MLA

by Mrinal Bhan

Submission date: 08-May-2023 12:32AM (UTC+0530)
Submission ID: 2086656757
File name: MLA_Final_1.docx (560.59K)
Word count: 2929
Character count: 16593

fPI: A Novel Index for Predictive Analysis of Parkinson's Disease Using Acoustic Sound Feature

[1]Gautam Gupta
Data Science & AI
International Institute of Information Technology
Naya Raipur, India
gautam21102@iiitnr.edu.in

[2]Mrinal Bhan
Data Science & AI
International Institute of Information Technology
Naya Raipur, India
mrinal21102@iiitnr.edu.in

[3]Sahil Nimsarkar
Data Science & AI
International Institute of Information Technology
Naya Raipur, India
sahiln21102@iiitnr.edu.in

***Abstract* — Parkinson's disease is a severe neurological disorder that can lead to major disability, reducing the quality of life, and with no known cure. It is caused by the lack of dopamine, a chemical responsible for transmitting messages between the brain and nervous system to control body movements. Most people with Parkinson's Disease have speech disorders, and it typically occurs in individuals over the age of 60. However, some people may experience early onset before the age of 50. Currently, there is no cure for PD, but medications and surgery may provide some relief for motor symptoms. This project used several machine learning classification algorithms to predict if someone had PD, comparing SVM , Decision Trees, Random Forest, K-nearest neighbors, and ensemble techniques. The Extreme Gradient Boosting (XGBoost) classification algorithm achieved a 96.5% test accuracy rate, which was the highest compared to other algorithms. The performance of these models was assessed with a reliable dataset from the UCI Machine Learning Repository.**

Keywords— Parkinson's, Classification, Support Vector Machine, Random Forest, Decision Trees, XGBoost

I. INTRODUCTION

Parkinson's disease is a neurological ailment that impairs the body's motor abilities over time and gets worse. The body may tremble subtly in some places as one of its first symptoms. The disease leads to tightness and difficulty with movements. It is a debilitating condition that can't be diagnosed with a blood test.As existing diagnosis systems for Parkinson's are expensive and unreliable, there is a need for a faster and more cost-effective tool. This research project uses various machine learning algorithms to detect the existence of PD through the assessment of acoustic sound patterns. The dataset used to evaluate accuracy contains features derived from voice samples of both affected and unaffected individuals. Models such as SVM, Decision Trees, Random Forest, and XGBoost are used to diagnose PD as accurately as possible. A dataset taken from UCI consisting of several features from different voice samples of Parkinson's patients and unaffected subjects is used in model training and deployment. This data with its different frequencies can be given to the model to identify those with Parkinson's and the results can be displayed in an application that can be used by the patient to access the outcome.

Our motivation for choosing this project is based on the need for accurate detection of Parkinson's disease (PD) using speech processing algorithms. There have been a number of research focused on identifying Parkinson's disease using various classifiers and speech tasks. For instance, Little et al. [1] conducted an evaluation of methods for differentiating PD cases from healthy individuals by identifying dysphonia, touching a classification accuracy of 91% using SVM classifier. With sustained vowels, words, and sentences from PD patients and controls, Sakar et al.'s [2] voice experiments were designed, and they used an SVM classifier to obtain 77.5% accuracy. In a comparison of artificial neural networks (ANN), regression, and ANN classifiers, Das [3] found that the ANN classifier produced the most optimal outcomes, with a classification score of 92.9%. Other studies used Multilayer Feedforward Neural Networks (MLFNN) [4], algorithm for k-nearest neighbours in rotation-forest ensemble [5], Sparse multinomial logistic regression and linear logistic regression [6], and deep neural networks [7]. These studies demonstrate the potential of methods for speech signal processing for the accurate detection of PD, and we aim to contribute to this field by exploring the effectiveness of different classifiers and speech tasks for PD diagnosis. The use-case of machine learning algorithms in the medical industry is given below in Fig 1;

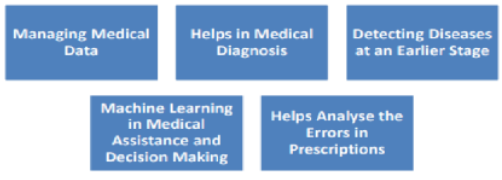

Fig I: Use-case of ML

II. DATASET

The University of California, Irvine (UCI) created the data set used for analysis, which contains speech signals. It includes 195 samples of features from 31 individuals, with 23 of them having Parkinson's disease and 8 being from control group. The primary purpose of the data set is to differentiate between healthy individuals and those with PD, based on a "status" column which is mapped as 1 for Parkinson's affected and 0 for healthy. Each patient was recorded around 6 times, and the data set contains approximately 75% of cases with PD and 25% of cases that are healthy.

