Feature Selection Techniques to Choose the Best Features for Parkinsons Disease Predictions Based on Decision Tree

Yulianti[1*], A N Syapariyah[1], A Saifudin[1], T. Desyani[1]
[1]Informatics Engineering, Universitas Pamulang Jalan Raya Puspitek 46, Tangerang, Banten 15310, Indonesia

*yulianti@unpam.ac.id

Abstract. Parkinson is a disease that is caused by nerve cell damage in the brain and incurable. Knowing about Parkinson disease is very important so that medical action can be taken to prevent Parkinson's getting worse. The dataset that uses to analysis for Parkinson disease using machine learning algorithms has many features. The dataset with many features can increase complexity, but not all features have a positive influence on the results of the analysis. Irrelevant features can reduce model performance. This research proposes to apply feature selection to choose features that have a positive effect so that the performance of the model does not decrease. The experiment results show that the application of feature selection can lead to better model performance.

1. Introduction

The increasing population of the elderly in the world in recent years has caused Parkinson's disease to become one of the diseases caused by the disease which is a neurological disease after Alzheimer's disease in elderly people [1]. Indication of Parkinson's disease includes stiff and slow muscles, tremors and inappropriate movements that the patient wants. That because of the brain's basal ganglia degeneration and lack of the neurotransmitter dopamine [2]. Reducing the growth of this disease is needed because there's no cure for Parkinson's. To decrease the growth of Parkinson's disease needs early diagnosis. But the diagnosis of this disease cannot be done only by the laboratory test [3]. At present invasive techniques and empirical experiments are the general methods applied to diagnose Parkinson's. It is necessary to review diagnostic techniques other than invasive and empirical tests. So there are cheaper, more manageable and more accurate methods [4].

Traditional diagnosis not proper for the early detection of Parkinson's because this requires many observations about daily activities, abilities, and other neurological parameters to evaluate the progress of Parkinson's. Based on research that has been done before, it has been found that Artificial Intelligence (AI) and Machine Learning have the potential for good classification. Classification systems can improve the accuracy and reliability of diagnosis and also decrease errors and a more efficient system [5]. The use of the Decision Tree has been practiced successfully for medical predictions and reliable decision-making techniques [6]. But need other features so that the results of the production are better.

How to detect the best Machine Learning algorithm techniques that can handle the early diagnosis of Parkinson's is a problem that exists today. Feature Selection Technique is the best feature in Machine Learning that can handle the early diagnosis of this disease. Another name for Feature Selection is the attribute selection technique. The process of automatically selecting attributes from a dataset helps a lot in the problem of predictive modelling [2].

Based on the prediction problem of Parkinson's disease above, the Machine Learning algorithm will be implemented with a feature selection technique using a dataset downloaded from the UCI Machine Learning Repository. By using a classification algorithm, the Decision Tree. With the hope of obtaining the best features to solve the prediction problem of Parkinson's disease.

2. Methodology

This research about Parkinson's disease prediction uses sound recordings that have been recorded using related tools. The voice data is a general dataset contained in the University of California, Irvine repository or UCI Machine Learning and can be accessed or downloaded easily so that research can be compared with one another.

The following point form the dataset:

