iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** I **Month of publication:** January 2021

DOI: https://doi.org/10.22214/ijraset.2021.32907

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 9 Issue I Jan 2021- Available at www.ijraset.com

A Comparative Study of Different Machine Learning Algorithms in Detection of Parkinsons Disease

Mansi Singh[1], Naveen Mathews Renji[2]
[1, 2]*Computer Science and Engineering Department, Ramaiah Institute of Technology, Bangalore, India*

Abstract: ***Parkinson's Disease (PD) persistent consideration is constrained by lacking, irregular manifestation checking, rare access to mind, and meager experiences with human services experts prompting poor clinical dynamic and imperfect patient wellbeing related results.Advanced approaches have empowered target and remote checking of impaired motion function with the guarantee of significantly changing the indicative, observing, and helpful detecting in PD. We demonstrated that by using a variety of upper limb functional tests Motor_UPDRS. The objective of this paper is to provide preliminary evidence that machine learning systems allow one to determine whether a person is suffering from Parkinson's disease or not and different features of the disease using various machine learning algorithms .Diagnosis of the Parkinson disease through machine learning provides better understanding from Parkinson's Disease dataset in the present. Jupyter notebook has been used in the present experimentation for the statistical analysis, classification, Evaluation of supervised and unsupervised learning methods. Voice dataset for Parkinson disease has been taken from UCI Machine learning repository from Center for Machine Learning and Intelligent Systems. A study on feature relevance analysis and the accuracy using different classification methods was carried out on Parkinson data-set.***

Keywords: Parkinson's disease diagnosis, voice dataset, movement disorder, machine learning models Mathematical model, Feature extraction, subjective rating scales, Training

I. INTRODUCTION

Parkinson's disease is a neurological issue. The primary signs are issues with movement . Dopamine is created in a part of the brain called the "substantia nigra." In Parkinson's, the cells of the substantia nigra start dying due to decrease in dopamine levels. At the point when they have dropped 60 to 80 percent, symptoms of Parkinson's begin to show up.

The specific reason for Parkinson's is obscure. It might have both genetic and environment components. A few researchers imagine that viruses can trigger Parkinson's also. Low levels of dopamine and norepinephrine, a substance that controls dopamine, have been connected with Parkinson's.Proteins called " Lewy bodies" have additionally been found in the cerebrums of individuals with Parkinson's. Researchers don't know the role of Lewy bodies in the advancement of Parkinson's.

There's no particular test for the diagnosis of Parkinson's.Diagnosis is made dependent on wellbeing history, a physical and neurological test, just as a survey of signs and indications. A dopamine transporter (DAT) sweep may likewise be utilized. While these tests don't affirm Parkinson's, they can help preclude different conditions and bolster the specialist's finding.

Biomarkers for human voice that can offer in-sight into neurological scatters, for example, Parkinson's disease (PD), due to their underlying neuromuscular function.PD causes vocal weakness that impacts "speech", "motor skills" and different functions like conduct, temperament, sensation and thinking. Tele_monitoring of the parkinson's disease utilising voice estimation has an essential job in its initial conclusion of PD.Historically, PD has been hard to amount and doctors have focused on some symptoms while disregarding others, depending principally on subjective rating scales. Because of the decrease in motor control that is the sign of the disease , voice can be utilised as a way to recognise and analyse PD. With progressions in innovation and the prevalence of sound gathering gadgets in day by day lives,models that can translate the sound information into an diagnostic tool for healthcare services experts would conceivably give diagnosis that are less expensive and progressively precise.

II. LITERATURE SURVEY

A. Intelligent Parkinson Disease Prediction ,Analysis of voice recordings is important in the present decade to understand and diagnose Parkinson's diseases.The method provides the diagnosis of PD using voice dataset through machine learning algorithms.This paper has shown that k-NN has 82.5% of accuracy and SVM has shown 88.9% accuracy based on results.

B. Parkinson's disease Gait Classification,UiTM-Svm as classifier performed better compared to ANN specifically for data fusion of gait parameters.Both classifiers attained higher classification rate. Initial results proved that basic spatiotemporal contributed as the best input feature based on perfect accuracy ,sensitivity and specificity rate.

C. In a hybrid intelligent system , the accuracy of the method measured by MAE for the Total_UPDRS and Motor_UPDRS were obtained MAE = 0.4656 and MAE = 0.4967. The result of experimental analysis demonstrates that the proposed method is effective in predicting UPDRS. The method has the potential to be implemented as an intelligent system for Parkinson's disease prediction in healthcare.

D. Diagnosis of Parkinson Disease Using Machine Learning and Data Mining Systems,data and Survey graph provide statistical analysis on the voice data so that the healthy and Parkinson patients would be correctly classified. KStar and NNge algorithms have good accuracy.