Family	Attribute	Description
Vocal quality measures	MDVP: Fo (Hz)	Average vocal fundamental frequency
	MDVP: Fhi (Hz)	Maximum vocal fundamental frequency
	MDVP: Flo (Hz)	Minimum vocal fundamental frequency
Pitch Local perturbation measures	MDVP: Jitter(%) MDVP: Jitter(Abs) MDVP: RAP MDVP: PPQ Jitter: DDP	Several measures of variation in fundamental frequency
Amplitude local perturbation measure	MDVP: Shimmer MDVP: Shimmer(dB) Shimmer:APQ3 Shimmer:APQ5 MDVP: APQ Shimmer: DDA	Several measures of variation in amplitude.
Noise features	NHR, HNR	Two measures of the ratio of noise to tonal components in the voice
Non-linear measure	PDE, D2	Two nonlinear dynamical complexity measures
	DFA	Signal fractal scaling exponent
	Spread1, Spread2, PPE	Three nonlinear measures of fundamental frequency variation
Predictor	Status	Health status of the subject

Fig II: UCI Parkinson's disease dataset

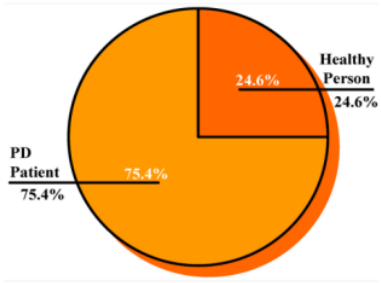

Fig III: Health Status of PD Patient

III. KEY CONTRIBUTION

The key contributions of the proposed work include:

1) We formulate a novel index called Frequency Parkinson's Indicator *(fPI)* derived from the features of the dataset, which is used to train different models.

$$fPI = \log_{10}(D2 * DFA) \times spread2$$

The index we proposed, is a combination of three different features: D2, DFA, and spread2. The selection of these parameters was based on previous research findings and their potential to distinguish PD patients from healthy individuals. This index was derived through a process of feature engineering, which involves selecting, transforming, and combining different features to create a new feature that may better capture the underlying relationship between the input data and the output variable. Our finding has important clinical ramifications because Parkinson's disease is a degenerative neurological condition that affects millions of people worldwide. PD must be precisely and speedily diagnosed in order to be effectively treated and controlled. However, current diagnostic methods are limited in their accuracy, and there is a need for more reliable and non-invasive diagnostic tools. Our proposed PDI offers a novel approach to PD classification using frequency parameters that are easily accessible and non-invasive. The high accuracy achieved by the PDI suggests that it has significant clinical potential as a diagnostic tool for Parkinson's disease. fPI has potential applications beyond the classification of Parkinson's disease. It is based on the analysis of nonlinear dynamics in speech signals.

2) Based on speech metrics, we categorized various machine learning models and examined their efficacy in diagnosing Parkinson's illness.

3) This research also discusses various machine learning (ML)-based frameworks for the diagnosis of Parkinson's disease with the aim of improving Parkinson's disease data.

4) The article concludes by highlighting the difficulties and outlining suggestions for further research.

IV. METHODOLOGY

The proposed methodology for the detection and prediction of Parkinson's Disease is shown in Figure IV:

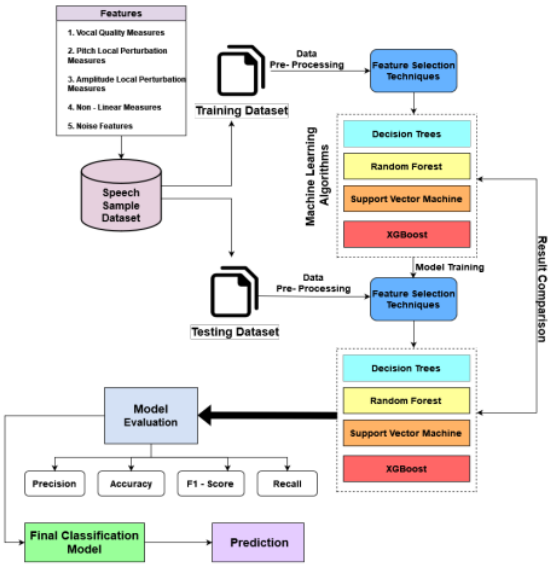

Fig IV: Flow Diagram of Procedure

A. Data Pre – Processing:

The features produced from the dataset were highly dimensional and unstandardized. Hence, dimensionality reduction techniques such as Principal Component Analysis (PCA) and Standard Scaler had been used as pre-processing tools for the data. After data pre-processing, exploratory data analysis was also done to get an overview of the used features and their correlations.

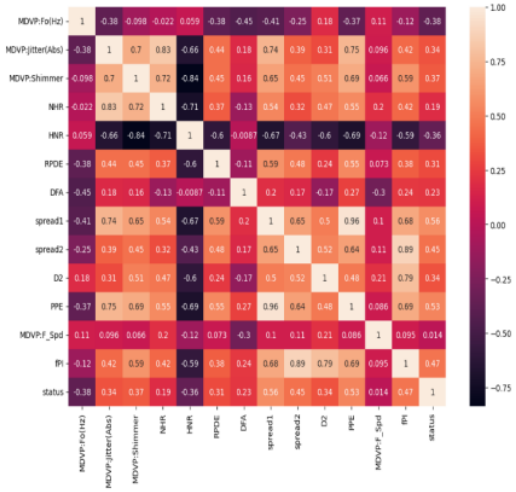

Fig VI: Correlation Heatmap of features

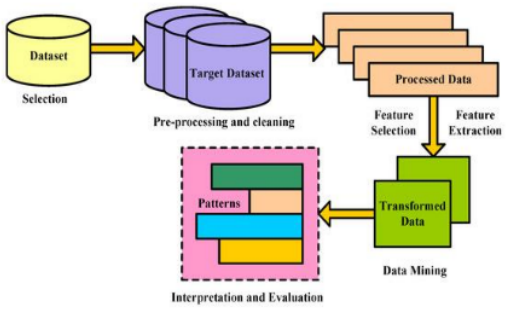

Fig V: Feature Selection and Pre-processing on dataset

B. Formulation of novel index (fPI):

A novel index was created in order to create a new feature for training machine learning models and validating their performances. The index was given by the equation:

$$fPI = \log_{10}(D2 * DFA) \times spread2$$

D2 and DFA are measures of the nonlinear dynamics of the signal, which capture the complexity of the system underlying the speech production process. Specifically, D2 measures the degree of divergence of nearby trajectories in phase space, while DFA characterizes the scaling properties of the signal. These features are widely used in the analysis of nonlinear systems and have been shown to be useful in the classification of Parkinson's disease. Spread2, on the other hand, is a measure of the spread of the signal in a nonlinear feature space. This feature is related to the fundamental frequency variation in the speech signal and is also a useful discriminator of Parkinson's disease. The log transformation of the product of D2 and DFA in the proposed index is a way of capturing the nonlinear interactions between these two features, which may be important for understanding the complexity of the system underlying the speech production process. The multiplication of this product by spread2 further emphasizes the importance of fundamental frequency variation in Parkinson's disease and may help to differentiate between healthy individuals and those with Parkinson's disease.

C. Machine Learning Algorithms

1) Decision Trees: The Decision Tree is a classifier that has a tree-like structure consisting of two types of nodes - Decision nodes and Leaf nodes. It begins with a starting point that branches into two or more outcomes. The Leaf nodes provide the classification or the attribute value of the input example. The DT Classifier is suitable for both classification and regression, although it is more commonly used for

classification. The tree starts from the root i.e. the top most node and progresses to the relevant branch based on the test value until it reaches the leaf node. The Decision Tree is an important machine learning technique as it simplifies a complex problem into a more manageable one.