1. Status
   Status from patient states declare healthy or has been identified as Parkinson's disease.
2. Gender
   Gender of patients consisting male or female.
3. Pitch Local Perturbation
   Pitch Local perturbation is a measurement of the pitch frequency with frequency equal to f_0 this measurement includes measurements jitter of relative (Jitter_Rel), jitter absolute (Jitter_Abs), (Jitter_RAP) jitter of relative average perturbation, Jitter of pitch perturbation quotient (Jitter_PPC).
4. Amplitude Perturbation
   Amplitude Perturbation is a measurement of the amplitude this measurement includes measurements shimmer of local (Shim_Loc), shimmer in decibel (Shim_dB), 3-point, 5-point and 11-point amplitude perturbation quotient (Shim_APQ5), (Shim_APQ11).
5. Harmonic Noise Ratio
   HNR is a measurement of the noise ratio in this measurement include the frequency 0-500 (HNR05), 0-1500 (HNR15), 0-2500 (HNR25), 0-3500(HNR35), 0-3800 (HNR38).
6. Recurrence Period Density Entropy
   This is a method for dynamical system (RPDE)
7. Detrended Fluctuation Analysis
   This is a method includes determining the statistical self-affinity of a signal (DFA).
8. Pitch Period Entropy
   Pitch Period Entropy is a measurement of a dysphonia (PPE).
9. Glottal-to-Noise Excitation Ratio
   This is a parameter used to designate whether a given sound signal comes from vocal cord vibration or turbulent sound.
10. Mel Frequency Cepstral Coefficient
    This is a method used to measure the signal from sound this measurement including Mel Frequency Cepstral Coefficient based on spectral measures of order (MFCC 0-13) and including Mel Frequency Cepstral Coefficient based on spectral measures of order (Delta0-13).

The dataset consists of 48 attributes of the results from the study. But simplified to 46 attributes including labels. Simplification of these attributes is due to 2 attributes that have nothing to do with this study.

Detecting early Parkinson's disease precisely and accurately is the purpose of this study. So that it can prevent this disease from becoming more severe due to the absence of a cure for this disease. This research is expected to increase the value of accuracy. Figure 1 is the proposed framework of this research.

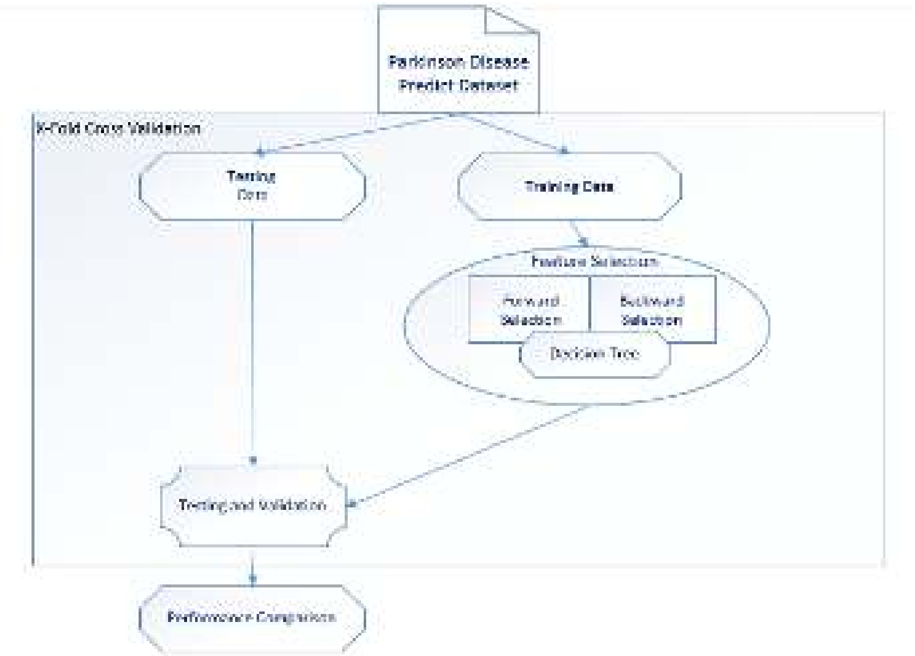

Figure 1 Parkinson disease predict label

The dataset can be classified using various classification methods, one of which is the Decision Tree. Decision Tree is a tree that has a node where each node represents the test attribute and leaf node shows classification. Examples of classification testing start from the root and then test attribute values and sort them to the branch that ends up reaching the leaf node that provides classification [7]. Here is a Decision Tree formula:

$$\boldsymbol{Entropy\ (S) = \sum_{i=1}^{n} - pi \times log_2\, pi} \tag{1}$$

Where S is case set and A is features. N is number if partition S and pi is proportion id Si to S. To measure the performance of the model used the results of validation. The confusion matrix is used to measure the performance of the model. A confusion matrix is a useful tool for analyzing how well classifiers can recognize tuples/features of different classes [8]. Confusion matrix also provides performance appraisal of classification models based on the number of objects predicted correctly and incorrectly [9]. Table 2 is a table of confusion matrix.