E. Parkinson's disease prediction using Bayesian network classifiers consists of four classification models namely naïve Bayes, multivariate filter-based naïve Bayes, filter selective naïve Bayes and support vector machines, SVM have been implemented to evaluate their ability to differentiate between cognitively intact patients with Parkinson's disease (PDCI), PDMCI and PDD.

F. Using an optimized crow search algorithm, an accuracy of 100% is achieved in prediction of PD. It also helps individuals at an early stage to get proper treatment. The implementation of OCSA has been measured for 20 datasets and the results have been compared with the original chaotic crow search algorithm(CCSA). The nature inspired algorithm reveals that accuracy can be maximised and number of features selected can be minimised by selecting an optimal subset of features.

G. Detecting and monitoring the symptoms of Parkinson's disease using smartphones, in a pilot study, these recordings can differentiate those with and without PD.These recordings may be able to predict motor scores in Parkinson disease.These results require further evaluation and confirmation in larger studies.

H. Using Smartphones and Machine Learning to Quantify Parkinson Disease Severity, a smartphone-derived severity score for Parkinson disease is feasible and provides an objective measure of motor symptoms this could be valuable for clinical care and therapeutic development.

III. PROBLEM FORMULATION

Parkinson’s disease is a brain disorder leading to shaking, stiffness and less dopamine which causes movement problems in the person suffering from Parkinson's disease.There are no accurate medical tests to detect and identify the disease.Therefore, there is a need for the detection and identification of parameters in a person suffering from PD.Implementation of a system for the detection and identification of the Total and motor UPDRS attributed to neurological conditions. Various machine algorithms are used on the Parkinson's disease dataset to develop a model capable of determining whether a person is suffering from PD by using the voice recordings.

IV. METHODOLOGY

A. Dataset Information

The dataset contains a variety of biomedical voice measurements from people that have paralysis agitans through tele-monitoring devices for remote symptom progression monitoring. The recordings were automatically captured within the comfort of the patient's homes.

The attributes in the table are "subject number", "subject age", "subject gender", "time interval from baseline recruitment date", "motor UPDRS", "total UPDRS". Each row corresponds to one of 55,875 voice recordings from these individuals. The objective is to predict the motor and total UPDRS(Unified Parkinson Disease Rating Scale) scores from the voice recordings.

B. Libraries

A number of libraries have been imported that are already developed in Python for the implementation of the Machine Learning models.

1) Pandas is used to access the .csv file of our dataset.

2) Matplotlib is used to plot the graph and ‘ply’ is used as the variable name for using this library in our code.

3) Numpy is used to convert the data into structured arrays for practical use by sklearn library.

C. Building a Prediction Model

Once data has been collected, predictive modelling is necessary.This needs to be done to increase the efficiency and accuracy of the model.Only the attributes which play a key role in the analysis of data needs to be considered.In our dataset, we have dropped the “Subject” attribute as there is no correlation of it with the target class.

D. Developing a Model:

The dataset is split into training and testing sets , The training data is used to develop the model by running our algorithm, the testing data is used on the model is correctly predict the accuracy or precision of the model The data is split in a ratio in such a way that prediction does not overfit or underfit values and the correct values will be obtained.

The train_test_split() is an inbuilt function from scikit learn for splitting the x and y variables data. The test_size parameter is used to split the entire dataset into test data and remaining as training data. Setting random_state as 0 will prevent random values to be taken from the dataset.

E. Feature Scaling

It is used here to standardize the independent features present within the data within a fixed range. It is performed to handle highly varying magnitudes or continuous values and units.

1) *StandardScaler:* It standardizes a feature by subtracting the mean and then scaling to unit variance.
2) *Fit_transform:* It is used to calculate the means of columns from data and then replace the missing values. Therefore, for training sets both calculation and transformation is necessary.

V. ALGORITHMS

A. Regressions Algorithms

Linear Regression, Multivariate, Regularised Regression

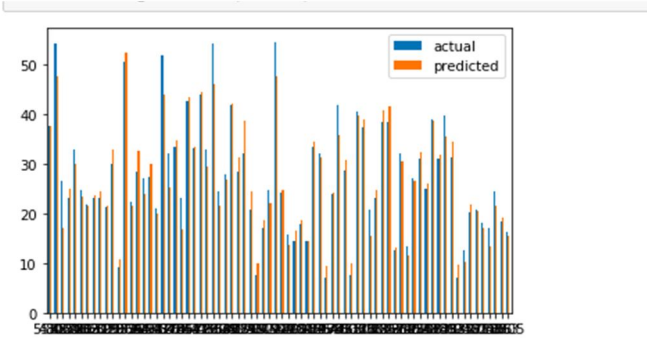
1) *Linear Regression:* In simple linear regression, each experiment is of two values.The dependent variable is one and the Independent variable is the other.

