

Robust attribute selection to improve the Parkinson's disease classification: a hybrid approach

Ameer K. Al-Mashanji¹, Laith Hamid Alhasnawy², Aseel Hamoud Hamza³

¹Department of Software, College of Information Technology, University of Babylon, Babylon, Iraq

²Department of Computer, College of Science, University of Babylon, Babylon, Iraq

³College of Law, University of Babylon, Babylon, Iraq

Article Info

Article history:

Received Nov 15, 2022

Revised Feb 23, 2023

Accepted Mar 24, 2023

Keywords:

Attribute selection

Classification

Machine learning techniques

Parkinson's disease

Performance measures

ABSTRACT

Parkinson's disease (PD) is known to be a neurodegenerative syndrome that progresses chronically. As a result of the damage or death of brain neurons that generate dopamine patients tend to face difficulty when performing simple everyday tasks like walking, writing, or speaking. The main contribution of this work presents a hybrid method for improving predicting PD. This methodology has been obtained by means of testing a number of different combinations of classification algorithms and approaches for selecting attributes. A total of three attributes selection methods (correlation, information gain, and variance threshold) and three classifiers (decision trees (DT), naive bayes (NB), and support vector machine (SVM)) have been adopted. The speech data set provided by University of California-Irvine (UCI) machine learning (ML) repository is adopted to analyze the performance of different combinations. The combination of information gain and DT classifier achieved the best performance rather than other combination methods, reaching a classification accuracy of (97.43%). Finally, an additional comparison of the performance analysis with the results of previous studies was made and it was found that the proposed methodology proved to outperform the results of other studies conducted in this field.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Laith Hamid Alhasnawy

Department of Computer, College of Information Technology, University of Babylon

Babylon, Iraq

Email: laith.alhasnawi4@uobabylon.edu.iq

1. INTRODUCTION

Parkinson's disease (PD) can be defined as a neurodegenerative syndrome that leads to the progressive deterioration of motor abilities as a result of the damaged brain cells responsible for the production of dopamine [1]. Common symptoms include shakiness, difficulty moving, behavioral issues, depression, dementia, tremor, handwriting alterations, muscle rigidity, and posture/balance impairment. The main symptoms all together are also known as Parkinsonism or Parkinson's syndrome [1]. Alterations in a patient's voice are a commonly occurring symptom whose identification could be done by means of analyzing the patient's speech data. It has been observed that the patient's voice is affected gradually along as the disease intensifies and they may start stuttering [2].

A Parkinson's syndrome affects both male and female patients, and tends to develop after the age of 60. However, there are cases of PD in patients before the age of 50 [3]. The fact that the (early) diagnosis of Parkinson's is rather challenging has been the motivation to develop a decision support system (DSS) for helping the medical staff in diagnosing Parkinson's. Such a system could function as a second opinion in

diagnosing Parkinson's, as the use of machine learning (ML) reduces the likelihood of errors [4]. Since researchers have used several ways such as single photon emission computed tomography (SPECT), magnetic resonance imaging (MRI), and handwritten images, as well as changes in a speech called dysphonia for PD's diagnosis, this work involves the use of speech changes in diagnosing PD [5]. Different ML techniques have been used over time in building DSSs, such as preprocessing, attribute selection, classification, and validation steps. ML helps in analyzing disease patterns in medical data sets, as well as making decisions in a shorter time [6].

Pre-processing techniques cover procedures like data normalization, attribute selection, and balancing. Attribute selection decreases computational expenses and expands its accuracy. First, the attributes are selected via three attribute selection methods: correlation, information gain, and variance threshold. The reduced attribute subset was adopted to train and test the classifiers in identifying the ideal combinations of attribute method and classifiers. Second, the Parkinson's speech dataset was found to be of no balance, as 147 out of 195 samples were from individuals suffering from Parkinson's. Therefore, shuffle was applied for treating the lack of balance. At last, a performance analysis for the three classifiers (naive bayes (NB), decision trees (DT), and support vector machine (SVM)) is conducted on full and reduced attribute sub-sets. It is noticed that combining the information gain algorithm with the DT classifier leads to more favorable results than the other methods.

