

Research Article

Biocomposite's Multiple Uses for a New Approach in the Diagnosis of Parkinson's Disease Using a Machine Learning Algorithm

Abdallah Al-Husban,¹ Mustafa Mahdi Abdulridha,^{2,3} A. A. Hamad Mohamad ^{4,5} and Abdelrahman Mohamed Ibrahim ⁶

¹Department of Mathematics, Faculty of Science and Technology, Irbid National University, P.O. Box: 2600 Irbid, Jordan

²Department of Medical Instruments Engineering Techniques, Al-Farahidi University, Baghdad 10021, Iraq

³Department of Medical Laboratory Techniques, Al-Turath University College, Baghdad 10021, Iraq

⁴Department of Medical Laboratory Techniques, Dijlah University College, Baghdad 10021, Iraq

⁵The University of Mashreq, Research Center, Baghdad, Iraq

⁶Accounting and Financial Management, School of Management Studies, University of Khartoum, Sudan

Correspondence should be addressed to Abdelrahman Mohamed Ibrahim; amibrahim@uofk.edu

Received 9 June 2022; Revised 26 June 2022; Accepted 3 July 2022; Published 14 July 2022

Academic Editor: Rabia Rehman

Copyright © 2022 Abdallah Al-Husban et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Neurodegenerative diseases drastically affect human beings without distinction; it does not matter if they are male or female. Sometimes, it is not clear why a person in their life developed a well-known disease in the world such as Parkinson's disease (PD). Nowadays, various novel machine learning-based algorithms for evaluating Parkinson's disease have been designed. The most recent strategy, which was developed using deep learning and can forecast the severity of Parkinson's disease, is the one described here. To identify this disease, a thorough medical history, previous treatment history, physical examinations, and some blood tests and brain films must be completed. Diagnoses are more critical since they are less expensive and less time-consuming. Voice data from 253 people used in the current study corroborates the doctor's diagnosis of Parkinson's disease. To acquire the best results from the data, preprocessing is done. To perform the balancing procedure, a systematic sampling strategy was used to select the data that would be analyzed. Several data groups were constructed using a feature selection technique based on the label's effect strength. Classification algorithms and performance evaluation criteria employ DT, SVM, and kNN. The classification algorithm and data group with the highest performance value were chosen, and the model was created due to this selection. The SVM approach was employed when constructing the model, and 45% of the original data set data were used. The data was sorted from most relevant to least important. 86% performance accuracy was achieved, in addition to excellent results in all other areas of the project. As a result, it has been established that medical decision support will be provided to the doctor with the assistance of the data set obtained from the speech recordings of the individual who may have Parkinson's disease and the model that has been developed.

1. Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder due to the loss of neurons in the substantia nigra, which decreases dopamine levels, an important neurotransmitter whose primary function is the correct control of movements [1]. It is a chronic and incurable disease that

manifests itself through a progressive loss of the ability to coordinate actions, presenting several peculiar characteristics such as tremor at rest, slowness in the initiation of movements, difficulty in speaking, and muscular rigidity. PD is characterized as slowness of movement (bradykinesia), tremors, and convulsions [2]. In addition to these, sleep disturbance, symptoms of depression, and speech disorder are

observed [3]. Speech disorder includes difficulties affecting social life, such as low voice, dull speech, inability to start speaking, pronunciation errors, and inability to adjust the volume while speaking [4].

A simple test cannot determine whether a person has PD or not. A neurologist doctor requests biochemical tests and brain tomography from patients to diagnose the disease and understand whether another disease condition causes the disease. In addition, some physical examinations are required to evaluate functional adequacy of the legs and arms, muscle condition, free gait, and balance. As the patients are usually aged greater than 60, the required tests are complex for people. Because of all these difficulties, simpler and more reliable methods are needed to diagnose PD [5–7].

