iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** V **Month of publication:** May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5379

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Dimensionality Reduction Method for Prediction of Parkinson Disease using Speech Data

Jayashri P[1], Rajalakshmi V[2]
[1, 2]*Department of Computer Science & Engineering, Sri Venkateshwara College of Engineering*

Abstract: ***Neurological diseases are risky if the symptoms are not detected. It is important to detect the symptoms accurately at an early stage. Parkinson's disease (PD) is a significant nervous system disorder affecting ten million people worldwide. Parkinson disease's can be identified by three characteristics: speech, memory and movement disorder. Speech and memory data is used to predict the presence of the disease. Movement disorders are a less reliable resource for detection. Collection and processing of speech data are less intensive. Since memory disorder symptoms are detected at a later stage, the challenge increases. Generally, classification and regression techniques are implemented for the identification of PD patients. The proposed work focuses on using Principal component analysis on a combination of speech and memory information. The technique is used for reducing the dimension of the features. Traditional classification algorithms have been applied and compared. The accuracy of the dimensionality reduction techniques is improved to 97% by using the Random Forest approach.***
Keywords: Parkinson disease, Random forest, speech, Principal component analysis, Dimensional reduction

I. INTRODUCTION

The human brain has brain cells that contain dopamine. The Substantia nigra is one of the neurons in the specific area of the brain. Dopamine is the major one where the brain contacts with other parts of the body and provides smooth movement in contraction and expansion of the parts.

The degeneration in the dopamine from the human brain leads to slow in the movement of the body parts.

Parkinson's disease is a neurodegenerative disorder that reduces dopamine production in the brain and leads to bradykinesia. The major symptoms are tremor is the involuntary shaking due to muscle relaxation and contraction, rigidity is the stiffening of body parts followed by bradykinesia.

The other symptoms include gait, difficulty in speech, emotional changes, sleep disruption. The diagnosing of the disease at an early stage is difficult because the biomarker is not visible at the stages. The different stages in Parkinson's disease are diagnosed using the UPDRS (Unified Parkinson's Rating Scale).

The stages are early Parkinson disease, healthy patients, REM (rapid eye movement) facing difficulty Parkinson disease. The genetics and environmental factors are the major source for the cause of PD. The reason for the disease is not properly found by the researcher.

Age is also an important constraint in the cause of the disease. The age of the patient above 50 has a major risk of getting the disease. Men have a higher PD than women. The person works in the pesticide and the herbicide area reports the chance of the disease. The original cause of the disease is not found out.Worldwide, Parkinson disease affects ten million people. Men are more affected in PD than women.

The disease occurs age before 50 is termed as early offset. In India, one lakh people are affected by Parkinson disease. The disease rate is increasing year by year. In 2030, the rate of the PD increases by 1.5 million expected by the researcher. In recent, death has increased without the proper recognition of the disease in the patient.

The speech is affected by basal ganglia failing to reproduce movement leads to suppression of vocal cords and facing difficulty in pronunciation. The symptoms related to speech are softening of voice and the person takes long breaths, the words may run into one another, the words are expressed rapidly.

The symptoms related to memory are finding the word to speak is difficult and difficult to participate in conversations. The other symptoms are hypokinetic dysarthria, voice quality becomes low, hypokinetic articulation term as loss of movement, hypophonia means the absence of coordination in vocal muscle, mono-pitch is the clarity of voice in the same pitch, mono loudness, and deficits in timing.

Speech and voice abnormalities were the major biomarker in Parkinson's disease and are noticed in 90% of people. The reading and monologue data is used to detect the illness in Parkinson disease. The traditional machine learning algorithm is applied to the combined data. Then, the features are selected from the combined data using PCA. The features in the dataset are reduced by Principal component analysis by using dimensionality reduction. Using the PCA for feature projection and applied random forest as a classifier to predict the accuracy.

II. RELATED WORK

In 2019, Amr Gaballah proposed a technique to identify Parkinson disease using amplification devices that distinguish healthy patients with the disease patient. The dataset consists of 11 participants and the age range between 58 to 80 years contains male and female participants. The regression method such as support vector regression, GVR, Deep neural network applied and found the correlation with ratings.

In the same year, John Prince presented an ensemble learning method to handle the patients with the missing data. The dataset is gathered from the mpower application found in the iPhone. The features like tapping, memory, speech and walking capture remotely through the iPhone. The classification is based on the complete and incomplete dataset. The different classification algorithm is applied and found the classification accuracy rate.