The regression model is given by the mathematical expression:

$$y = \beta 0 + \beta 1 x + \varepsilon$$

where , β0 is the y-intercept, β1 is the slope

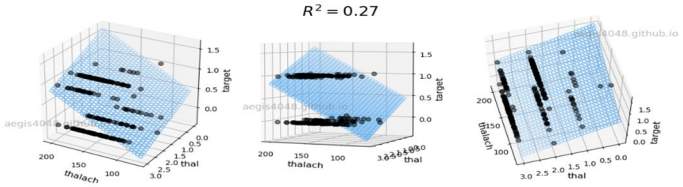
Our model predicts the dependent variable “Y_pred2" using the test values of the independent variable.We can see the variance for the predicted values(Y_pred2) and actual test value of dependent variable(Y_test). Inbuilt methods does the math with the predefined formula for each value.



We achieved an accuracy of 92% for our Linear Regression model, because our regression model contains independent variables which are statistically significant. The statistical significance indicates that changes in the independent variables(all the attributes in our dataset correlate with shifts in the dependent variable (status). This signifies that our model explains a good proportion of the variability in the dependent variable.

This accuracy basically determines the variation from the predicted values from the actual true values of our data. The linear regression model getting an accuracy of 92% indicates that the model is able to correctly predict the values 92% of the time.

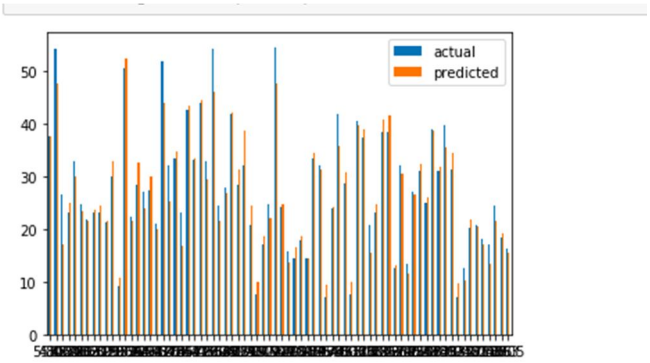
2) *Multivariate Regression:* When there are more than one independent variable, the data must be utilized in multivariate analysis. This is often called multiple regression.



Here we have considered the motor_UPDRS, total_UPDRS and the 'status' as the attributes to plot the 3D graph.We can observe that the multiple regression area falls into the aspect where the status of the patient is 1 when the total and motor UPDRS corresponds to Parkinson's infected patients and it regreses towards 0 when the total and motor UPDRS score falls under a healthy patients score.

The R squared score here is only 0.27, this indicated that the correlation between the hyperplanes of the dependent variables and independent variables. The value of the dependent attribute 'status' is related by an R squared score of 0.27 with the independent attributes motor_UPDRS and total_UPDRS.

3) *Ridge Regression:* It is a type of Regularized regression, which is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are all unbiased, but their variances are large, this is so that they may be far from the true value. The values become dependent on the value of lambda.



We accomplished an accuracy of 92% for our Ridge regression model, on the grounds that our regression model contains independent variables which are factually important. The measurable significance shows that adjustments in the independent variables (all the attributes in our dataset associate with shifts in the dependent variable(status).This connotes that your model clarifies a decent extent of the changeability in the dependent variable. This precision fundamentally decides the variety from the predicted qualities from the real obvious data of our information. The ridge regression model getting a precision of 92.17% demonstrates that the model can effectively predict the 'status' 92.17% of the time. This is because of the correct connection between the independent variables that don't turn into an obstruction for the regression formula from failing.

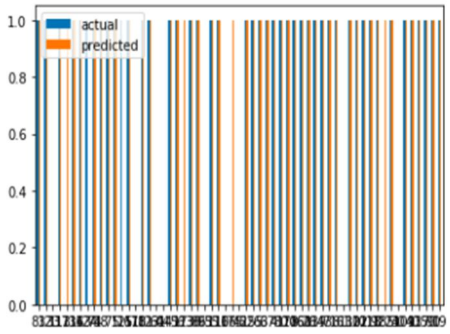
B. Logistic Regression

Logistic Regression is a parametric classification model. It works only with binary data and outputs a categorical prediction.

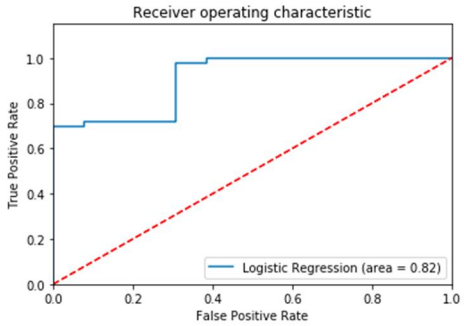
We considered the "status" column out of all the 22 attributes to predict whether a person is healthy or not. The "status" column is a binary attribute and therefore we didn't need to reduce it to categories.

Predict variable (desired target): Y contains the "status", which tells whether the person is healthy or not (binary: "1", means "Parkinson's disease ", "0" means "healthy").