For a decade, many researchers have shown great interest in offering a solution for diagnosing Parkinson's by voice. Initially, voice recordings made in the laboratory used a series of characteristics extracted to use as predictors to classify Parkinson's patients and healthy controls. These recordings, in general, used to be sustained vowels since, as demonstrated [6, 7], they offer more information than words or short phrases. Usually, a selection of the extracted features is carried out to improve the effectiveness of the data mining methods (kNN, SVM, or random forests). In this way, the characteristics were analyzed, and, to avoid redundancies and simplify the problem, those that were strongly correlated were eliminated. There are various algorithms to follow for the selection, and we found an excellent comparison of four of them in [8]. This selection of characteristics is still present and occupies most of the articles, and there are even specific ones such as [9]. Using these traditional methods, it is possible to discriminate whether a patient suffers from PD or not. It is even possible to predict their Unified Parkinson's Disease Rating Scale level; it is a scale for assessing Parkinson's disease that measures motor and nonmotor symptoms and is very useful for monitoring patients. It should be noted that until now, the vast majority of studies have evaluated their models using cross-validation or similar, without noticing that different recordings of the same individual are found both in training and in the test. This may be a significant reason why they get such optimistic results.

Apart from these, there are also studies on nonsmall data sets [7]. A comparison was made between the data set in which the adjustable Q -factor wavelet transform was used and the data set in which this transform was not used. It has been stated that the conversion increases the accuracy rate, which is one of the performance criteria, albeit at a low rate. To get more relevant results, it is necessary to use more than one data set in a larger data set and balance the data set [10, 11] used the subjects' walking data to diagnose PD. The data set was grouped according to the age factor. The proposed model is constructed using the Dual Density 1-D Wavelet Transform method. Salari et al. found high accuracy rates in different studies conducted recently [12]. A comprehensive data set was used in the studies, but balancing was not done. The data sets generally used in the literature are fundamental vocal frequency, the amount of variation in frequency, and variation in amplitude as features extracted for use in machine learning.

2. Materials and Method

The investigation followed the flow chart depicted in Figure 1. Separate groupings of similar attribute values were created to ensure an even distribution of data. The data sets were sorted from most relevant to least relevant using a feature selection technique. These ordered data groups were divided into feature groups at a certain percentage, and the performances of each data group were evaluated with classification algorithms. Necessary operations from the arrangements in the data set to the performance evaluation stage were carried out in the MATLAB program.

2.1. Parkinson's Disease Data Set. The University of Baghdad's College of Medicine's Machine Learning Repository provided us with the data we needed for our investigation. The study included 188 Parkinson's disease patients (107 males and 81 females) and 64 control subjects (23 males and 41 females). Involved individuals are in the range of 33 to 87 years old. The label in the data set represents only one patient group and zero healthy groups. There were 757 measurements recorded from the 253 participants, who were asked to repeat the/a/vowel three times each. From the data collected, a tag was one of 753 attributes created.

2.2. Data Preprocessing. The researchers developed several processes [13] to prepare the data set for analysis. The following sections outline the data preprocessing steps used in this study.

2.2.1. Separating the Data Set into Related Attribute Groups. There are 753 features in the raw data set that can be traced back to certain feature groups. In addition to basic characteristics like intensity parameters and formant frequencies (formal phonetic frequencies), wavelet features and MFCCs (Mel-Frequency Cepstral Coefficients) are some of the features that can be discovered (wavelet properties) [14]. Because of the similarity in the characteristics of the new time features, it was decided to consolidate the three categories (intensity, formant frequency, and bandwidth parameters). As shown in the schematic in Figure 2, data groupings were constructed that contained 5 core attribute groups and a group that included all attributes.