In 2016, Achraf Benba proposed distinguishing between the neurological disease and Parkinson disease patients. The voice sample is gathered from the 50 subjects using microphone and voice recording is processed in various conditions. The five different supervised classification algorithms are performed on the data. Using the linear SVM kernels, the performance metrics are calculated and compared with the existing techniques.

Ferdous Wahid proposed the techniques that classify the Parkinson disease using the Gait features. The database consisted of 26 age controls suspected of Parkinson disease and 23 Parkinson's subjects. The classification accuracy of 92.6% obtained in random forest approach then the gait feature normalizes the accuracy rate is reduced to 80.4% using support vector machine and 86.2% using Kernel Fisher Discriminant.

Bryan T.cole presents a technique that uses the wearable sensors to track the Tremor and Dyskinesia symptoms from the participants. The database consists of 12 subjects of which 8 are Parkinson patients and 4 are healthy. The dynamic machine learning algorithm is applied for classification of which deep neural network provides better classification accuracy.

Martinez Manzanera uses orientation sensors to compute the score for the Bradykinesia using UPDRS. The database comprises both the male and female and age limit above 64. Using the feature selection method, the seven features were used to reduce the error rate compared to the normal classification algorithms.

III.IMPLEMENTATION METHODOLOGY

The method consists of three steps including:

A. Dataset Description

The dataset was obtained from the UCI machine learning data repository and it was created by HlavniÄka et al. The data consists of 130 patients that come under 3 categories. The healthy patients, early-stage and patient-facing difficulty are the categories. The dataset consists of 130 samples of which 30 patients were early Parkinson disease symptoms, 50 patients were developing high risk in Parkinson disease and 50 were healthy controls.

The twelve attributes are Entropy of speech, rate of speech, acceleration of speech, duration of pause interval, duration of the voiced interval, decay of voiced fricatives, relative loudness of respiration, pause interval of respiration, rate of speech respiration, the latency of respiration exchange, the gap in between voiced interval, duration of the unvoiced stop. The patient performs reading the passage and expresses about the job, family and current activities up to 80 words. The patients are examined by the speech specialist and undergo training by the specialist.

In the reading passage, the patient read the passage for a period of 90 sec. The various metrics are measured in the recorded data. In this method no memory is involved. The person reads the content specified by the specialist.

In the monologue, the patient tells about the job, family and current activities up to 80 words. In this method the memory is involved. The time is taken for the thought and expression of the word. The monologue is processed differently from the reading passage data. The features and description are shown in table I.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Table I
Features And Description

FEATURES	DESCRIPTION
Entropy of speech Timing	Measure peakedness of disturbance
Rate of speech	No of words in a minute
Acceleration of speech	Time taken become fast or slow
Duration of pause interval	Time to pause for a certain time
Duration of voiced interval	Time taken for voice to deliver
Decay of voiced fricatives	Delay in pronounce F, S, Z, V
Relative loudness of respiration	Respiration between the speech
Pause interval of respiration	Respiration during the pause timing
Rate of speech respiration	Respiration rate during speech
Latency of respiration exchange	Time taken for expiration and inspiration
Gap in between voiced interval	Gap in-between speech
Duration of unvoiced stop	Time taken to start the next speech

B. Disease Classification

The Machine learning algorithm uses different methods to classify the Parkinson patients. The supervised learning method is used to train a module and perform classification. Then the test data evaluates of a final module fit on the training data. The four classification algorithms are used in the project. Random forest, Support vector machine, Naïve Bayes and K nearest neighbor are the classification algorithms used.

1) *Support vector machines:* SVM are one of the supervised classification techniques that separate the classes by using the hyperplane. The distance between the hyperplane is larger than the error rate is low then the accuracy is improved. The mapping of a higher dimension to the lower dimension is done by using a kernel function.
2) *Random forest:* RF is the supervised classification techniques used in different areas. Random forest is derived from the decision tree and it consists of the group of the decision tree. Classification and regression trees are used in random Forest. In a random forest, the multiple decision trees are combined for the correct classification of the dependent and independent variables. The vote is provided by each tree based on the classification features by the independent variable. The majority of voters decided on the final results.
3) *K nearest neighbor algorithm:* KNN is a method that performs both classification and regression. The model assigns the values for the features and calculates the distance between the features then the values are arranged in ascending order. Based on the vote the values are classified into different classes.
4) *Naive Bayes:* NB is the classification algorithm of independent variables. The Bayes probability works by calculating the maximum likelihood. The conditional probability is applied to calculate the probability for each hypothesis, hence a large dataset can be handled by the classifier.