2) Random Forest: One of the most commonlyused supervised ML classifier called Random Forest can be applied to both classification and regression tasks. It makes use of ensemble learning, which combines different classifiers to increase accuracy and resolve complex issues. Numerous decision trees operating on various dataset subsets make up the algorithm. The algorithm estimates the mean value of the predictions from each tree in order to increase the model's accuracy. Random Forest generates predictions from each tree and combines them to create a final output depending on the majority vote, in contrast to single decision trees.

3) Support Vector Machine: This classifier is a versatile method that can be used to address both classification and regression problem statements. In a classification problem, SVM aims to separate or classify two distinct classes with the aid of a hyperplane. To achieve this, SVM creates two marginal lines and a hyperplane that is some distance away from the lines. This makes it easier to linearly separate the classification points. Furthermore, SVM generates two parallel planes that pass through the nearest point of each of the two classes, known as the support vectors. The distance between the planes is referred to as the marginal distance, which acts as a cushion to more effectively divide a point into the appropriate classes. SVM selects the best hyperplane with the maximum marginal distance.

4) eXtreme Gradient Boost: The effective ensemble learning technique XGBoost is applied to classification and regression issues. Its name, "Extreme Gradient Boosting," refers to the way the gradient boosting decision tree algorithm is implemented. To repair the mistakes caused by the preceding tree, the technique employs numerous decision trees that are successively trained. It also uses regularization techniques to avoid overfitting and improve generalization. Compared to other algorithms, XGBoost provides better performance due to its ability to handle high-dimensional data and its capacity to work with missing values. In classification tasks, XGBoost is particularly useful as it can handle imbalanced datasets by assigning higher weights to underrepresented classes.

D. Model Evaluation:

Several metrics that have been used to assess the effectiveness of machine learning models are listed in Fig. V. AUC score, F1 score, and recall are some of the measures most frequently preferred to assess the effectiveness of machine learning models. Precision measures the fraction of genuine positives over the total number of anticipated positives. On the otherhand, accuracy measures the percentage of correctly predicted occurrences out of all cases. Recall is measuring the ratio of true positives to all other positives. The F1 score, which combines precision and recall, offers a fair assessment of a model's performance. AUC score gives us a means to assess a model's capacity to distinguish between healthy and non-heatlhy patients by measuring the area under the Receiver Operating Characteristic (ROC) curve. These measurements are frequently combined to give a thorough insight of a model's performance.

PERFORMANCE METRIC	DEFINITION
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
F1 – Score	$\frac{2 * (Precision * Recall)}{(Precision + Recall)}$
Precision	$\frac{TP}{TP + FP}$
Sensitivity (Recall)	$\frac{TP}{TP + FN}$
Specificity (TNR)	$\frac{TN}{TN + FP}$
AUC	The two – dimensional area under the ROC curve

Fig VII: Performance metrics used for evaluation of ML models.

V. RESULTS

In this study, we explored the performance of various ML algorithms, in predicting Parkinson's disease based on a dataset of clinical and demographic features of patients. The results obtained from the experiments demonstrate the effectiveness of these ML techniques in accurately predicting the presence of Parkinson's disease.

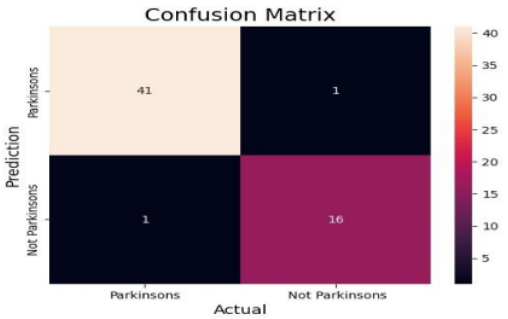

Fig VIII: Confusion Matrix for Prediction

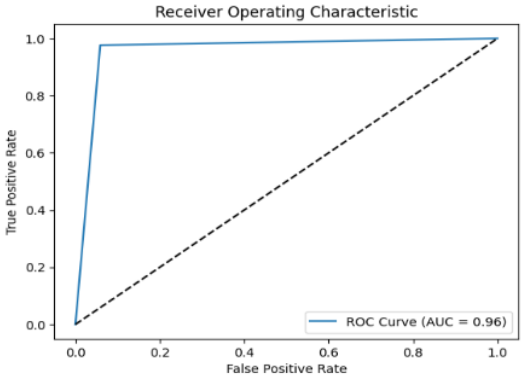

Fig IX: ROC Curve (TPR vs FPR)