2.2.2. Balancing the Data Set. According to [15], an "unbalanced data set" refers to a data collection that has values for each label class that differs from one another. It is possible that a nonbalanced data collection can result in false accuracy values utilized in performance evaluation, resulting in wrong conclusions being drawn. It was decided to apply a systematic sampling approach to eliminate this unwanted circumstance. This process balanced the system by employing the undersampling method, and the unbalanced state was then retrieved by reversing the procedure. In the down-sampling approach, the outnumbered tag class is regarded as the undersampled tag class, and vice versa. The data set for this investigation consisted of 757 measurements, with 192 having healthy labels (0) and 564 having patient labels (Figure 3). Following the completion of the balancing

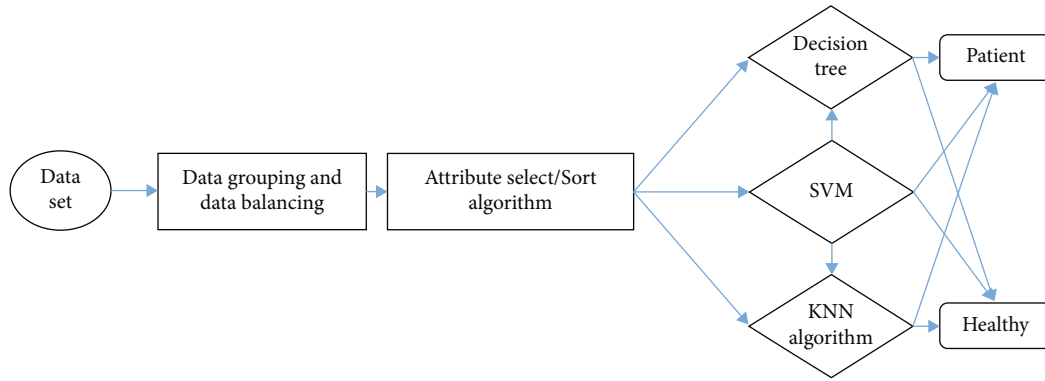


FIGURE 1: Flow chart.

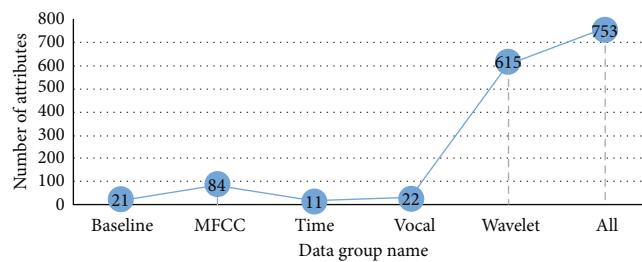


FIGURE 2: Data set attribute groups.

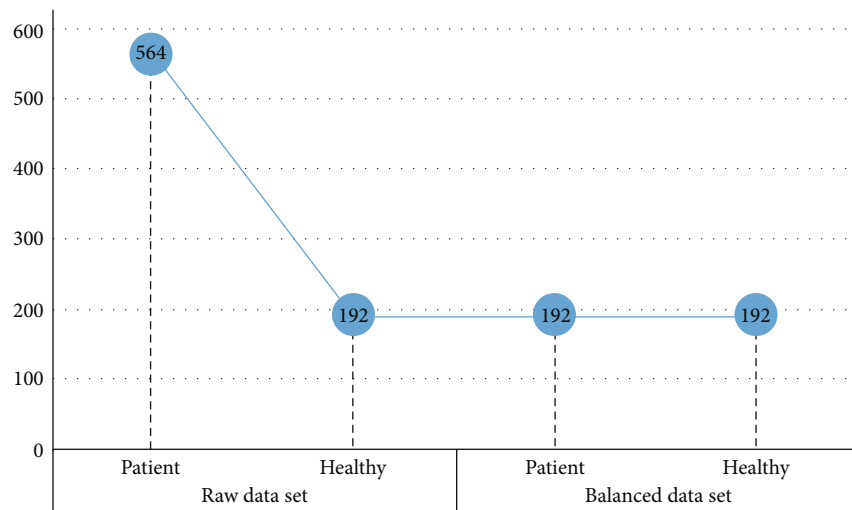


FIGURE 3: Schematic representation of data set balancing.

procedure, 192 healthy labels and 192 patient labels were obtained as a result of the process.