C. Principal Component Analysis

PCA uses a method dimensionality reduction to reduce the dimension of the feature in the dataset using a proportion of variance. The term principal component is the method that transfers the correlated variable to the limited number of uncorrelated variables. PCA is the compression technique that reduces the features without the loss of information.

The PCA reduces the dimension based on the variance. The variance is estimated by the Eigenvalue and the trace of the covariance matrix. The PCA computed by the Eigenvalue is divided by the covariance matrix. The highest variance is noted in the first PC and the succeeding principal components are orthogonal to the first PC but the variance is lower than the preceding one. PCA improves the accuracy in the data than the machine learning techniques. The Eigenvalue and Eigenvector of the covariance are the important approaches in the principal component.

Dimensionality reduction algorithm is represented in the following steps:

1) Normalize the data and calculate the mean of the X as the data.
2) Subtract the mean value from the X and make it as X.
3) Compute the covariance matrix from the X.
4) Calculate the Eigenvalues and Eigenvectors from the covariance matrix.
5) Principal component (PC) is determined from the Eigenvalue and eigenvector.

IV. RESULTS AND DISCUSSIONS

A. Machine Learning Algorithms

The four classification algorithms are used for the classification. They are Random Forest, Support vector machine, Naive B ayes, K nearest neighbor. The metrics are calculated for these algorithms and find the perfect algorithm that provides a more accurate result. By comparing the accuracy of the above algorithms, the Random forest provides better results. So, the Random forest is preferred to provide the accurate result for the prediction of Parkinson disease.

The reading passage and monologue data are fed into a model. The model split 80% of entries as training and 20% as testing. Sensitivity and specificity are the terms used to calculate accuracy. The model is trained and the classification result is predicted. The various algorithms are used for the classification of Parkinson disease with their performance metrics for reading passage and monologue. The second column denoted the specificity and followed by the sensitivity. The accuracy is estimated in the last column. The random forest is an ensemble technique generating an accuracy of 83% for both reading passage and monologue data.

Table II shows the traditional machine learning algorithms for reading passage data. The performance is evaluated by the Sensitivity, Specificity, and Accuracy.

TABLE II
Performance For Reading Passage Data

ALGORITHM	ACCURACY	SENSITIVITY	SPECIFICITY
RF	0.830	0.769	0.871
SVM	0.769	0.794	0.760
KNN	0.759	0.738	0.784
NAIVE BAYES	0.609	0.493	0.795

The traditional machine learning algorithms results for monologue are shown in Table III. The module calculates the metrics such as sensitivity, specificity, accuracy.

TABLE III
Performance For Monologue Data

ALGORITHM	ACCURACY	SENSITIVITY	SPECIFICITY
RF	0.833	0.782	0.806
SVM	0.794	0.818	0.785
KNN	0.623	0.738	0.784
NAIVE BAYES	0.623	0.506	0.771

B. Performance for Principal Component Analysis

PCA performs dimensionality reduction and the technique is to act as a data compression method. For Parkinson disease twelve features are considered which is reduced to six. The conversion of these twelve features into six Principal components is done using a proportion of variance formula. It eliminates the highest correlation between the features by projecting the features in 6-dimensional spaces. In these, the entropy of speech timing and decay of voiced fricatives, duration of voiced duration and duration of unvoiced duration and latency of respiration exchange and decay of voiced fricatives are highly correlated. The projection technique is used in converting the correlated to uncorrelated variables without the loss of information in the higher dimension. The technique is similar to linear regression by projecting the higher dimension to the lower dimension.

The idea of the PCA is a projection of the correlated feature to the uncorrelated features. The twelve features are converted into the six principal components. The selection of the principal component is computed by the Eigenvalue to the trace of the covariance matrix. The proportion of the variance is ordered in such a way that the highest variance in the first and succeeding has the variance less than the preceding. 91% of the variance is covered by the first 6 principal components.

The features are plotted in the correlation plot. The correlation plot represents the correlated values between the features. The value between the features is high, then the dependence is high, and it affects the prediction of the results. Fig. 1 represents the correlation between the features.