We can now observe the plot of the graph the depicts both the actual values as well as the predicted values by the logistic model.



This is the correlation between the actual values and the predicted values of the logistic regression model.

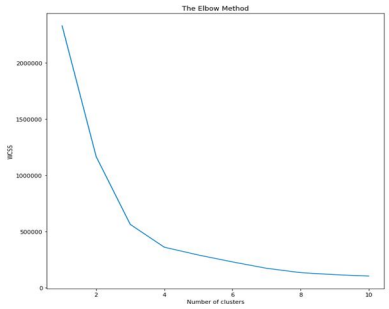


The receiving operating characteristic(ROC) curve is a tool used with binary classifiers. The dotted red line represents the ROC curve is a random classifier. The blue line represents the logistic regression. The blue line is further away from the red dotted line which shows that it is a good classifier.
Logistic regression was able to attain similar accuracies to linear, multiple and ridge regression but its accuracy was lower by 2% comparatively. This can be attributed to overfitting due to logistic regressions over-confident models.
Logistic regression attained an accuracy of 90%.

C. Clustering Algorithms

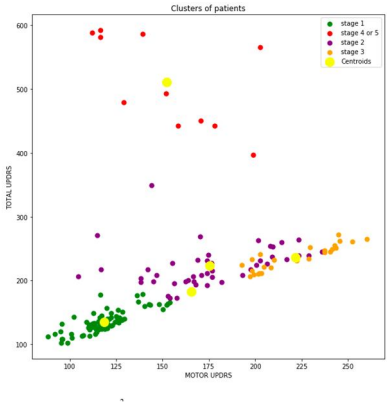
1) K means Clustering Algorithm: The K-means clustering algorithm begins with the random initialisation step. We have considered all the 21 attributes for our training set 'X' because it is relevant data. The independent variable 'y' is "status". We have taken into consideration 11 iterations to get the most accurate model. We have made clusters for the different stages of Parkinson's disease .

a) Elbow Method: The elbow method for all the values of the k calculates an average score for all clusters.Our line chart resembles an arm, then the "elbow" (the point of inflection on the curve) is a good indication that the underlying model will fit best at that point. The scoring parameter metric is set to distortion, which calculates the sum of squared distances from each point to the assigned center. From this method we come to the conclusion that 4 clusters can be formed for our model from the below graph.



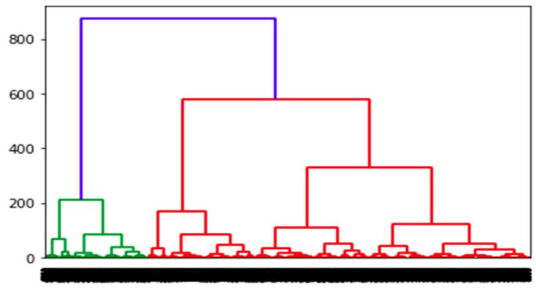
From our model we have obtained 4 clusters mainly which indicate the stages of Parkinson's disease in humans :

- *Stage 1 (Green Cluster):* During this initial stage, the person will have mild symptoms that generally do not interfere with daily activities. Tremor and most other movement symptoms occur on only one side of the body only. Changes in posture, walking and facial expressions occur during this stage.
- *Stage 2 (Purple Cluster):* Symptoms start getting worse during this stage. Tremor, rigidity and most movement symptoms affect both sides of the body. Walking problems and poor posture might be apparent. The person will still be able to live alone, but daily tasks are more difficult and lengthy.
- *Stage 3(Orange Cluster):* During mid-stage, loss of balance and slowness of movements are very significant. Falls are very common. The person is still fully independent, but symptoms will significantly impair activities such as dressing and eating.
- *Stage 4 and 5 (Red Cluster):* During this stage, symptoms will get severe and limiting. It is possible to stand without assistance, but movement requires a walker. The person needs help with activities of daily living and will not be able to live alone.This is the most advanced and debilitating stage. The person requires a wheelchair or is bedridden.

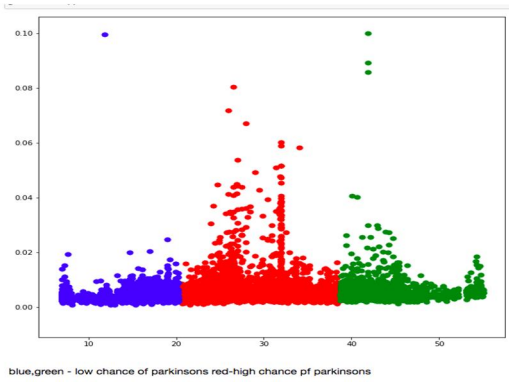


The graph above shows the number of patients suffering from the different stages of Parkinson's disease. The accuracy of our model is 50.2%. Even though the different clusters are visible to us by eye. This is because in K Means clustering, the algorithm confers equal radius to each cluster sphere as it considers Euclidean distance which results in the algorithm testing the result space as isotropic. Because of this equal radius, spherical assumption is violated and the algorithm predicts in a non-intuitive manner. Another reason for the low accuracy may be due to the number of K groupings in our algorithm.