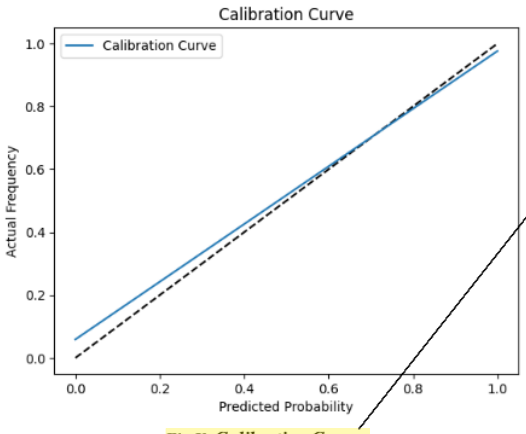

Fig X: Calibration Curve

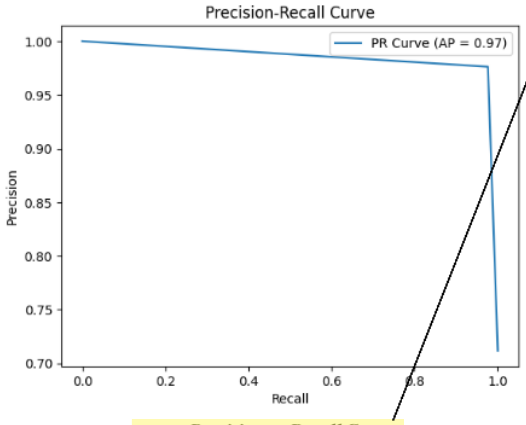

Fig XI: Precision vs Recall Curve

It can be Inferred from the given figures such that XGBoost outperforms all the Other Models implemented with Accuracy Score around 96.6 %, Precision Score around 97.6%, F1 Score around 97.6%. Hence Considering all the Performance Metrics, XGBoost t is the most obvious choice considering the fact that it has outperformed and can be trusted on such a problem which doesn't allow even a small gap for error considering it is a part of a field like Health Care.

Model	*Accuracy*	*Precision*	*Recall*	*F1 Score*	*AUC Score*
SVM	0.83	0.9	0.85	0.87	0.81
Decision Tree	0.898	0.95	0.904	0.926	0.893
Random Forest	0.932	0.975	0.928	0.951	0.934
XGBoost	0.966	0.976	0.976	0.976	0.958

Fig XII: Obtained Results

VI. CONCLUSION

One of the most serious neurodegenerative illnesses with no known cure is Parkinson's disease, thus prevention is crucial. In this experiment, we used a variety of prediction models, including SVM, Decision Trees, Random Forest, and XG Boost, to predict Parkinson's disease. These models are used to train the dataset, and we also compared the many models created using the various methods to choose the one that matches the data the best. Utilizing evaluation criteria that effectively predict the disease, such as Accuracy, F1-score, etc. is the goal. We made advantage of the Kaggle-available Speech dataset of Parkinson's patients, which includes the patients' speech attributes. More than 20 features and 30 patient information make up the dataset. The top five features, as determined via feature selection, are used to construct the models. With an accuracy of 96.6% based on these results, XG Boost excels above the other machine learning methods. We developed a technique that is extremely accurate at predicting Parkinson's disease.