Gupta *et al.* [7] utilized two ML algorithms in analyzing the artificial neural networks (ANN) and random forest (RF) classifiers for predicting PD. The data set used by the authors in their experiment is obtained from the repository located at the University of California-Irvine (UCI). The adopted dataset contains 754 attributes without missing values. The class labels (0) and (1) indicate whether or not the disease occurs. The principal component analysis (PCA) is applied for selecting the optimal attributes in the classifying process. The experimental results indicate that using ANN and PCA combined leads to better results than using it in combination with the RF classifier. Senturk [8] made use of ML algorithms for diagnosing PD. The attributes were chosen via the recursive feature elimination (RFE) method, to determine the best attributes. ANN, SVM, and the regression tree were implemented in the classification process. Combining RFE and SVM realized an accuracy rate of 93.84%.

Tuncer *et al.* [9] used vowels to diagnose Parkinson's syndrome. The attributes were selected using the relief-based method. Eight classification algorithms were used in their work. The k-nearest neighbor (KNN) classifier achieved an accuracy of 92.46% and thereby outperformed the rest of the classifiers. Sharma *et al.* [10] used a variety of ML algorithms for diagnosing Parkinson's syndrome. They used the PD speech datasets which are provided by the UCI's ML repository. The authors implemented the algorithm of modified gray wolf optimization to select the best attributes. Three classifiers have been adopted: RF, KNN, and DT. The experimental results showed that the best accuracy achieved by the classifiers based on the speech dataset is 93.87%. The content of this article is divided in the following way: section 2 describes outlines of materials and methods, section 3 states the observations made throughout the experiment and followed by analysis, section 4 states the results and discussion, and at last, section 5 states the conclusion.

2. MATERIALS AND METHOD

2.1. Parkinson's-speech dataset

Studies indicate a constant patterning of vocal deterioration in the main cases of Parkinson's. Therefore, this work addresses the distinction between patients who suffer from Parkinson's from those who are healthy, via the analysis of patients' speech signals [11]. The benchmark Parkinson's-speech dataset used in the present paper is an open access dataset and can be downloaded freely available at the UCI. It contains 195 instances with 23 numeric attributes for Parkinson's patients whose voice have been recorded for study purposes.

The data indicates the status by means of binary values: (0) states that the patients suffer from PD. The proposed method is examined by means of the same UCI dataset [11]. Table 1 states the information on the dataset. Table 2 depicts the statistical issues of classes in the dataset. Figure 1 shows the bar chart of distribution classes in the dataset, while the description of the 23 attributes are shown detail in Table 3.

Table 1. Description of selected datasets

Name of dataset	Parkinson speech
Number of instances	195
Number of attributes	23
Class variable	Healthy and Parkinson

Table 2. The statistics of classes in the dataset

Class	Instances	Distribution (%)
Parkinson (1)	147	75.38
Healthy (0)	48	24.62
Total	195	100