2.3. Feature Selection/Ranking Algorithm

2.3.1. *Fisher Attribute Sorting Algorithm.* The number of features affects machine learning’s positive and negative performance. The feature selection technique is carried out to reduce the negative effects on the environment. This approach generates a grade from relevant to irrelevant based on the power of any attribute in the tag estimate process.

The researcher can include as many features in the data set as he wants, ordered from the most relevant to the most irrelevant. Thus, it can get more accurate results without unnecessary data and have a faster program cycle. In this study, the feature selection algorithm was used and as shown in Figure 4, 5% and 10% of the data set, starting from the most relevant. The model was set up according to the performance values by taking 50% of it.

2.4. *Classification Algorithms.* Classification processes in our study were implemented with a decision tree (DT), support

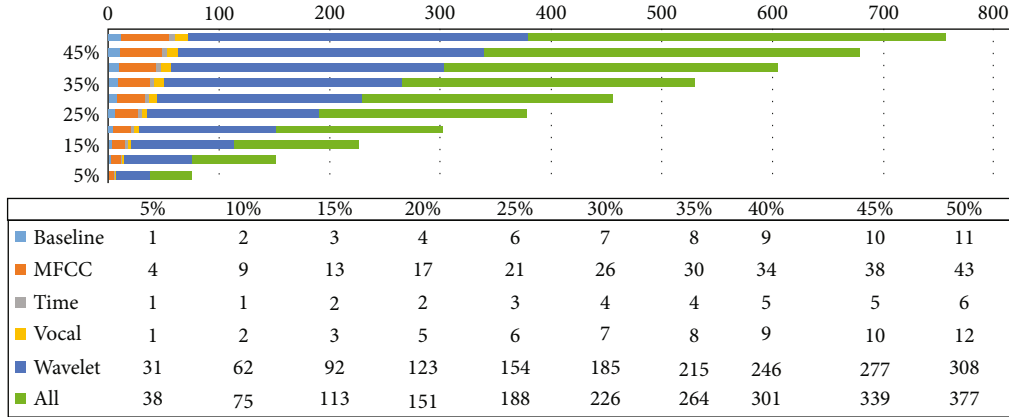


FIGURE 4: Attribute ranking in data groups.

vector machines (SVM), and k nearest neighbor (kNN) algorithms. The flow chart steps in Figure 5 were applied for the classification process. Half of the data was used in model creation and the other half to test the model to perform classification in the data set. For each data group, the training data set was created with the help of the systematic sampling method. The remaining data set was used in the testing phase. The performance evaluation criteria of the model built on the test data were tested.

2.4.1. Decision Tree (DT). The decision tree algorithm's fundamental structure is composed of several components, including roots, branches, nodes, and leaves, to mention a few. When the tree structure is constructed, each attribute is assigned to a particular node in the hierarchy. Between the root and the nodes, there are branches to consider. It is sent from one node to the other via branches from each node. The selection is made in the tree based on the most recently visited leaf [12]. The critical reasoning in forming a decision tree structure can be explained as follows: at each node reached, the relevant questions are asked, and the final leaf is reached in the shortest amount of time and space possible based on the responses provided. The responses to the questions serve as the foundation for creating models. It is determined whether or not this trained tree structure performs as expected by using test data, and the model is employed if it gives the desired result.

2.4.2. Support Vector Machines (SVM). Support vector machines (SVM) are used to separate data belonging to two classes in the most ideal way. This separation is performed with the appropriate linear and nonlinear lines. The learning data closest to the hyperplane are called support vectors. The maximum distance between the support vectors is determined, and a curve is fitted in between. This curve is accepted as the generalized solution for the classification process [12, 13]. The SVM method is one of the best and simplest algorithms among the supervised learning methods. The SVM algorithm develops a suitable classification method using the training data set. Then, it tries to classify the test data set with the minimum error with the method it has developed. SVM is used effectively in regres-

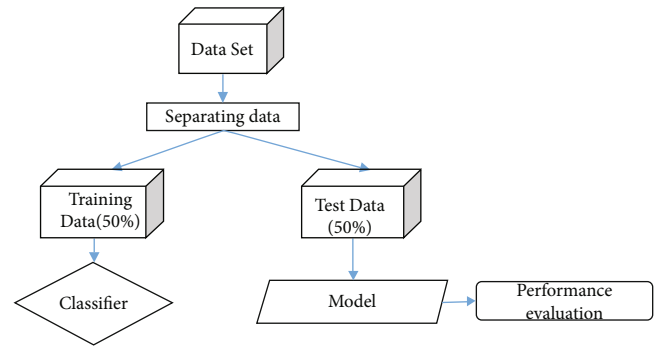


FIGURE 5: Separating the data set into training and test data sets.