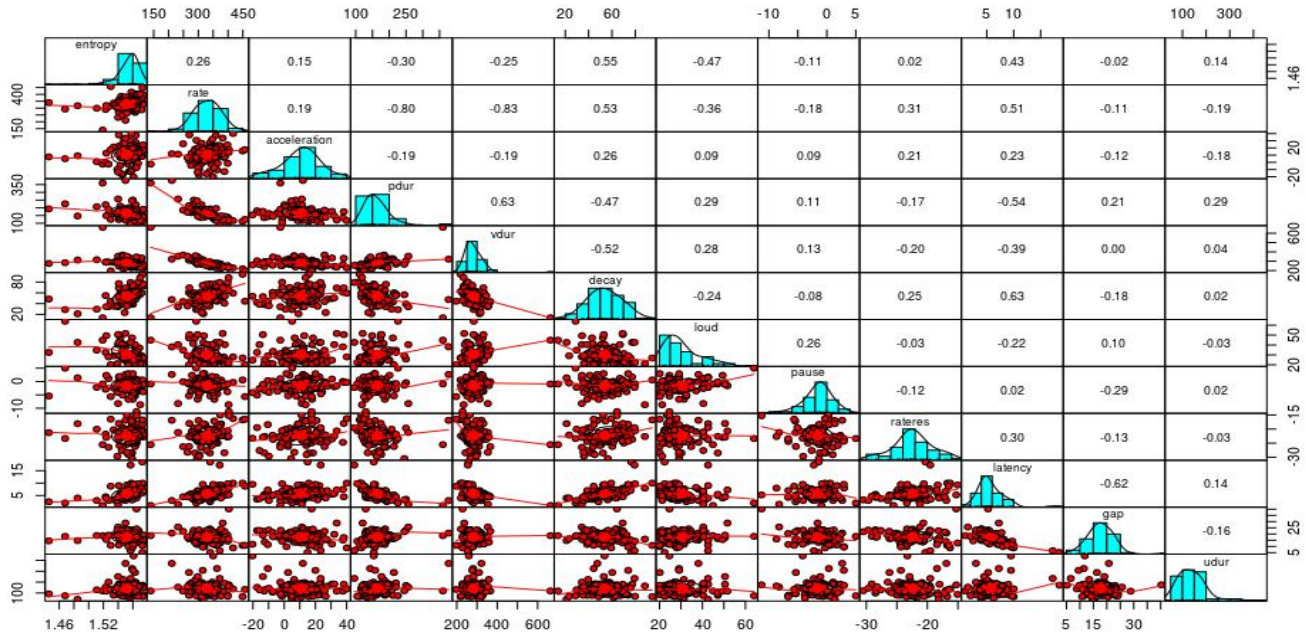


Fig. 1 Correlation between the features

Fig.2 represents the uncorrelated plot. The PCA is used to transform the correlated features to the uncorrelated features. In an uncorrelated plot, the values tend to 0. The uncorrelated features in the plot improve the performance. Fig 2 represents the uncorrelated plot.

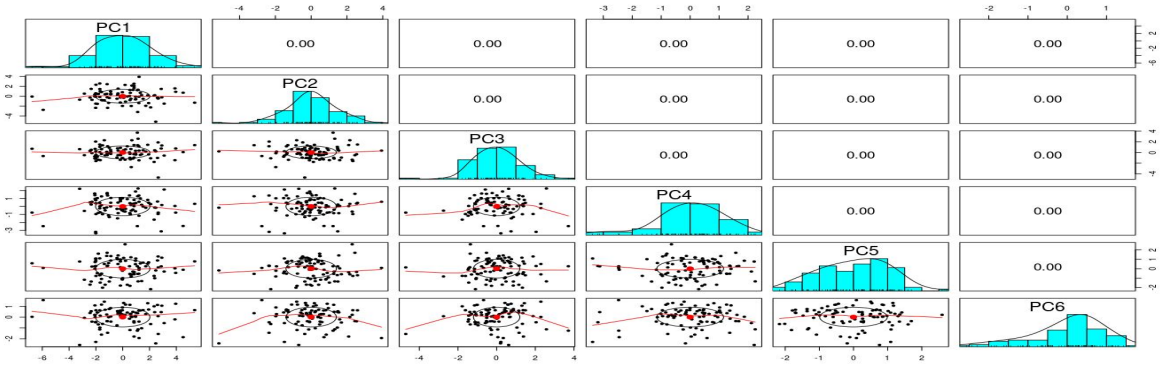


Fig. 1 Uncorrelated plot

The count of the Principal component is determined by the proportion of variance. The graph plots the Principal component against the proportion of variance. The selection of the principal component is based on the proportion of variance. Fig.3 represents the count of PC (Principal component).