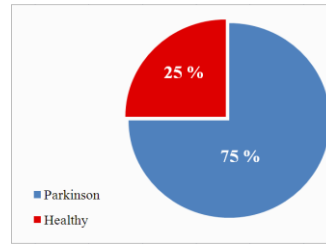


Figure 1. Class distribution in the dataset

Table 3. Attribute description

Attribute name	Description of abbreviation
#1_(MDVP-F0 (Hz))	Multidimensional voice_program represents average-vocal fundamental frequencies
#2_(MDVP-Fhi (Hz))	Multidimensional voice_program represents maximum-vocal fundamental frequencies
#3_(MDVP-Flo(Hz))	Multidimensional voice_program represents minimum-vocal fundamental frequencies
#4_(MDVP-Jitter (%))	MDVP_jitter in percent
#5_(MDVP-(Abs))	MDVP_absolute jitter in micro-seconds
#6_(MDVP RAP)	MDVP_relative amplitude perturbation
#7_(MDVP-PPQ)	MDVP_period perturbation quotient
#8_(Jitter-DDP)	Average absolute difference of differences between cycles, divided by the average period
#9_(MDVP-Shimmer)	MDVP_local shimmer
#10_(MDVP-Shimmer (dB))	MDVP_local shimmer in decibels
#11_(MDVP-APQ)	MDVP_amplitude perturbation-quotient
#12_(Shimmer-APQ3)	3-Point_amplitude perturbation-quotient
#13_(Shimmer APQ5)	5-Point amplitude perturbation quotient
#14_(Shimmer-DDA)	Average absolute difference between consecutive differences between the amplitude of consecutive periods
#15_(NHR)	Noise harmonic-ratio
#16_(HNR)	Harmonics noise-ratio
#17_(DFA)	Detrended fluctuation analysis
#18_(Spread1)	Fundamental frequencies nonlinear measures
#19_(Spread2)	Nonlinear measures of fundamental frequencies
#20_(D2)	Correlation dimensions
#21_(PPE)	Pitch period-entropy
#22_(RPDE)	RPDE_recurrence period density entropy
#23_(Status)	(0) Healthy; (1) Parkinson

2.2. Attribute ranking method

Attributes in this type of method are selected based on specific performance metrics with no regard to prediction algorithms. Therefore, these methods are used before the prediction models [12]. Three ranking methods have been implemented to evaluate and rank each attribute in the Parkinson's-speech dataset [13]. The attribute ranking methods adopted within this system are outlined in the following sections.

2.2.1. Correlation method

This method individually measures the correlation between each attribute in the dataset and the target class [14]. The attribute weight ranges between 1 and -1, so that the attribute is considered very weakened if its weight is close to zero, meaning that the attribute is not related to the target class, while it is considered very strong if its weight is close to ± 1 , meaning that the attribute is highly related to the target class [14]. The correlation between each attribute and the target class is calculated in (1):

$$cor(x, y) = \frac{\sum (X_i - \underline{X})(Y_i - \underline{Y})}{\sqrt{\sum (X_i - \underline{X})^2} \sqrt{\sum (Y_i - \underline{Y})^2}} \quad (1)$$

where X is representing the attribute, Y is representing the target class, \underline{Y} is representing the average of the target class, and \underline{X} is representing the average of the attribute.

2.2.2. Information gain method

It is an essential and commonly used method for selecting attributes. The significance of attributes is determined in comparison with the general class. In case the information gain value of an attribute exceeds a particular threshold, it is considered to be an important attribute. Therefore, it is often adopted in reducing the

dimensions and increasing the efficiency of the classifying process. The information gain of each attribute with the target class could be obtained using (2) [15]:

$$IG(S, t) = E(S) - E(S|t) \quad (2)$$

where $E(S)$ is the entropy of a random variable S (target class) and $E(S|t)$ is the conditional entropy of S given the value of the attribute (t).

2.2.3. Variance threshold method

Variance threshold method is an attribute selection method that removes all the low variance attributes from the dataset that are of no great use in modeling. Constant attributes show constant values in all observations of the dataset. These attributes provide no information that allows ML models to predict the target efficiently [16]. The variance for each attribute is calculated in (3):

$$\text{Variance} = \sum_{i=1}^n (Xi - \underline{X})^2 / n \quad (3)$$

where Xi is the values of an attribute, \underline{X} is the mean, and n the number of instances.

2.2.4. Decision trees technique

DT can be described as a prediction model used in the mapping of observations made of a certain item, to conclude upon the item's target values. The structure of DT includes root, internal, and leaf nodes. It could be defined as some sort of flow chart with a tree-like structure, whereby internal nodes denote test conditions on attributes, branches represent the results of test conditions, and the leaf-terminal nodes are all assigned class labels [17]. The highest top-node is known to be root of the DT. Overall, this type of structures has a "divide and conquer" approach, whereby all of the paths form decision rules by themselves. The benefits of DT include the fast classification processes, strong learning abilities, and relatively simple structures [17].