TABLE 1: Training and test set distribution.

	Education (50%)	Testing (50%)	Total
Patient	86	86	192
Healthy	86	86	192
Total	192	192	384

sion analysis as well as classification problems. Most of the objects in the data sets used cannot be separated by linear vectors [9]. If objects cannot be separated with the help of a linear vector, a nonlinear support vector machine algorithm is used for classification. To classify a data set with objects, a size transformation is performed.

2.4.3. k NN (k Nearest Neighborhood). In k NN classification, classification is made according to the nearest neighbors. For nearest neighbors, the value of k may change to decide how to classify an unknown event; it determines how many values are considered neighbors. In the presence of an unknown sample, a nearest neighbor classifier explores the pattern space in search of the k training samples most similar to the unknown sample. The distance measures Euclid, Minkowski, and Manhattan are all employed to calculate the nearest neighbor [9]. In this case, the unknown instance is assigned the most prevalent class among its nearest

TABLE 2: Classifier achievement data for each data group.

Data set	Baseline data group			MFCC data group		
Performance criteria				Classifier algorithms		
	Decision tree (DT)	kNN	SVM	Decision tree (DT)	kNN	SVM
Accuracy rate (%)	71.35	68.75	66.67	69.79	75	73.44
Sensitivity	0.75	0.7	0.72	0.6	0.72	0.7
Specificity	0.68	0.68	0.61	0.79	0.79	00.78
<i>F</i> -measurement	0.72	0.69	0.66	0.69	0.75	0.73
Kappa	0.43	0.38	0.33	0.4	0.5	0.47
Area under the ROC curve	0.71	0.71	0.68	0.72	0.72	0.72
Data set	Time data group			Vocal data group		
Performance criteria				Classifier algorithms		
	Decision tree (DT)	kNN	SVM	Decision tree (DT)	kNN	SVM
Accuracy rate (%)	66.15	67.71	72.4	61.98	64.06	63.02
Sensitivity	0.73	0.82	0.8	0.64	0.58	0.64
Specificity	0.59	0.53	0.65	0.6	0.7	0.63
<i>F</i> -measurement	0.65	0.65	0.72	0.62	0.64	0.63
Kappa	0.32	0.35	0.45	0.24	0.28	0.26
Area under the ROC curve	0.65	0.69	0.70	0.61	0.63	0.64
Data set	Wavelet data group			Whole data group		
Performance criteria				Classifier algorithms		
	Decision tree (DT)	kNN	SVM	Decision tree (DT)	kNN	SVM
Accuracy rate (%)	68.23	73.44	79.69	69.79	76.56	85.42
Sensitivity	0.81	0.76	0.97	0.6	0.81	0.94
Specificity	0.55	0.72	0.63	0.79	0.72	00.78
<i>F</i> -measurement	0.66	0.73	0.76	0.69	0.76	0.86
Kappa	0.36	0.47	0.59	0.4	0.53	0.72
Area under the ROC curve	0.69	0.74	0.82	0.70	0.78	0.89

neighbors. When $k = 1$, the unknown sample is assigned the class of the training sample closest to it in the design area. The time to classify a test sample with the nearest neighbor classifier increases linearly with the number of training samples retained in the classifier. It has a large storage requirement. It also performs poorly when different properties affect the result for different scopes. The parameter that can affect the performance of the kNN classification algorithm is the number of nearest neighbors to be used. A nearest neighbor is used by default [10]. The k nearest neighbor algorithm is created by calculating the neighborhood distances for each object. k parameter expresses how many neighbors will be classified in the algorithm [5]. Each object in the data set is checked to which class its k neighbors belong. The object is included in the class its neighbors are the most. In order to avoid equality, the value of k is generally chosen from odd numbers [6]. In this study, the k value was chosen as 3.

2.5. Performance Evaluation Criteria. A variety of performance evaluation criteria were employed in this study to assess the overall performance of models developed using decision trees and support vector machines. The details are given in the above sections. While classifying the data set, the training-test ratios were 50-50% (Table 1).

3. Results and Discussion

The goal of our research was to use machine learning to diagnose Parkinson's disease. For this purpose, classifier algorithms are used. Classifier algorithms are applied to certain data groups, and appropriate performance values are obtained. Classifier achievements are shown in Table 2. The classifiers give the data groups the highest accuracy rate: 71% for baseline, 75% for MFCC, 72% for time, 64% for vocal, 80% for wavelet, and 85% for all. For each data group, there are classifier achievements with a certain accuracy. However, the highest accuracy is seen in all data groups, including each data group. The highest accuracy rate was seen in support vector machines (SVM) among the classification algorithms. The 85% accuracy (highest value) was obtained when 45% of the data in the group named all was taken and the SVM algorithm was used for classification. It can be said that the better results when 45% of the data set is trained, rather than 50%, may be due to the ranking of the most relevant to the most irrelevant features in the data set. The support vector machine for sensitivity, *F*-measure, kappa, and AUC provides the highest success when looking at other performance criteria. Only for the originality criterion does the decision tree algorithm achieve a slightly

higher success rate (0.02). Other performance criteria of the group with successful results are as follows; sensitivity, 0.94; specificity, 0.078; F -measure, 0.86; kappa, 0.72; and area under the ROC curve, 0.86.

4. Conclusion

In our study, it was aimed to benefit from machine learning in diagnosing PD. The data set used in machine learning consisted of only the analysis of the voice recordings of the patients. In this way, the diagnostic process will be shorter and less costly. It will also reduce the workload of clinical staff and enable patients to have an easier diagnosis process.

Many studies in the literature on PD diagnosis [13–16] created a data set by conducting the subjects to diagnose PD. The generated data set was grouped according to the age factor and analyzed using the Double Intensity 1-D Wavelet Transform method [7]. However, it has been studied on a small data set with few features.

In another study, Kumar et al. established a model with 87% accuracy using artificial neural networks. Two data sets were used in the established model. One of the data sets consists of 23 features, while the other consists of 26 features [6]. The models with a high accuracy rate will not provide the same accuracy in large data sets. The number of data in the data set we use is quite large. Therefore, the models created in this article can produce more reliable results. Many data sets currently available in the literature have an uneven distribution. In the studies, a model was created without eliminating this imbalance [14] found high accuracy rates in their different studies. A comprehensive data set was used in the studies, but balancing was not done. We think that the data used in this study will work more stable because the model is created by balancing the subsampling method. In models created with unstable data, the system produces results prone to data with the excess amount [7–9]. The models proposed in this study are one step ahead of the literature, built with balanced data sets.

In some studies in the literature, the results are given as the average of training and test performances [11]. In this regard, the results of these studies should be discussed. In this article, the test data were evaluated independently. Therefore, articles with high accuracy rates were relatively low. However, they are acceptable values and are more reliable than other studies [16].

This study obtained the best result when the first 45% of the features ordered according to the feature selection algorithm were taken. The accuracy rates decrease when working with more data groups, and the cycle speed slows down. Looking at the classification algorithms for each data group, the best performance values were seen in the support vector machine algorithm (Table 2). These results were obtained when the data set in the whole group was classified with the most relevant features of 45% (accuracy: 85%, sensitivity: 0.94, specificity: 0.078, F -measure: 0.86, kappa: 0.72, Auc: 0.86). With the created model, it is concluded that medical decision support can be provided to the doctor by facilitating the difficult and costly diagnostic process of diagnosing PD.