VII. FUTURE SCOPE

In the future, incorporating multi-modal datasets could significantly improve the accuracy and reliability of machine-learning models for Parkinson's disease detection. This would involve incorporating important factors such as handwritten text and gait analysis into the datasets used for training the models. Additionally, there is also room for the betterment the accuracy of the used machine learning models for PD detection. Exploring different algorithms like Artificial Neural Networks (ANN), Naïve Bayes Classifier, etc., and optimizing the hyperparameters could be a potential avenue for improving the accuracy of the model. By exploring these avenues, we can expect significant progress toward better and more reliable methods for Parkinson's disease detection, leading to better patient outcomes and quality of life. Another

promising direction for future research could be the development of a real-time Parkinson's monitoring system- a type of system that would continuously monitor the symptoms of Parkinson's disease in patients and provide real-time alerts when there is a deterioration in their condition. Such a system could be invaluable for patients, allowing them to take proactive measures and seek timely medical intervention when necessary. Thus, by exploring this direction, we could make significant strides toward more effective Parkinson's disease management and treatment.

VIII. Acknowledgment

We extend our sincere gratitude to Dr. Anurag Singh and Mr. Shresth Gupta, our guides, and mentors for their unwavering support, valuable guidance, and the opportunity to delve into such an interesting subject. Their unwavering dedication to creative thinking and hard work has been a source of inspiration for us. Without their invaluable guidance, we would not have been able to create such a project. Furthermore, we would like to extend our heartfelt appreciation to IIIT NR for providing us with all the necessary facilities to conduct this research. Their support has been instrumental in the successful implementation of this project. Finally, we would like to express our appreciation to all those who contributed to this research work, directly or indirectly. Their support has been invaluable in shaping our knowledge and understanding.

IX. References

[1] *Bhattacharya, Ipsita & Bhatia, Mohinder: Pal Singh. (2010). SVM classification to distinguish Parkinson's disease patients. 14. 10.1145/1858378.1858392.*

[2] *Little, S., Pogosyan, A., Neal, S., Zavala, B., Zrinzo, L., Hariz, M., Foltynie, T., Limousin, P., Ashkan, K., FitzGerald, J., Green, A. L., Aziz, T. Z., & Brown, P. (2013). Adaptive deep brain stimulation in advanced Parkinson's disease. Annals of neurology, 74(3), 449–457. https://doi.org/10.1002/ana.23951*

[3] *Das, Resul. (2010). Das, R.: A comparison of multiple classification methods for diagnosis of Parkinson's disease. Expert Systems with Applications 37, 1568-1572. Expert Systems with Applications. 37. 1568-1572. 10.1016/j.eswa.2009.06.040.*

[4] *Olanrewaju, Rashidah & Sahari, Nur & Aibinu, Abiodun & Hakiem, Nashrul. (2014). Application of neural networks in early detection and diagnosis of Parkinson's disease. 2014 International Conference on Cyber and IT Service Management, CITSM 2014. 10.1109/CITSM.2014.7042180.*

[5] *Gök, Murat. (2015). An ensemble of k-nearest neighbours algorithm for detection of Parkinson's disease. International Journal of Systems Science. 46. 10.1080/00207721.2013.809613.*

[6] *Mandal, I., & Sairam, N. (2013). Accurate telemonitoring of Parkinson's disease diagnosis using a robust inference system. International journal of medical informatics, 82(5), 359–377. https://doi.org/10.1016/j.ijmedinf.2012.10.006.*

[7] *Shahid, A. H., & Singh, M. P. (2020). A deep learning approach for prediction of Parkinson's disease progression. Biomedical engineering letters, 10(2), 227–239. https://doi.org/10.1007/s13534-020-00156-7 M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989*

MLA

ORIGINALITY REPORT

8%	6%	5%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchgate.net Internet Source	1%
2	Vivek Tiwari, R. S. Thakur. "chapter 118 Knowledge-Based Forensic Patterns and Engineering System", IGI Global, 2018 Publication	1%
3	journals.lww.com Internet Source	1%
4	Das, R.. "A comparison of multiple classification methods for diagnosis of Parkinson disease", Expert Systems With Applications, 201003 Publication	<1%
5	Submitted to PES University Student Paper	<1%
6	Submitted to SASTRA University Student Paper	<1%
7	Amin ul Haq, Jian Ping Li, Bless Lord Y. Agbley, Cobbinah Bernard Mawuli, Zafar Ali, Shah Nazir, Salah Ud Din. "A survey of deep	<1%

learning techniques based Parkinson's disease recognition methods employing clinical data", Expert Systems with Applications, 2022
Publication

8	downloads.hindawi.com Internet Source	<1 %
9	Lipsita Sahu, Rohit Sharma, Ipsita Sahu, Manoja Das, Bandita Sahu, Raghvendra Kumar. "Efficient detection of Parkinson's disease using deep learning techniques over medical data", Expert Systems, 2021 Publication	<1 %
10	digitalcommons.usf.edu Internet Source	<1 %
11	ebin.pub Internet Source	<1 %
12	journals.pan.pl Internet Source	<1 %
13	ntc-container.com Internet Source	<1 %
14	ouci.dntb.gov.ua Internet Source	<1 %
15	Tapan Kumar, Pradyumn Sharma, Nupur Prakash. "Comparison of Machine learning models for Parkinson's Disease prediction", 2020 11th IEEE Annual Ubiquitous Computing,	<1 %

	Electronics & Mobile Communication Conference (UEMCON), 2020 Publication	
16	cs.umd.edu Internet Source	<1%
17	sonnati.wordpress.com Internet Source	<1%
18	www.nature.com Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

MLA

PAGE 1

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Possessive

Article Error You may need to use an article before this word.

P/V You have used the passive voice in this sentence. You may want to revise it using the active voice.

Article Error You may need to use an article before this word.

S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

Wrong Article You may have used the wrong article or pronoun. Proofread the sentence to make sure that the article or pronoun agrees with the word it describes.

Article Error You may need to remove this article.

Possessive

Article Error You may need to use an article before this word.

Article Error You may need to use an article before this word.

Possessive

Article Error You may need to use an article before this word.

PAGE 2

Article Error You may need to use an article before this word.

Possessive

Article Error You may need to use an article before this word.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

PAGE 3

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Article Error You may need to use an article before this word.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Article Error You may need to remove this article.

P/V You have used the passive voice in this sentence. You may want to revise it using the active voice.

Frag. This sentence may be a fragment or may have incorrect punctuation. Proofread the sentence to be sure that it has correct punctuation and that it has an independent clause with a complete subject and predicate.

Article Error You may need to remove this article.

Article Error You may need to use an article before this word.

Article Error You may need to remove this article.

Frag. This sentence may be a fragment or may have incorrect punctuation. Proofread the sentence to be sure that it has correct punctuation and that it has an independent clause with a complete subject and predicate.

Sentence Cap. Review the rules for capitalization.

PAGE 4

Missing "," Review the rules for using punctuation marks.

Compound These two words should be a compound word.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Wrong Article You may have used the wrong article or pronoun. Proofread the sentence to make sure that the article or pronoun agrees with the word it describes.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Article Error You may need to use an article before this word.

Article Error You may need to remove this article.

PAGE 5

Prep. You may be using the wrong preposition.

Run-on This sentence may be a run-on sentence.

Article Error You may need to remove this article.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Article Error You may need to remove this article.

Wrong Article You may have used the wrong article or pronoun. Proofread the sentence to make sure that the article or pronoun agrees with the word it describes.

PAGE 6

Frag. This sentence may be a fragment or may have incorrect punctuation. Proofread the sentence to be sure that it has correct punctuation and that it has an independent clause with a complete subject and predicate.

Article Error You may need to use an article before this word. Consider using the article **the**.

Frag. This sentence may be a fragment or may have incorrect punctuation. Proofread the sentence to be sure that it has correct punctuation and that it has an independent clause with a complete subject and predicate.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Article Error You may need to use an article before this word.

Article Error You may need to use an article before this word. Consider using the article **the**.

Frag. This sentence may be a fragment or may have incorrect punctuation. Proofread the sentence to be sure that it has correct punctuation and that it has an independent clause with a complete subject and predicate.

Article Error You may need to use an article before this word. Consider using the article **the**.

Proofread This part of the sentence contains an error or misspelling that makes your meaning unclear.

Article Error You may need to use an article before this word. Consider using the article **the**.

Sp. This word is misspelled. Use a dictionary or spellchecker when you proofread your work.

Article Error You may need to use an article before this word.

Article Error You may need to use an article before this word.