2) *Hierarchical Agglomerative Clustering:* We use complete linkage and operate on our given dataset using euclidean distance and observe that in the dendrogram, the highest vertical distance that does not intersect with any clusters, is the middle Red one. Given that 3 vertical lines cross the threshold (when the linkage distance is the highest, then this threshold value is optimal, because the dissimilarity among classes is maximal). Thus with this information we are able to deduce that the optimal number of clusters is 3.



The x-axis contains the samples and y-axis gives the distance between these samples. The vertical line with the maximum distance is the blue line and hence we can decide a threshold of 400 and cut the dendrogram. We have two clusters as this line cuts the dendrogram at two points. We applied hierarchical clustering for 3 clusters.



We clearly visualise three clusters which indicates

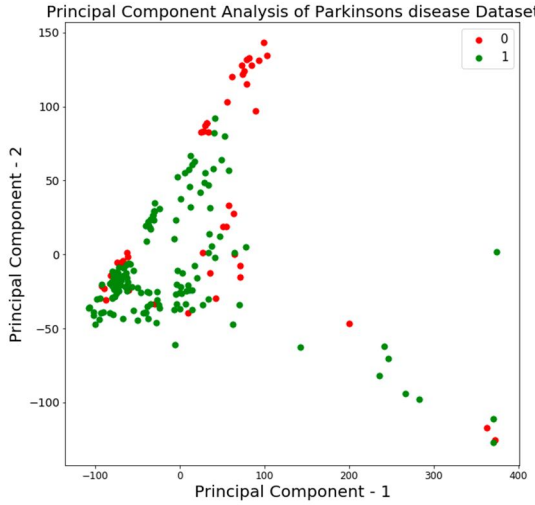
a) Blue cluster and green cluster represent that there are lower chances of a person getting Parkinson's disease.

b) Red cluster represents that there are high chances of a person getting Parkinson's disease.

D. Dimensionality Reduction

1) 6.Principal Component Analysis: PCA is used to identify patterns in data on the basis of correlation between features. We successfully converted the dataset into a PCA model with namely two principal components. Using explained_variance_ratio_ function we can see that the first principal component contains 72.8% of the variance and the second principal component contains 21.8% of the variance. Together, the two components contain 94.6% of the information. Almost 94% of the model was fit using these two components losing 6% of the efficiency.

This section is plotting 2D data.The graph shows that the classes seem well separated from each other.



In the graph above the Red Class is represented by zero which means that the person is healthy.And the green class is represented by 1 which indicates that the person has Parkinson's disease. PCA is a multivariate ordination analysis.It has ordered the samples in a plane defined by two axes PCA 1 and PCA 2 according to their continuous values.

E. Classification Algorithms

1) Decision Tree: We have taken all the other 21 attributes as our independent variables X because all of those features have an effect on whether the person is suffering from Parkinson's disease or not . Here, "status" is the dependent variable "Y". We have considered 27% of our dataset as a test sample and after multiple trial and errors, this was found to be the optimal split ratio. The classification rate is 84%. - This means the decision tree is able to correctly predict the target class (if the person suffers from PD or not) from the given attributes.

	precision	recall	f1-score	support
0	0.61	0.85	0.71	1802
1	0.94	0.82	0.88	5686
accuracy			0.83	7488
macro avg	0.78	0.84	0.79	7488
weighted avg	0.86	0.83	0.84	7488

a) Precision is the ability of a classiifer not to label an instance positive that is actually negative.We have got a weighted avg precision of 89%.

b) Recall is the ability of a classifier to find all positive instances.Our model for all instances that were actually positive, 82% was classified correctly.

c) The F1 score is a weighted harmonic mean of precision and that the best score is 1.0 and the worst is 0.0. The weighted average of F1 is 86%.

We are able to attain a reasonable good score because a decision tree traces out all possible outcomes and considers every path to the target class until it reaches a conclusion with the highest precision.

In our case, we were able to achieve higher accuracy from regression algorithms than this decision tree algorithm, mainly because of the sensitivity of the decision tree meaning a small change in the data leads to a great change within the decision tree. This leads to untable outputs which often lack the accuracy they receive from other classification algorithms. Continuous valued attributes also present a challenge for decision trees and in our dataset , the majority of the attributes have a continuous value in the range of values.

F. Artificial Neural Network

The dataset is split into training and testing data and the training data is fed into a neural network. The neural network was created using python library keras with a tensorflow backend. This module enabled us to create multiple hidden layers within the neural network each having a randomized filter that upon backpropagation over 100 epochs.Our 2 hidden layers use relu activation to ensure that the function maintains the values between 0->x (x being the max input value).