Table 1 Table of confusion matrix

Class		Actual	
		True	False
Prediction	True	TP (True Positive)	FP (False Positive)
	False	FN (False Negative)	TN (True Negative)

The performance of the model can be seen from the value of Accuracy or AUC. To calculate the performance of the model the following equation can be used [8].

$$\boldsymbol{Accuracy = \frac{TP+TN}{TP+TN+FP+FN}} \tag{2}$$

$$\boldsymbol{TP_{rate} = \frac{TP}{TP + FN}} \quad (3)$$

$$\boldsymbol{FP_{rate} = \frac{FP}{FP + TN}} \quad (4)$$

The AUC can be calculated based on the approximate average trapezoidal plane for curves made by TP_{rate} and FP_{rate} [9].

$$\boldsymbol{AUC = \frac{1+ TP_{rate} - FP_{rate}}{2}} \quad (5)$$

3. Result and Discussion

Based on the model proposed in figure 1, to find out the performance of the basic model applied by the Decision Tree algorithm as a classification without being optimized. The second model is the integration of Decision Tree with Forward Selection. And the third model integrates Decision Tree and Backward Selection.

Table 2 Average of model performance

Model	**Performance**	
	Accuracy	**AUC**
Decision Tree	64.17%	0.6417
Forward Selection + Decision Tree	71.74%	0.7171
Backward Selection + Decision Tree	69.42%	0.6942

Based on figures 2 and 3, it can be seen about Parkinson's prediction using the Decision Tree algorithm and the Forward Selection feature selection technique get better values than other models. The Performance Accuracy Model and AUC are the same for all.