2.2.5. Naïve bayes technique

NB is a simple classification method that is based on identifying the probabilistic relations among classes and attributes [18]. It depends on the bayesian theory for computing the target probability using values of certain predictors or attributes. It is more mode favorable than other probability classifiers, as it computes the most likely output using the provided input [19], [20].

2.2.6. Support vector machine

SVM is a classifying algorithm used with both non-linear and linear data. It works by transforming the originally used training data into higher dimensions via non-linear mapping. Next, the model aims to identify the hyper-plane linear optimal separation [21]. Using suitable non-linear mapping towards higher dimensions that are more sufficient, the hyper-plane can separate the data of two classes. To classify data, SVM maximizes the margins of both classes and minimizes classification errors. Other applications of SVM include cases of regression [22], [23].

2.3. Performance measures

The present study adopts four common performance metrics for evaluating the accuracy of classification algorithms. The confusion matrix which appears in Figure 2 records the correct and incorrect classification results to measure the quality of the classifier [24], [25].

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Figure 2. Class distribution in the dataset

where TP is true-positive, FP is false-positive, FN is false-positive, and TN is true-negative.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (4)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (5)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (6)$$

$$F - Measure = \frac{2 * (Recall * Precision)}{(Recall + Precision)} \quad (7)$$

3. THE PROPOSED METHOD OF THE STUDY

The architecture of the proposed methodology involves for three stages to achieve the goal of this study. In the first step, ranking methods have been applied for attribute selection. In the second step, classification models are applied for the prediction task. Finally, the classification models are evaluated based on various measures. The block diagram of the suggested method stages is explained in Figure 3.

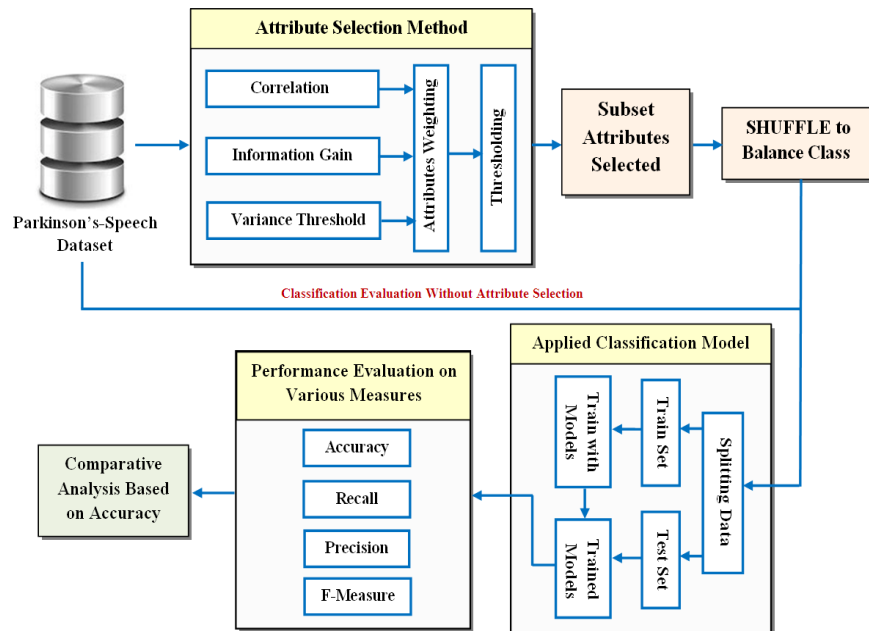


Figure 3. The architecture of the proposed method

Stage 1 attribute selection: several attribute selection methods are used on the Parkinson's-speech dataset to reduce the attribute space. Thus, a subset of the most important attributes is selected among the original ones. The attribute selection methods used are correlation, information gain, and variance threshold. These methods are applied before the classification model for selecting the attributes according to the performance measures, with no regard to the classification algorithms. The key role of the attribute selection method is identifying the most important attributes which directly affect the target class (Parkinson and healthy). These methods evaluate the attributes and give a different rank value for each one of them. All weak attributes have been deleted through a predefined threshold. As for the class imbalance problem, shuffle has been used to handle this issue. Stage 2 prediction stage: this stage represents the most important step in the proposed method. Three different classification models have been used for validating how accurate the selection of these attributes, ensuring that these selected attributes are indeed most likely to influence the target class (Parkinson and healthy). Stage 3 evaluation of prediction model: in this stage, accuracy, recall, precision, and F-measure performance measures are utilized for measuring the efficiency of the classification models.

4. RESULTS AND DISCUSSION

The suggested methodology follows the concept of classification tasks to classify the class label (Parkinson or healthy) in the Parkinson's-speech dataset. The hold-out-validation method (80% for training and 20% for testing) is used for validating the results. At first, the weight of each attribute is calculated using the correlation method. Next, the top (12) attributes are determined depending on a predefine threshold value. In the information gain method, the weight for each attribute is computed, and any attribute with a weight less than the predefine threshold value is discarded. The output of this method performs well, as only 5 attributes are selected. In the variance threshold method, the variance value for each attribute is computed and any attribute that does not achieved a predefine threshold is neglected from the dataset. The yielded results from the variance threshold method are 12 attributes. Table 4 presents the selection of the attributes via the attribute selecting method.

Table 4. Attributes selected by attribute selection methods

Attribute selection method	No of attribute selected	Attributes name
Correlation	12	Shimmer: DDA, Shimmer: APQ3, MDVP: Shimmer(dB), Shimmer: APQ5, HNR, MDVP: APQ, MDVP: Shimmer, MDVP: Flo(Hz), MDVP: Fo(Hz), spread2, PPE, spread1
Information gain	5	PPE, spread 1, MDVP: Fo(Hz), spread 2, MDVP: APQ
Variance threshold	12	MDVP: Fo(Hz), MDVP: Fhi(Hz), MDVP: Flo (Hz), MDVP: Shimmer(dB), NHR, HNR, RPDE, DFA, spread1, spread2, D2, PPE

After the selection of attributes via several attribute selection methods, the efficiency of three ML classifiers via differing attribute sub-sets was evaluated. It has been found the results for all classifiers with attribute selection methods archive the best accuracy. Table 5 states the efficiency rates of NB, DT, and SVM classifiers for all attributes once and again for the reduced attribute sub-sets.

Table 5. Performance of classifiers with attribute selection methods

Attribute selection algorithm	All attributes	Correlation	Information gain	Variance threshold
NB classifier				
Accuracy (%)	69.23	82.05	89.74	84.61
Precision (%)	75	81	87	82
Recall (%)	78	88	93	89
F-measure (%)	69	81	88	83
DT classifier				
Accuracy (%)	84.61	92.30	97.43	94.87
Precision (%)	82	92	98	97
Recall (%)	84	89	95	91
F-measure (%)	83	90	97	93
SVM classifier				
Accuracy (%)	87.17	89.74	94.87	92.30
Precision (%)	92	94	97	95
Recall (%)	79	82	91	86
F-measure (%)	83	86	93	90

Figure 4 presents an analysis whereby the enhancement in classification accuracy is compared. It draws a comparison between three classifiers via the attribute selection methods, as an improvement has been observed in the reduced attribute sub-sets. It has been found that the information gain method has a better performance than the alternative selecting methods having rates of 89.74%, 97.43%, and 94.87% for NB, DT, and SVM, respectively.

Table 6 compares the suggested methodology and the methodologies in previous studies. Figure 5 illustrates graphically the accuracy improvement of the proposed methodology as compared to the previous methodologies which has been implemented through other authors. All codes conducted to implement the proposed hybrid methods were executed in Python language (version 3.7) with jupyter notebook lab under Windows 64-bit OS environment, Intel Core i7 processor, 6 GB memory, and a NVIDIA GeForce GTX 2 GB graphics.

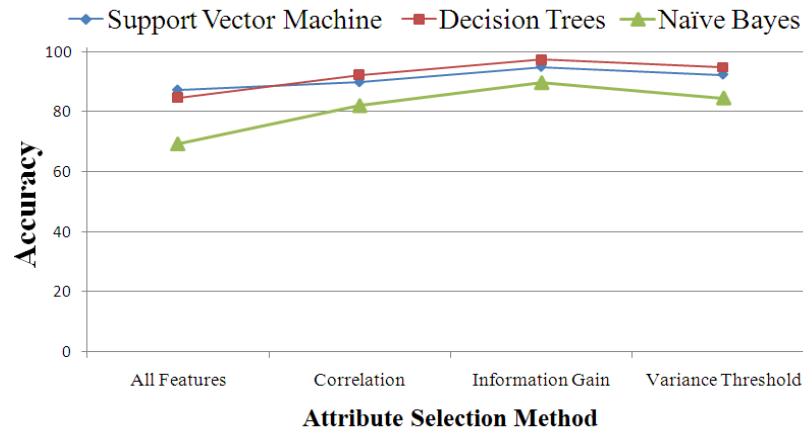


Figure 4. Comparative analysis of attribute selection methods with classifiers

Table 6. Performance comparison with previous studies

Reference	Attribute selection methods	Classifies model	Accuracy (%)
[17]	Cuttle-fish algorithm	KNN DT	92.19
[10]	Grey-wolf algorithm	RF, KNN, DT	93.87
[8]	RFE and attribute significant algorithm	SVM, classification trees, ANN	93.84
[18]	Genetic algorithm, extra tree, and mutual information	NB, RF KNN	95.58
Proposed method	Correlation, information gain, and variance threshold	SVM, DT, NB	97.43

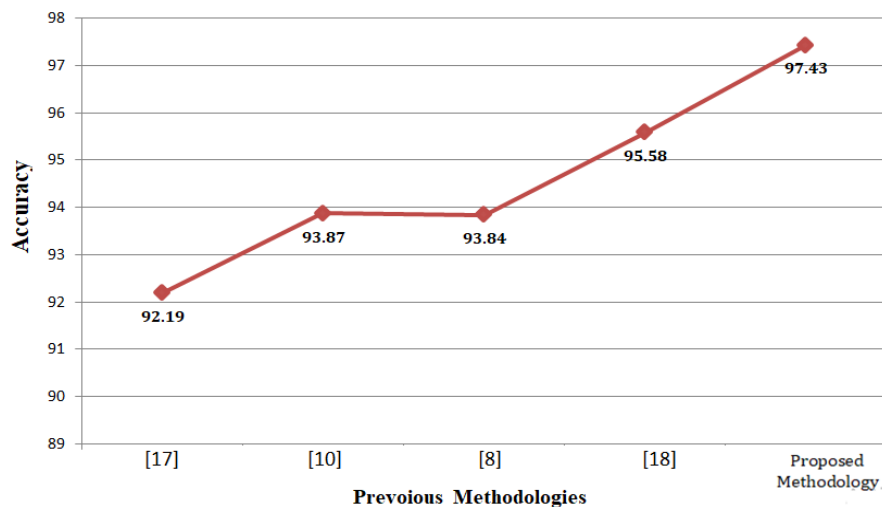


Figure 5. Comparison of accuracy with previous studies

5. CONCLUSION

This study contributed to presenting a proposed hybrid approach to improve the accuracy of classification (class label: Parkinson or healthy) in the Parkinson's-speech dataset. No particular method has been assigned for selecting a universal attribute and a universal classifier for a medical data set. To find the best results, researchers have to try different methods to achieve the best combination. The main aim of using attribute selecting methods is to select the best subset of attributes by eliminating the attributes which no predictive information. The results indicate that using the attribute selecting method is beneficial due to the reduction in time and increase in simplicity and accuracy. The hybrid method that has been introduced in this work has proven to yield better results than alternative approaches, realizing an accuracy of 97.43%. It can therefore be concluded that the proposed method does not substitute the healthcare experts, but rather functions as a second opinion in diagnosing PD. Further research is aimed to study the efficiency of the suggested methodology on other speech and voice data sets.




ACKNOWLEDGEMENTS

The authors thank the Department of Software, College of Information Technology for their constant encouragement to complete this manuscript.




REFERENCES

- [1] F. N. Emamzadeh and A. Surguchov, "Parkinson's disease: Biomarkers, treatment, and risk factors," *Frontiers in Neuroscience*, vol. 12, pp. 1–14, 2018.
- [2] C. O. Sakar *et al.*, "A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform," *Applied Soft Computing Journal*, vol. 74, pp. 255–263, 2019, doi: 10.1016/j.asoc.2018.10.022.
- [3] S. G. Reich and J. M. Savitt, "Parkinson's Disease," *Medical Clinics of North America*, vol. 103, no. 2, pp. 337–350, 2019, doi: 10.1016/j.mcna.2018.10.014.
- [4] T. A. Zesiewicz, Y. Bezchlibnyk, N. Dohse, and S. D. Ghanekar, "Management of Early Parkinson Disease," *Clinics in Geriatric Medicine*, vol. 36, no. 1, pp. 35–41, 2020, doi: 10.1016/j.cger.2019.09.001.
- [5] C. Kotsavasiloglou, N. Kostikis, D. H. -Varsakelis, and M. Arnaoutoglou, "Machine learning-based classification of simple drawing movements in Parkinson's disease," *Biomedical Signal Processing and Control*, vol. 31, pp. 174–180, 2017, doi: 10.1016/j.bspc.2016.08.003.
- [6] S. Raval, R. Balar, and V. Patel, "A Comparative Study of Early Detection of Parkinson's Disease using Machine Learning Techniques," in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, 2020, pp. 509–516, doi: 10.1109/ICOEI48184.2020.9142956.
- [7] I. Gupta, V. Sharma, S. Kaur, and A. K. Singh, "PCA-RF: An Efficient Parkinson's Disease Prediction Model based on Random Forest Classification," *Arxiv-Computer Science*, vol. 1, pp. 1–10, 2022.
- [8] Z. K. Senturk, "Early diagnosis of Parkinson's disease using machine learning algorithms," *Medical Hypotheses*, vol. 138, pp. 1–5, 2020, doi: 10.1016/j.mehy.2020.109603.
- [9] T. Tuncer, S. Dogan, and U. R. Acharya, "Automated detection of Parkinson's disease using minimum average maximum tree and singular value decomposition method with vowels," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 211–220, 2020, doi: 10.1016/j.bbe.2019.05.006.
- [10] P. Sharma, S. Sundaram, M. Sharma, A. Sharma, and D. Gupta, "Diagnosis of Parkinson's disease using modified grey wolf optimization," *Cognitive Systems Research*, vol. 54, pp. 100–115, 2019, doi: 10.1016/j.cogsys.2018.12.002.
- [11] V. Despotovic, T. Skovranek, and C. Schommer, "Speech Based Estimation of Parkinson's Disease Using Gaussian Processes and Automatic Relevance Determination," *Neurocomputing*, vol. 401, pp. 173–181, 2020, doi: 10.1016/j.neucom.2020.03.058.
- [12] B. Remeseiro and V. B. -Canedo, "A review of feature selection methods in medical applications," *Computers in Biology and Medicine*, vol. 112, pp. 1–35, 2019, doi: 10.1016/j.combiomed.2019.103375.
- [13] D. Gupta, S. Sundaram, A. Khanna, A. E. Hassaniien, and V. H. C. de Albuquerque, "Improved diagnosis of Parkinson's disease using optimized crow search algorithm," *Computers and Electrical Engineering*, vol. 68, pp. 412–424, 2018, doi: 10.1016/j.compeleceng.2018.04.014.
- [14] O. Cigdem and H. Demirel, "Performance analysis of different classification algorithms using different feature selection methods on Parkinson's disease detection," *Journal of Neuroscience Methods*, vol. 309, pp. 81–90, 2018, doi: 10.1016/j.jneumeth.2018.08.017.
- [15] A. A. Imran, A. Rahman, H. Kabir, and S. Rahim, "The Impact of Feature Selection Techniques on the Performance of Predicting Parkinson's Disease," *International Journal of Information Technology and Computer Science*, vol. 10, no. 11, pp. 14–29, 2018, doi: 10.5815/ijitcs.2018.11.02.
- [16] Y. S. Ambarwati and S. Uyun, "Feature Selection on Magelang Duck Egg Candling Image Using Variance Threshold Method," in *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 2020, pp. 694–699, doi: 10.1109/ISRITI51436.2020.9315486.
- [17] D. Gupta *et al.*, "Optimized cuttlefish algorithm for diagnosis of Parkinson's disease," *Cognitive systems research*, vol. 52, pp. 36–48, 2018.
- [18] R. Lamba, T. Gulati, H. F. Alharbi, and A. Jain, "A hybrid system for Parkinson's disease diagnosis using machine learning techniques," *International Journal of Speech Technology*, vol. 25, no. 3, pp. 583–593, 2022, doi: 10.1007/s10772-021-09837-9.
- [19] M. M. Saritas and A. Yasar, "Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 7, no. 2, pp. 88–91, 2019, doi: 10.18201/ijisae.2019252786.
- [20] U. N. Dulhare, "Prediction system for heart disease using Naive Bayes and particle swarm optimization," *Biomedical Research*, vol. 29, no. 12, pp. 2646–2649, 2018.
- [21] W. Deng, R. Yao, H. Zhao, X. Yang, and G. Li, "A novel intelligent diagnosis method using optimal LS-SVM with improved PSO algorithm," *Soft Computing*, vol. 23, no. 7, pp. 2445–2462, 2019, doi: 10.1007/s00500-017-2940-9.
- [22] S. Huang, C. A. I. Nianguang, P. P. Pacheco, S. Narandes, Y. Wang, and X. U. Wayne, "Applications of support vector machine (SVM) learning in cancer genomics," *Cancer Genomics and Proteomics*, vol. 15, no. 1, pp. 41–51, 2018, doi: 10.21873/cgp.20063.
- [23] L. H. S. Vogado, R. M. S. Veras, F. H. D. Araujo, R. R. V. Silva, and K. R. T. Aires, "Leukemia diagnosis in blood slides using transfer learning in CNNs and SVM for classification," *Engineering Applications of Artificial Intelligence*, vol. 72, pp. 415–422, 2018, doi: 10.1016/j.engappai.2018.04.024.
- [24] B. Seref and E. Bostanci, "Performance Comparison of Naïve Bayes and Complement Naïve Bayes Algorithms," in *2019 6th International Conference on Electrical and Electronics Engineering (ICEEE)*, 2019, pp. 131–138, doi: 10.1109/ICEEE2019.2019.00033.
- [25] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018.




BIOGRAPHIES OF AUTHORS

Ameer K. Al-Mashanji    is a lecturer at the University of Babylon, College of Information Technology, Department of Software, Iraq. He obtained his B.Sc. and M.Sc. degrees in 2008 and 2020, respectively. His main research interests include data mining, bioinformatics, computer vision, and machine learning. He can be contacted at email: amir.mashanji@uobabylon.edu.iq.



Laith Hamid Alhasnawy    is a lecturer at the Department of Computer, College of Science, University of Babylon, Iraq. He obtained his B.Sc. and M.Sc. degrees in 2008 and 2020, respectively. His main research interests include advance network, artificial intelligent, and machine learning. He can be contacted at email: laith.alhasnawi4@uobabylon.edu.iq.



Aseel Hamoud Hamza    is a lecturer at the College of Law, University of Babylon. She received her B.Sc. from the College of Information Technology at University of Babylon and completed her M.Sc. in the field of information security from the Department of Computer Science, College of Women's Sciences, University of Babylon, from 2017 to 2019, and she worked in the field of teaching from 2019 to 2022 in University of Babylon. She can be contacted at email: aseel.hamod@uobabylon.edu.iq.