We use an adam optimizer to ensure weights of each neuron in each layer in each epoch is changed according to the parameters without succumbing to timing issues on reaching the global minimum. The output layer finally projects the probabilities of the input setting the correct class as the output .

We were able to produce an output towards our target class with an accuracy of 98.84%.

Artificial Neural Network using backpropagation has enabled us to achieve the highest accuracy model , this is mainly because of the work of the 2 hidden layers and because of our mini-batches of size 10, we were able to minimize computation time as well as optimize the efficiency and accuracy of the network . The model was run for 100 epochs with batch size of 10 until it attained an accuracy of 98.70%.

G. Support Vector Machine

	precision	recall	f1-score	support
0	0.87	0.65	0.75	1802
1	0.90	0.97	0.93	5686
accuracy			0.89	7488
macro avg	0.88	0.81	0.84	7488
weighted avg	0.89	0.89	0.89	7488

Observing the confusion_matrix(y_test,y_pred) and classification_report(y_test,y_pred) - we are able to clearly see how well the model was trained and tested. Performed greatly in its prediction capabilities with a high number of true positives and negatives and very low count of false positives and false negatives.

SVM classifiers outshine other classifiers if there is a clear boundary of separation by the hyperplane. But in this case, due to many data points lying within the close proximity of the boundary of separation, it is unable to classify as accurately. SVM also takes more computational time on datasets with a high size.

After fitting the model and creating its kernel with a boundary separation line The SVM algorithm gave an accuracy of 89.329%.

H. Adaboost Algorithm

	precision	recall	f1-score	support
0	0.88	0.87	0.87	3685
1	0.96	0.96	0.96	11291
accuracy			0.94	14976
macro avg	0.92	0.91	0.92	14976
weighted avg	0.94	0.94	0.94	14976

A classification report is used to measure the quality of predictions from the Adaboost classification algorithm.

From the dataset used, “status” is the target class

0 - indicates that the person is healthy

1- indicates that the person is suffering from PD

It iteratively corrects the mistakes of the weak classifier and improves accuracy by combining weak learners. We can observe a very good precision in the predicted output. Adaboost hyper parameter was specified at 300 in our case. This high accuracy in the model can be credited to Adaboost's advantage of preventing overfitting.

When we previously computed the accuracy of the decision tree model prior to boosting, we only got an accuracy of 84.9% which was low compared to other classifier, after Adaboost classifier was applied to decision tree classifier we were able to increase the accuracy of the model by 10% which is very significant rise in the accuracy.
The accuracy of this algorithm is 94%.

I. Naive Bayesian Algorithm

Common applications include filtering spam, classifying documents, sentiment prediction. It works on conditional probability. From the conditional probability, we can calculate the probability of an event using its prior knowledge.

	precision	recall	f1-score	support
0	0.51	0.63	0.57	3045
1	0.87	0.80	0.84	9435
accuracy			0.76	12480
macro avg	0.69	0.72	0.70	12480
weighted avg	0.78	0.76	0.77	12480

A classification report is used to measure the quality of predictions from the Naive Bayesian algorithm.

From the dataset used, “status” is the target class

0 - indicates that the person is healthy

1- indicates that the person is suffering from PD

We can observe that compared to all the other algorithms,Naive bayesian classifier has the lowest accuracy in its prediction model as shown above. This mainly because of its feature independence assumption and treatment of all attributes equally regardless of their importance. This means that, no matter how less a particular attribute may vary in importance from another attribute which in our case, the important attribute could be (total_UPDRS or motor_UPDRS), it will still be weighted with equal importance .
Naive Bayes is generative model, this means that it determines the Class label of the input based on the bayes rules calculate p(y | x), and then picking the most likely label of ‘status’.
The accuracy of this algorithm is only 76.28%.

J. K Nearest Neighbour Algorithm

KNN has been used in statistical estimation and pattern recognition.To select the K that’s right for your data, run the KNN algorithm several times with different values of K and choose the K that reduces the number of errors we encounter while maintaining the algorithm’s ability to accurately make predictions when it’s given data it hasn’t seen before.

	precision	recall	f1-score	support
0	0.94	0.87	0.90	1802
1	0.96	0.98	0.97	5686
accuracy			0.95	7488
macro avg	0.95	0.93	0.94	7488
weighted avg	0.95	0.95	0.95	7488

A classification report is used to measure the quality of predictions from the KNN algorithm.

From the dataset used, “status” is the target class

0 - indicates that the person is healthy

1- indicates that the person is suffering from PD

KNN trains by computing the euclidean distance between the instances and their respective attributes from other instances. This means that the instance will be labeled according to its closest neighbour. Our X values are computed for their distance from neighbours, there will be a disadvantage of computational time but this can be overlooked due to the fact that we were able to achieve a very high accuracy form the algorithm. The main advantage of KNN and the reason it fared better than other algorithms is because of its lack of assumptions on the data being processed.
KNN was able to correctly classify 95.41% of the time because of its efficient similarity measure criteria for every case.