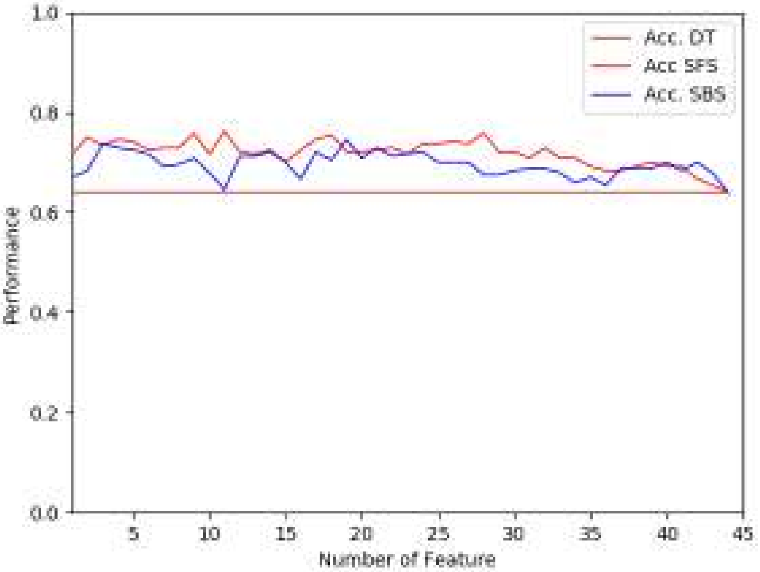

Figure 2 Graph of accuracy models

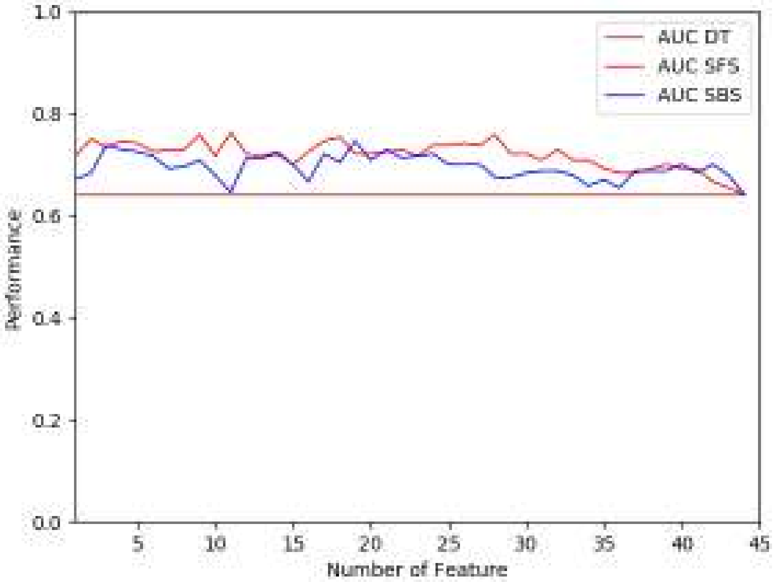

Figure 3 Graph of AUC models

Based on the number of results of the performance validation of the Accuracy and AUC models having the same value, it can be seen that the model that implements the Decision Tree and integrated with the Forward Selection technique shows better performance results compared to other models.

4. Conclusion

Prediction of Parkinson's disease is one of the important topics because early detection is needed so that the developmental disease can be prevented as early as possible because there is no cure for this disease. The proposed model shows the results that no model produces a very good performance. The results of this study indicate that a model that integrates Decision Tree and Forward Selection provides better performance values. The proposed model can help predict Parkinson's disease better. For further research, we suggest applying bagging techniques to reduce misclassification.

References

[1] Aich S, Sain M, Park J, Choi K W and Kim H C 2018 A mixed classification approach for the prediction of Parkinson’s disease using nonlinear feature selection technique based on the voice recording *Proc. Int. Conf. Inven. Comput. Informatics, ICICI 2017* 959–62

[2] Soliman A B, Fares M, Elhefnawi M M and Al-Hefnawy M 2016 Features selection for building an early diagnosis machine learning model for Parkinson’s disease *2016 3rd Int. Conf. Artif. Intell. Pattern Recognition, AIPR 2016* 133–6

[3] Sonu S R, Prakash V, Ranjan R and Saritha K 2018 Prediction of Parkinson’s disease using data mining *2017 Int. Conf. Energy, Commun. Data Anal. Soft Comput. ICECDS 2017* 1082–5

[4] Ul Haq A, Li J, Memon M H, Khan J, Din S U, Ahad I, Sun R and Lai Z 2019 Comparative Analysis of the Classification Performance of Machine Learning Classifiers and Deep Neural Network Classifier for Prediction of Parkinson Disease *2018 15th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. ICCWAMTIP 2018* **1** 101–6

[5] Aich S, Kim H C, Younga K, Hui K L, Al-Absi A A and Sain M 2019 A Supervised Machine Learning Approach using Different Feature Selection Techniques on Voice Datasets for Prediction of Parkinson’s Disease *Int. Conf. Adv. Commun. Technol. ICACT* **2019-February** 1116–21

[6] Albu A 2017 From logical inference to decision trees in medical diagnosis *2017 E-Health Bioeng. Conf. EHB 2017* 65–8

[7] Gavankar S S and Sawarkar S D 2017 Eager decision tree *2017 2nd Int. Conf. Converg. Technol.*

I2CT 2017 **2017-January** 834–40
[8] Jiawei H, Kamber M, Han J, Kamber M and Pei J 2012 *Data Mining: Concepts and Techniques*
[9] Gorunescu F 2011 *Data mining: concepts and techniques*