Data Availability

The data underlying the results presented in the study are available within the manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

References

- [1] G. Amali, K. Umadevi, and B. Varghese, "Prediction of Parkinson's disease using machine learning techniques on speech dataset," *Research Journal of Pharmacy and Technology*, vol. 12, pp. 644–648, 2019.
- [2] R. Blanco and A. Sánchez-Ferro, "The Diagnosis of Parkinson's Disease: Current Clinical Practice and Future Trends," in *Pathology, Prevention and Therapeutics of Neurodegenerative Disease*, pp. 103–116, Springer, Singapore, 2019.
- [3] E. Celik and S. Omurca, "Improving Parkinson's disease diagnosis with machine learning methods," in *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pp. 1–4, Istanbul, Turkey, 2019.
- [4] Z. K. Senturk, "Early diagnosis of Parkinson's disease using machine learning algorithms," *Medical Hypotheses*, vol. 138, p. 109603, 2020.
- [5] A. H. Hameed, E. A. Mousa, and A. A. Hamad, "Upper limit superior and lower limit inferior of soft sequences," *International Journal of Engineering and Technology (UAE)*, vol. 7, no. 4.7, pp. 306–310, 2018.
- [6] T. Kumar, P. Sharma, and N. Prakash, "Comparison of machine learning models for Parkinson's disease prediction," in *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pp. 0195–0199, New York, NY, USA, 2020.
- [7] D. A. Al-Mashhadany, L. Q. Ismael, and A. M. Zaki, "Seroprevalence of Helicobacter pylori among human in Erbil governorate, Kurdistan Region, IRAQ," *Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences*, vol. 4, p. 268, 2019.
- [8] R. Mathur, V. Pathak, and D. Bandil, "Parkinson disease prediction using machine learning algorithm," in *Parkinson disease prediction using machine learning algorithm: Proceedings of ICETEAS 2018*, pp. 357–363, Warsaw, Poland, 2019.
- [9] A. A. Hamad, M. L. Thivagar, M. B. Alazzam, F. Alassery, F. Hajje, and A. A. Shihab, "Applying dynamic systems to social media by using controlling stability," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 4569879, 7 pages, 2022.
- [10] I. Nissar, W. A. Mir, and T. A. Shaikh, "Machine learning approaches for detection and diagnosis of Parkinson's disease - a review," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 898–905, Coimbatore, India, 2021.
- [11] A. Qasim and A. Al-Taweel, "Caries experience among Parkinson disease patients in Baghdad - Iraq," *International Journal of Science and Research (IJSR)*, vol. 6, 2017.
- [12] N. Salari, M. Kazeminiya, H. Sagha, A. Daneshkhah, A. Ahmadi, and M. Mohammadi, "The performance of various machine learning methods for Parkinson's disease

- recognition: a systematic review,” *Current Psychology*, vol. 140, pp. 1–24, 2022.
- [13] M. Singh, “A comparative study of different machine learning algorithms in detection of Parkinsons disease,” *International Journal for Research in Applied Science and Engineering Technology*, vol. 9, no. 1, pp. 755–766, 2021.
- [14] T. V. Sriram, M. V. Rao, G. S. Narayana, D. S. Kaladhar, and T. P. Vital, “Intelligent Parkinson disease prediction using machine learning algorithms,” *IJEIT.*, vol. 3, pp. 212–215, 2013.
- [15] M. Thanoun and M. Yaseen, “A comparative study of Parkinson disease diagnosis in machine learning,” in *2020 The 4th International Conference on Advances in Artificial Intelligence*, pp. 23–28, London, United Kingdom, 2020.
- [16] H. Watanabe, K. Hara, M. Ito, M. Katsuno, and G. Sobue, Eds., “New diagnostic criteria for Parkinson’s disease: MDS-PD criteria,” *Brain and nerve = Shinkei kenkyū no shinpo*, vol. 70, no. 2, pp. 139–146, 2018.