K. Genetic Algorithm

A Genetic Algorithm is a heuristic search that is inspired by Charles Darwin's theory of natural evolution. This algorithm is the process of natural selection where the most suitable features are selected for for the next iteration to provide optimal results for a given dataset.

The output of the program is transformed through a sigmoid function in order to transform the numeric output into probabilities of each class. In essence this means that a negative output of a function means that the program is predicting one class, and a positive output predicts the other.

Here we are able to observe that the SymbolicClassifier was able to find non-linear decision boundaries. The genetic algorithm passes tthe dataset to the est() function which is part of the gplear module. In the est() function, multiple generations of the data are generated and these generations of the data are labelled accordingly by extrapolating its classes. The est() function can now after fitting and running genetic mutations on the real datasets and will now be able to predict values of the target class by only reading the values of the remaining X attributes. After testing the estimator on the remaining samples, we were able to have a prediction accuracy of 90.22%.

VI. RESULTS

A. Regression Algorithms

Algorithm Name	Accuracy / Score	Inference
Linear Regression	92%	Fared well and was able to predict the right target class similar to the other regression algorithms.
Multiples regression	R squared score of 0.27	Did not fare well because of using 'total_UPDRS' and motor_UPDRS as independent variables to predict the status of PD, and these two attributes have high correlation even though they are the most significant attributes.
Ridge Regression	92%	Fared well and was able to predict the right target class similar to the other regression algorithms.
Logistic Regression	90%	Fared well and was able to predict the right target class similar to the other regression algorithms.

B. Clustering Algorithms

Algorithm Name	Inference
K-means Clustering Algorithm	Though there was visual separation in the different clusters, the algorithm failed to perform well and was only able to predict the right cluster 50.2% of the time due to spherical assumption violation.
Hierarchical Agglomerative Clustering Algorithm	HAC was much more accurate in helping us visualize the 3 different clusters of where the two extreme clusters of blue and green were safe zones and the middle red cluster were patient with the 'status'=1. This is because HAC prioritises dissimilarity to classify and cluster data unlike K-means similarity criteria.

C. Dimensionality Reduction

Algorithm Name	Inference
Principal Component Analysis	PCA was able to determine two principal components with 72.8% and 21.8% variance (total 94.6%) , this means using only the 2 components to determine outputs would fare well and only around 6% of efficiency would be reduced.

D. Classification Algorithms

Algorithm Name	Accuracy	Inference
Decision Tree Algorithm	84.9%	It performed relatively inadequately because of a decision tree lacking good performance under continuous valued attributes being the input.
Artificial Neural Networks	98.7%	Performed the best among all the classifiers and regressors . This was due to the accuracy achieved by back-propagation through our network consisting of 2 hidden layers with an adam optimizer ensuring weights are calculated with efficiency for each neuron.
Support Vector Machines	89.32%	It took a comparatively much higher time to finish its computation and training and still couldn't determine a very accurate separation line between the instances.
Genetic Algorithm	90.22%	It performed quite well and took the least amount of time to train and make predictions , this is due to its advanced generative functionalities which make the estimator of the genetic algorithm capable of prediction faster and with higher accuracy than other classifiers.
Adaboost Algorithm	93.81%	By boosting the Decision tree algorithm with Adaboost, the accuracy was increased by almost 10% which is very significant.
Naive Bayesian Algorithm	76.28%	Fared the least well among all the algorithms, mainly due to the algorithm giving equal importance to all attributes and calculating p(y|x) to pick the most likely 'y' value with highly variable x values.
K-Nearest Neighbour	95.41%	The similarity criteria of measuring the nearest distance to determine the class label worked well as the same class instances had low differences in distance between their attribute values.