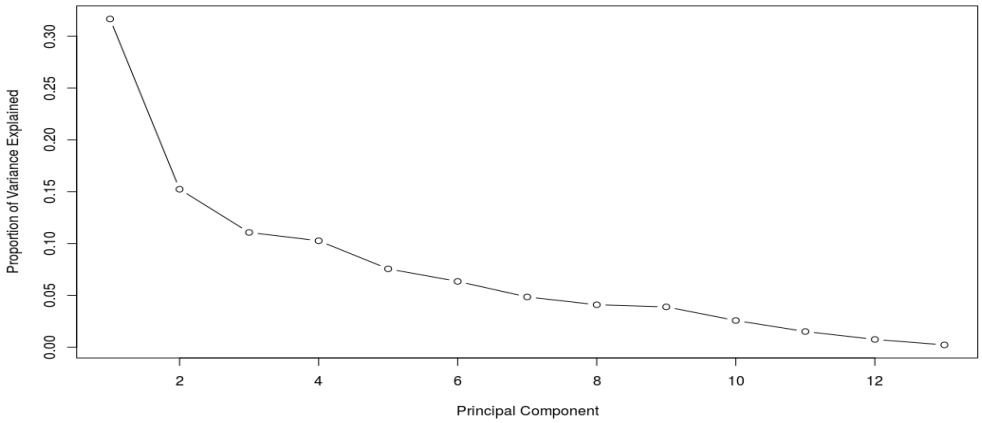


Fig. 3 Determine the count of Principal component

The performance for PCA is compared with the traditional machine learning algorithm. The random forest technique is compared in both the methods. The PCA technique is chosen to improve the accuracy with fewer features. The accuracy of PCA applied to the random forest is 97%. Table IV represents the performance for dimensionality reduction.

TABLE IV
Performance For Dimensionality Reduction

	WITHOUT PCA	WITH PCA
ACCURACY	0.830	0.974
SENSITIVITY	0.769	0.984
SPECIFICITY	0.871	0.964

V. CONCLUSIONS

Parkinson disease leads to death in the worst case, and also has life-impacting symptoms. Diagnosis plays a major role in helping patients to prevent and identify the disease at an early stage. The traditional machine learning algorithm classifies whether the patient has Parkinson disease or not. The irrelevant and dependent features in the data reduce the performance. The PCA is the technique used to reduce the features and provide accurate results. The methods use the dimensionality reduction technique for converting original features to the new features without the loss of data. In the future, more speech features are captured and applied to a deep learning algorithm for accurate and ambiguous result.

REFERENCES

[1] Amr Gaballah, Vijay Parsa, Monika Andreetta, Scott Adams, "Objective and Subjective Speech Quality Assessment of Amplification Devices for Patients With Parkinson's Disease", IEEE Transactions on Biomedical and Engineering, Vol. 27, No.6 , pp.1226 – 1235,2019.

[2] John Prince, Fernando Andreotti, and Maarten De Vos, "Multi-Source Ensemble Learning for the Remote Prediction of Parkinson's Disease in the Presence of Source-Wise Missing Data", IEEE Transactions on Biomedical Engineering, Vol. 66, No. 5, pp. 1402-1411, 2019.

[3] O. Martinez-Manzanera, E. Roosma, M. Beudel, "A Method for Automatic and Objective Scoring of Bradykinesia Using Orientation Sensors and Classification Algorithms", IEEE Transactions on Biomedical Engineering , Vol. 63 , No. 5, pp.1016 – 1024, 2016.

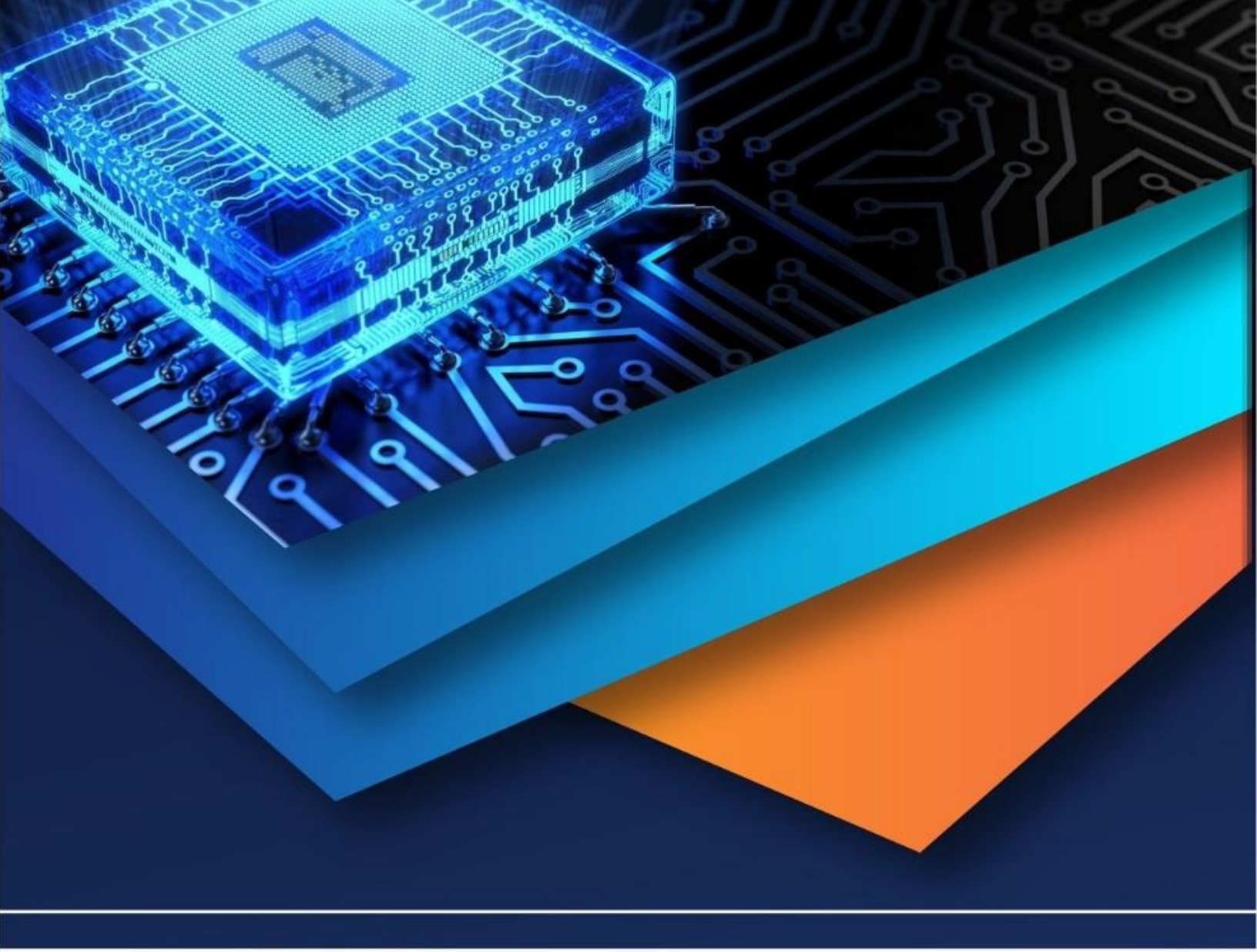
[4] N. Kostikis, D. Hristu-Varsakelis, M. Arnaoutoglou, and C. Kotsavasiloglou, "A Smartphone-Based Tool for Assessing Parkinsonian Hand Tremor ", IEEE Journal of Biomedical and Health Informatics, Vol. 19, No. 6, pp.1835-1842, 2015.

[5] Ferdous Wahid, Rezaul Begg, Chris J. Hass, Saman Halgamuge, David C. Ackland, "Classification of Parkinson's Disease Gait Using Spatial-Temporal Gait Features", IEEE Journal of Biomedical and Health Informatics , Vol. 19 , No. 6, pp.1794 – 1802, 2015.

[6] Achraf Benba, Abdelilah Jilbab, and Ahmed Hammouch, "Discriminating Between Patients With Parkinson's and Neurological Diseases Using Cepstral Analysis", IEEE Transaction on Neural Systems and Rehabilitation Engineering, Vol. 24, No. 10, pp.1100-1108,2016.

[7] Bryan T. Cole, Serge H. Roy, Carlo J. De Luca, S. Hamid Nawab, "Dynamical Learning and Tracking of Tremor and Dyskinesia From Wearable Sensors", IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 22 , No. 5, PP. 982 -991, 2014.

[8] Diane J. Cook, Maureen Schmitter-Edgecombe, and Prafulla Dawadi, "Analyzing Activity Behavior and Movement in a Naturalistic Environment Using Smart Home Techniques", IEEE Journal of Biomedical and Health Informatics, Vol. 19, No. 6, pp.1882-1892, 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY