl* vol. 37, no. 2, pp. 1568-1572, 2010.
- [6] R.A. Barker and S.B. Dunnett, "Functional integration of neural grafts in Parkinson's disease", *Nat Neurosci* vol. 2, no.12, pp. 1047-1048, 1999.
- [7] L.O. Ramig, C. Fox, S. Sapir, "Parkinson's disease: Speech and voice disorders and their treatment with the Lee Silverman Voice Treatment", *Seminars in Speech and Language* vol. 25, no. 2, pp. 169-180, 2004.

- [8] A.K. Ho, R. Ianse, C. Marigliani, J.L. Bradshaw, S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease", *Behav Neurol* vol. 11, no. 3, pp. 131-137, 1998.
- [9] K. Umapathy, S. Krishnan, V. Parsa, D.G. Jamieson, "Discrimination of pathological Voices Using a Time-Frequency Approach", *IEEE T Bio-Med Eng* vol. 52, no. 3, pp. 421-430, 2005.
- [10] J.I. Godino Lorente and P. Gomez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors", *IEEE T Bio-Med Eng* vol. 51, no. 2, pp. 380-384, 2004.
- [11] R. Shrivastav, "The use of an auditory model in predicting perceptual ratings of breathy voice quality", *J Voice* vol. 17, no. 4, pp. 502-512, 2003.
- [12] B.E. Sakar, M.E. Isenkul, C.O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, O. Kursun, "Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings", *Journal of Biomedical and Health Informatics* vol. 17, no. 4, pp. 828-834, 2013.
- [13] B. Harel, M. Cannizzaro, P.J. Snyder, "Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study", *Brain Cognition* vol. 56, no. 1, pp. 24-29, 2004.
- [14] B. Shahbaba and R. Neal, Nonlinear models using Dirichlet process mixtures, *Journal of Machine Learning Research* vol. 10, pp. 1829-1850, 2009.
- [15] P.F. Guo, P. Bhattacharya, N. Kharma, "Advances in detecting Parkinson's disease", *Medical Biometrics* vol. 6165, pp. 306-314, 2010.
- [16] M.R. Daliri, "Chi-square distance kernel of the gaits for the diagnosis of Parkinson's disease", *Biomedical Signal Processing and Control* vol. 8, no. 1, pp. 66-70, 2013.
- [17] D.C. Li, C.W. Liu, S.C. Hu, "A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets", *Artif Intell Med* vol. 52, no. 1, pp. 45-52, 2011.
- [18] C.O. Sakar and O. Kursun, "Teliagnosis of Parkinson's disease using measurements of dysphonia", *Journal of Medical Systems* vol. 34, no. 4, pp. 591-599, 2010.
- [19] A. Ozcift and A. Gulten, "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms", *Comput Meth Prog Bio* vol. 104, no. 3, pp. 443-451, 2011.
- [20] P. Luukka, "Feature selection using fuzzy entropy measures with similarity classifier", *Expert Syst Appl* vol. 38, no. 4, pp. 4600-4607, 2011.
- [21] G.S. Babu, S. Suresh, B.S. Mahanand, "A novel PBL-McRBFN-RFE approach for identification of critical brain regions responsible for Parkinson's disease", *Expert Syst Appl* vol. 41, no. 2, pp. 478-488, 2014.
- [22] F.J. Martinez-Murcia, J.M. Gorriz, J. Ramirez, I.A. Illan, A. Ortiz and The Parkinson's Progression Markers Initiative, "Automated Detection of Parkinsonism Using Significance Measures and Component Analysis in DatSCAN imaging", *Neurocomputing* vol. 126, pp. 58-70, 2014.
- [23] <https://archive.ics.uci.edu/ml/datasets/Parkinsons/>, Access time: 10.01.2019
- [24] P. Ivens, A. Paulo, C. Donald, "The use of machine learning algorithms in recommender systems: A systematic review", *Expert Syst Appl*, vol. 97, pp. 205-227, 2018.
- [25] L. Breiman, "Random forests", *Mach Learn*, vol. 45, no 1, pp.5-32, 2011.
- [26] Kleinbaum DG, Klein M, *Logistic Regression A Self-Learning Text*, 3rd Edition, Springer 2010.
- [27] F. Mordelet, J. Horton, A.J. Hartemink, B.E. Engelhardt, R. Gordán, "Stability selection for regression-based models of transcription factor-DNA binding specificity", *Bioinformatics*, 29:i117-i125, 2013.
- [28] R. Tibshirani, Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol*, vol. 58, pp. 267-288, 1996.
- [29] N. Meinshausen, P. Bühlmann, "Stability selection", *J. R. Statist Soc. B*, vol. 72, no. 4, pp.417-473, 2010.
- [30] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Waltham, MA, USA, 2012.
- [31] S.A. Shaikh, "Measures derived from a 2x2 table for an accuracy of a diagnostic test", *J Biom Biostat* vol. 2, no. 128, pp. 1-4, 2011.
- [32] C.J. van Rijsbergen, *Information retrieval*. 2nd ed. London: Butterworths, 1979.