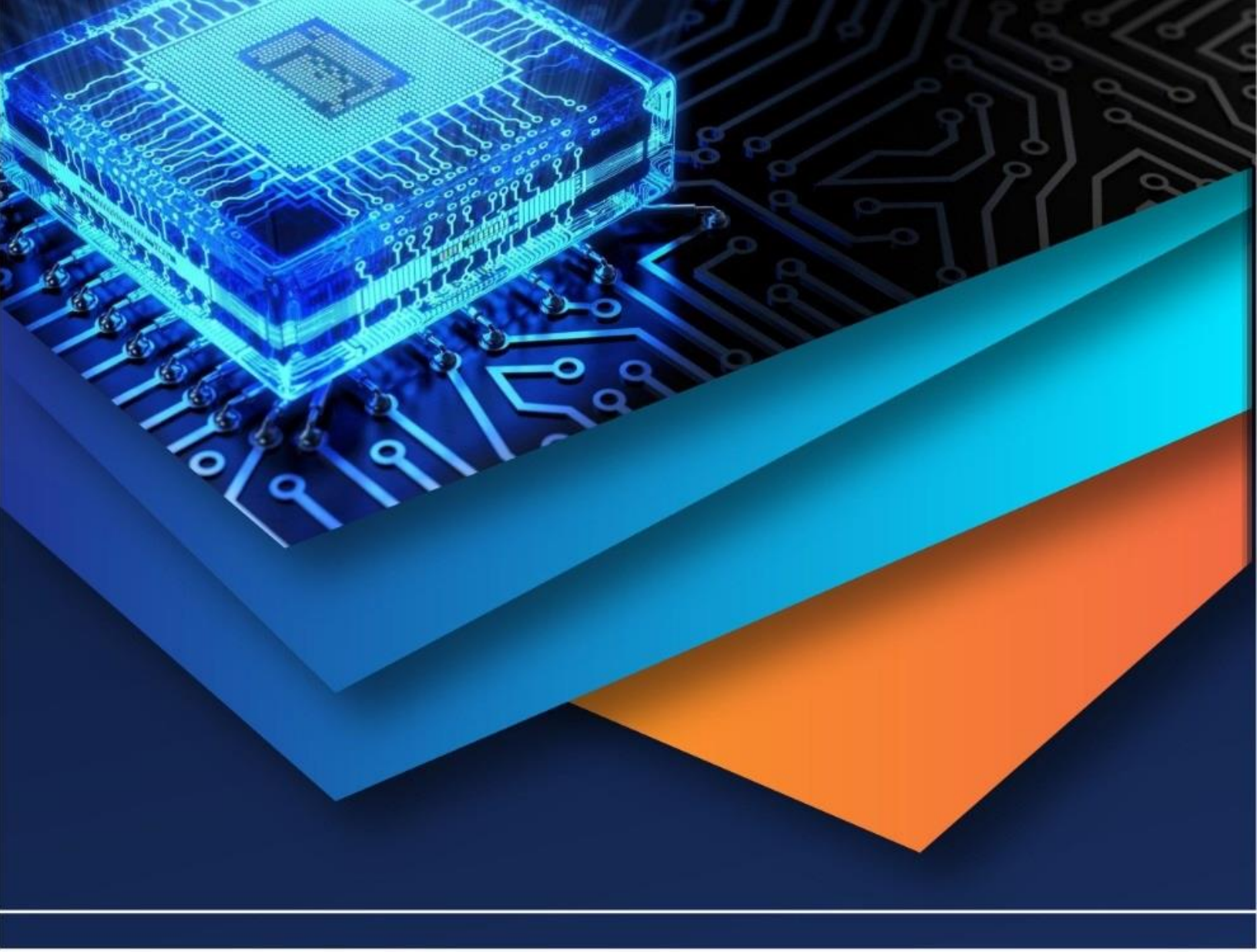
VII. CONCLUSION

This paper presented a review for the prediction of Parkinson's disease by using machine learning algorithms. A brief introduction of 12 machine learning based approaches used for the prediction of Parkinson disease.The summary of results obtained by various researchers is available in a literature survey to predict the Parkinson diseases is also presented.

We are able to infer from our experiments and analysis that though all the different classification algorithms perform well, some work better on a given dataset due to parameterized factors as well as algorithmic variation that affect the accuracy as well as precision of the prediction. Hence we are able to conclude that our Artificial neural network which used the back propagation method ,with 2 hidden layers running using backend tensor flow, keras library functions was able to secure the highest accuracy and produce the most precise model for classification of a patient into healthy or suffering from parkinson's disease. Our regression models gave similar outputs regardless of whether it was ridge regression , linear regression, multiple regression or logistic regression. This indicates that there is a strong relation between the entities that are porous through all the three regression algorithms.

REFERENCES

[1] J. A. Obeso, C. W. Olanow and J. G. Nutt, Levodopa motor complications in Parkinson's disease, 2000.
[2] H. Peng, F. Long and C. Ding, "Feature selection based on mutual information criteria of max-dependency max-relevance and min redundancy", IEEE Transactions on pattern analysis and machine intelligence, vol. 27, no. 8, pp. 1226-1238, 2005.
[3] L. Jeancolas, H. Benali, B. E. Benkelfat, G. Mangone, J. C. Corvol, M. Vidailhet, et al., "Automatic detection of early stages of Parkinson's disease through acoustic voice analysis with mel-frequency cepstral coefficients", 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), pp. 1-6, May 2017.
[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., "Scikit-learn: Machine learning in python", Journal of machine learning research, vol. 12, no. Oct, pp. 2825-2830, 2011.
[5] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology", IEEE transactions on systems man and cybernetics, vol. 21, no. 3, pp. 660-674, 1991.
[6] B. Pittman, R. Hosseini Ghomi and D. Si, "Parkinson's disease classification of mpower walking activity participants",



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY