Hybrid Predictive Model on Detection of Neurodegenerative Disorder using Machine Learning Classification Algorithms

OGUOMA IKECHUKWU STANLEY[1], AGBAKWURU A.O[2], AMANZE B.C[3]
[1]*Department of Computer Science, University of Agriculture and Environmental Science Umuagwo, Imo State, Nigeria*
[2, 3]*Department of Computer Science, Imo State University Owerri, Nigeria*

Abstract- ***The aim of this paper is to design a Hybrid Predictive Model on Detection of Neurodegenerative Disorder using Machine Learning Classification Algorithms, with major focus on Parkinson disease while the objectives is to ensure that the developed model can detect a patients speech movement pattern, speech motor speed and suggest major causes of the disease. The motivation towards this study is based on accuracy on Parkinson disease prediction by including more variables involving patient's family history, blood test and protein level so as to improve on the existing dataset used by (Yan et al., 2020). The study employed three machine learning classification algorithm methods which include: Support Vector Machine, Neural Network and Decision Tree algorithms. The data was analyzed with R and JASP while the experiments are done on the dataset which contains (Patients Medical History, Laboratory Tests, Imaging Studies, body movement, speech timing, and family health history, antiparkinson medication sourced from UCI and Kaggle machine learning repository. The result after the experiment shows that the use of a hybrid approach involving three classification algorithms in health related data prediction to develop a model called (Ogu-Ikem-Neurodegenerative-Parkinson-Model) is one of the best and more accurate method suitable for data prediction and hence has more percentage acceptance level when it comes to health issues, therefore it could be adopted for future use by medical practitioners to make decision on the subject matter.***

Indexed Terms- ***Artificial Intelligence, Machine Learning, Health Science, Classification Model, Parkinson disease prediction, health diagnosis and prediction of neurodegenerative disorder, Hybrid Predictive Model on neurodegenerative disorder.***

I. INTRODUCTION

Neurodegenerative diseases are a class of neurological disorders where neurons from the central nervous system die or are damaged causing severe disabilities, and eventually death. It can also be a type of disease in which cells of the central nervous system stop working or die (NCI, 2023). Neurodegenerative disorders usually get worse over time and have no cure. They may be genetic or be caused by a tumor or stroke. Neurodegenerative disorders also occur in people who drink large amounts of alcohol or are exposed to certain viruses or toxins. Examples of neurodegenerative disorders include amyotrophic lateral sclerosis, multiple sclerosis, Parkinson's disease, Alzheimer's disease, Huntington's disease, multiple system atrophy, tauopathies, and prion diseases. They are typically encountered in old age which might appear earlier. In the past years, their incidence increased significantly and it is expected that the increase will continue, as the world's population ages (Laske et al., 2015). Neurodegenerative diseases are problematic and can become a burden since their cause is unknown and no cure has been discovered. Treatments are currently targeting the alleviation of symptoms and due to recent advances in artificial intelligence, a significant help can come from the computational approaches targeting diagnosis and monitoring, e.g., detection of disease onset, characterization of the disease, improvement of the differential diagnosis, quantification of the disease progression, tracking of the medication effects. These tasks can be automated or at least improved with the help of machine learning algorithms and intelligent modeling tools. It has been estimated that nearly 6.8 million people expired every year due to neurological disorders (Zhang et al., 2017). The population keeps on mounting because of state-

of-the-art medical advancements and hygiene which affects; the ageing populations by increasing number of people suffer from neurodegenerative maladies. Therefore, it is crucial to diagnose the neuro related diseases at an early age to curtail the damages that the diseases impart on the human brain. Early detection of the diseases at a prior stage would warrant accurate diagnoses which will enable imparting correct treatment at an early stage. However, detecting neuro diseases at an early stage is challenging, not only for the affected individuals or their caregiver who may not recognize the initial symptoms but also for the clinicians who may not be able to diagnose the condition confidently. The symptoms of neurodegenerative diseases are generally voice impairment, loss of memory, difficulty in gait movement etc.

II. LITERATURE REVIEW

Machine learning algorithms can detect subtle changes in brain structure and function, helping to distinguish individuals with HD from healthy controls and providing a means to monitor disease progression over time. Additionally, machine learning models can analyze clinical data, including motor, cognitive, and psychiatric assessments, to identify relevant features and patterns that contribute to accurate diagnosis and prognosis (Lois *et al*., 2018). It also plays a crucial role in predictive modeling for HD risk assessment. By incorporating genetic data and other relevant factors, machine learning algorithms can predict an individual's likelihood of developing HD, aiding in early intervention and counseling.

Felix *et al*., (2019) also use Decision Tree which stands out as a highly effective tool in the diagnosis of Huntington's disease. Decision Tree achieved an impressive average accuracy of 100% in accurately classifying gait signals from subjects with HD. This remarkable accuracy underscores the robustness of the Decision Tree algorithm in distinguishing individuals with HD based on their gait dynamics. Additionally, the Decision Tree emerges as a pivotal machine learning algorithm employed for the prediction and identification of potential contributing genes in Huntington's disease (Cheng *et al*., 2020). According to (Mannini *et al*., 2016) Support Vector Machine (SVM) emerges as a crucial classifier for gait classification, playing a significant role in the context of Huntington's disease diagnosis, alongside other pathological conditions. The utilization of SVM to differentiate gait patterns among diverse clinical groups, including individuals with Huntington's disease, post-stroke patients, and healthy elderly individuals, employing data collected from inertial sensors.

III. METHODOLOGY

Three different classification algorithms was applied in this paper, Decision tree, neural network algorithms and support vector machine because of their ability to uncover or translate hidden pattern from a model and also accuracy in data prediction for effective decision making in real life.

IV. DECISION TREE ALGORITHMS

Decision tree (Han and Kamber; 2001) could be seen as a type of tree structure typically in a form of flowchart design. These tree structures are used to carry out classification and prediction modeling of objects in a class in a form of nodes and internodes. Both root and the internal nodes are taken as the test cases in the modeling process which in terms used as a separator with different features(Han and Kamber; 2001). According to (Sidana; 2017), these decision trees uses a classification or regression models to form a tree structure. The structure breaks down a particular data set into various smaller and smaller subsets as the associated decision tree development id in progress. The researcher further noted that the decision tree is build up with the nodes and leaf nodes, where the decision nodes has two or more different branches while leaf nodes shows the classification or decision results(Sidana; 2017) stated. Figure 1: Illustrate the structure of a decision tree

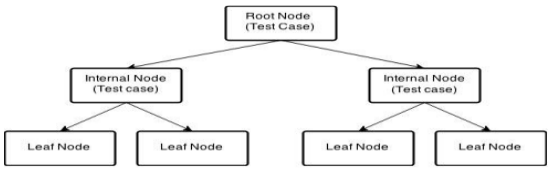

Figure 1: Illustrate the structure of a decision tree (Source: Pilani, Dubai, and Sumbaly; 2015)

Adoption of decision tree for this study did not just come but it was adopted because of its powerful technique for classification and prediction ability on a particular data set. Hence the identification and prediction of diabetes disease in a patient will have a very significant outcome after the analysis of the data set has been concluded and presented for future use. The proposed system diagram is shown in figure 2 below:

V. SUPPORT VECTOR MACHINE (SVM) ALGORITHM

This algorithm is a popular machine learning technique used for classification and regression analysis. Here are some common application areas of SVM algorithms: Image classification, Text classification, Bioinformatics, Financial forecasting and Medical diagnosis. These algorithms offer robust performance, especially in high-dimensional spaces, and are often used as a baseline for comparison with other machine learning techniques.

VI. NEURAL NETWORK ALGORITHM

ANALYSIS OF THE EXISTING SYSTEM

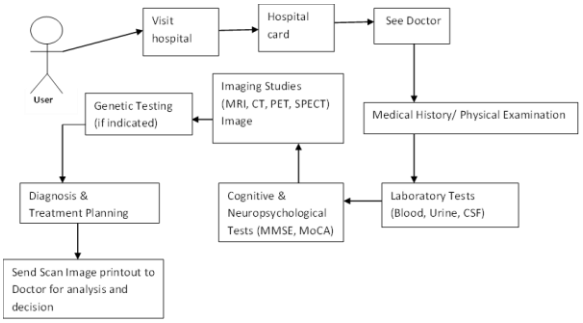

Figure 2: Analysis of the existing System

In this system analysis phase, a patient visits the hospital which a patient hospital card is requires before the patient can see the medical doctor for diagnosis and other necessary examination. The medical doctor first checks the patients' medical history before partaking into the physical examination which will enable the doctor to recommend for further medical examination either by Imaging Studies (MRI, CT, PET, SPECT) or Genetic Testing. Once the results of the genetic test or virtual scans are out, it will be forwarded to the medical Doctor for diagnosis and treatment plan can be initiated on the patient. The diagrammatical analysis of the existing system is shown figure 2 above.

VII. THE PROPOSED SYSTEM DIAGRAM

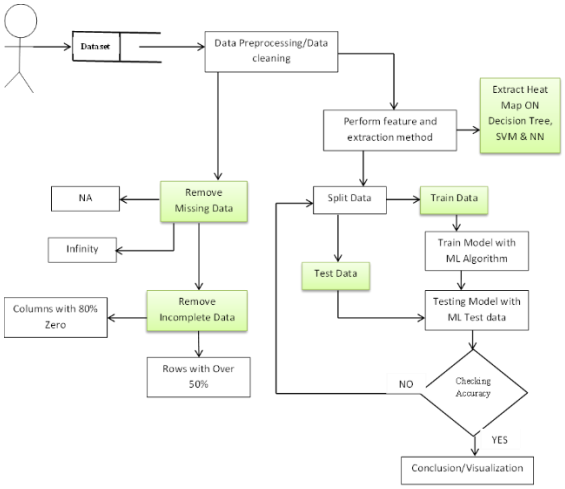

Figure 2: Diagram of the proposed model

The above diagram illustrates how the proposed system loads the Parkinson dataset into the JASP machine learning platform. The dataset was first of all undergo data preprocessing/data cleaning stage, it is here that the researcher understood and discovered some missing values which was removed and hence inclusion of the new collected data containing (blood information, protein level and patient movement). Then extraction and feature process are performed. The dataset was now Split into two parts (TEST and TRAIN) with a percentage rate of 20% and 80% respectively before checking on the extract heat Map on the three algorithms (Decision Tree, SVM and KNN before applying the various machine learning algorithms to produce the proposed predictive model called (Ogu-Ikem-Neurodegenerative-Parkinson-Model).

In summary, the researcher adopted the R and JASP analytical platform for the analysis while dataset was gotten from Kaggle repository and UCI machine learning repository and from locally hospital here in Owerri Municipal, Imo State Nigeria (Federal Medical Center FMC).

Table 1: SYSTEM ALGORITHM

INPUT	Kaggle repository and UCI machine learning repository and from locally hospital here in Owerri Municipal, Imo State Nigeria (Federal Medical Center FMC).
OUTPUT	Hybrid Predictive Model on Detection of Neurodegenerative Disorder using Machine Learning Classification Algorithms

	Model Name: (Ogu-Ikem-Neurodegenerative-Parkinson-Model).

VIII. RESULTS

EXPERIMENTS ON THE DATASET USING R

The first process was launching of the RStudio IDE after a successive launching, the following steps were done to design the model.

Step1: Loading packages to be used (that is libraries)
Step 2: Loading My Dataset to R Dataframe
Step 3: Exploring the data, at this stage, skimr::skim (Parkinson dataset (PD)) was used
Step 4: Converting outcome from numeric to factor and renaming them for easy understanding
Step 5: Changing some response by patients from alphabet to Numeric (YES = 1 and NO = 0)
Step 6: Plot Observations to view all the dataset
Step 7: Checking For Missing Values in Each Variable
Step 8: Replacing or removal of the missing values for each variable
Step 9: Normalizing the dataset

MODEL BUILDING

Step 10: The dataset was Split into two with the percentage of (80% = training and 20%=testing) respectively
Step 11: applying decision tree algorithm
Step 12: Using GINI MODEL to build (Ogu-Ikem-Neurodegenerative-Parkinson-Model).
Step 13: Analyzing the Prediction of the Model built With Gini Model
Step 14: Confusion Matrix
Hence the confusion matrix is used for more accuracy on the rate at which the model predicts user data.
Step 15: Validating the Model On the test Dataset
Step 11 was carried out again to use Neural Network and SVM algorithms on the same dataset in other to complete the hybrid use of the three algorithms and it is done following other steps below for a more accurate prediction of the model produced.

EXPERIMENTS ON THE DATASET USING JASP PLATFORM

The first process was launching of the JASP PLATFORM after a successive launching, the following steps were done to design the model.

Step 1: Loading the dataset from the location (Parkinson dataset (PD))
Step 2: Select machine learning packages
Step 3: Select first classification algorithm (Decision Tree)
Step 4: Set the target and features (clinical information, motor evaluation, Speech evaluation)
Step 5: Click to start the analysis on the dataset
Step 6: Download visualizations
Step 7: Start prediction accuracy by checking (F1 score, confusion matric, and Roc Curve and Decision Tree conditions).
Step 3: was carried out again to use Neural Network and SVM algorithms on the same dataset in other to complete the hybrid use of the three classification algorithms and it is done following other steps below for a more accurate prediction of the model produced which are predicted by looking at the output F1 score, ROC curve, confusion matric and decision tree models built.

EXPERIMENT OUTPUT

Figure 3: Basic Statistics of Diabetes data set (Fieldwork 2022)

Figure 4: Diabetes data structure (Fieldwork 2022)
Figure 5: Frequency Plot of Age against Pregnancy (Fieldwork 2022)
Figure 6: exploration of the dataset (Fieldwork 2022)
Figure 7: Gini model result on the data set (Fieldwork 2022)
Figure 7 above shows the decision tree model on how to predict if patient diabetes is positive or negative.

THE MODEL RULES USING R

If (YES) patient Glucose is >= 0.71 it means the patient has diabetes (Positive with 16%) but if (NO) we have (Negative with 84%) else If (NO) patient Age is >= 0.13 it means the patient diabetes is (Negative with 43%) else if (NO) patients BMI >=0.16 it means the patient diabetes is (Negative with 6%) else if (NO) patient Glucose >=0.36 it means the patient diabetes is (Negative with 6%) but if (YES) it means patient has diabetes (positive with 29%) else if (NO) patient DiabetesPedigreeFunction >=0.06 it means the patient diabetes is (Negative with 6%) but if (YES) patient diabetes is (Positive with 23%) else if (NO) patient BMI >=0.51 it means patient diabetes is (Positive with 20%) else if (YES) patient diabetes is (Negative with 2%) but if (YES) patient BloodPresure < 0.62 it means patient diabetes is (Positive with 17%) else if (NO) patient diabetes is (negative with 4%).

SUMMARY ON R

This is summary on how the experiment was done using the RStudio. The Basic Statistics of the dataset (diabetes) was show in figure 3 above listing all the roll counts and their various names and value while figure 4 shows the data structure of the dataset and their variables. The dataset was loaded into the

environment after which exploring of the dataset was done shown in figure 6. A plot of observations was done which shows various frequencies of all the variables shown in figure 5. Then there is need to check for missing values in each variable after which it was observed that there are excessive missing values in variable (skinthickness and insulin) which were removed and replaced with mean of each variable. So as to ensure that there is no missing decimal values in the dataset, conversion needs to be done where numeric values where converted to integer values. Because the data set needs to be transformed to 0 and 1 so as to enable easy scaling, hence normalization of the dataset was done by using the function(x). The Gini model was used in creating the model but first, the dataset was Split into two with the percentage of (75% = training and 25%=testing) respectively then a decision tree algorithm was applied on the training dataset of (75%), note: The GINI Model was adopted because of it presented a more clear model and hence make it easier for understanding of result when used to predict if patients has diabetes or not shown in figure 7 above after the model has been achieved, then Generating of the Frequency Table which will create tabular results of categorical variable of positive or negative throughput after testing on diabetes dataset and model. Then after the prediction result by the computer is on the dataset and model is made, there is need for more accuracy of the prediction on both dataset and model by applying a confusion matrix to guarantee the accuracy on the rate at which the model and predicts the dataset. To achieve a perfect result, a calculation of the confusion matrix is done by adding the dataset prediction positive value with negative value from the model then divide with the total number of the dataset. 142 +338/583 = 0.82345, this result shows that the applied testing on the trained (75%) dataset using the model is 82.3 accurate while the classification error is 0.176. Then testing the model on the test dataset (25%), then the model is rerun again on the test dataset, and it shows the prediction, then a confusion matrix is calculated using same formula 34 +102/185 = 0.7351 which is 74% accurate on test dataset.

In summary, R was used to create a model where the diabetes dataset was split into two. The model was applied on train dataset and test dataset and accuracy was determined on both before prediction of positive or negative patients on diabetic disease are achieved.

CONCLUSION

As earlier stated, that the aim of this work is to adopt Machine Learning and Data Mining tools in the identification and prediction of Diabetes Patients Using Classification Mining Algorithm. This research was able to show clearly how diabetes disease could be managed using prediction models. These models were able to predict the status of a patient's diabetes state efficiently and accurately.

RECOMMENDATION

The researcher therefore recommends the following:

1. Full adoption of machine learning tools should be used in solving real life challenging problems more especially in health related problems more especially in Nigeria.
2. Different models showed be employed and then compared against each other for accurate data prediction.
3. Other organizations should be encouraged to apply machine learning tools for easy decision making.
4. Other researches can be done in the area of predicting if a patient with diabetes can die within a specific date because of high increase in glucose in the system or not using other modeling tools like entropy.

REFERENCES

[1] Pravarti Jain And Santosh Kr Vishwakarma (2017) A Case Study on Car

[2] Evaluation and Prediction: Comparative Analysis using Data Mining Models, International Journal of Computer Applications (0975 – 8887) Volume 172 – No.9

[3] Diabetes UK, Understanding Diabetes-Your key to better health (2003), Prediction of Diabetes using Classification algorithm accessed from http://www.diabetes.or.uk/infocenter/pubs/understand.doc, Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K., (2016). Performance Analysis of

[4] DataMiningClassificationTechniquestoPredictDiabetes.ProcediaComputerScience82,115–121.doi: 10.1016/j.procs.2016.04.016.

[5] Sharief, A.A., Sheta, A., (2014). DevelopingaMathematicalModeltoDetectDiabetesUsingMultigeneGeneticProgramming.InternationalJournalofAdvancedResearchinArtificialIntelligence(IJARAI)3,54–59.doi: doi:10.14569/IJARAI.2014.031007.

[6] Pradhan, P.M.A., Bamnote, G.R., Tribhuvan., Jadhav, Chabukswar., Dhobale,V., (2012)

[7] A Genetic Programming Approach for Detection of Diabetes. International Journal of Computational Engineering Research2, 91–94.

[8] Tarik A. Rashid, S.M.A., Abdullah, R.M., Abstract, (2016). An Intelligent Approach for Diabetes Classification, Prediction and Description. Advances in Intelligent Systems and Computing 424,323–335.doi:10.1007/978-3-319-28031-8.

[9] Nai Arun, N., Moungmai, R., (2015). ComparisonofClassifiersfortheRiskofDiabetesPrediction.ProcediaComputerScience69,132 142.doi: 10.1016/j.procs.2015.10.014.

[10] Nai Arun, N., Sittidech, P., (2014). Ensemble Learning Model for Diabetes Classification

[11] Advanced Materials Research 931-932,1427–1431.doi: 10.4028/www.scientific.net/AMR.931-932.1427.

[12] Orabi, K.M., Kamal, Y.M., Rabah, T.M., (2016). Early Predictive System for Diabetes Mellitus Disease, in: Industrial Conference on Data Mining, Springer.Springer.pp.420–427.

[13] Priyam, A., Gupta, R., Rathee, A., Srivastava, S., (2013). Comparative Analysis of Decision Tree Classification Algorithms. International Journal of Current Engineering and TechnologyVol.3,334–337.doi: JUNE 2013, arXiv: ISSN2277-4106.

[14] Bamnote, M.P., G.R., (2014). Design of Classifier for Detection of Diabetes

[15] Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1, 763–770. doi:10.1007/978-3-319-11933-5.

[16] Esposito, F., Malerba, D., Semeraro, G., Kay, J., (1997). A comparative analysis of

[17] methods for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 476–491. doi:10.1109/34.589207.

[18] BITS Pilani, Dubai, BITS Pilani, Dubai and Ronak Sumbaly (2015) Diagnosis of Diabetes Using Classification Mining Techniques accessed from International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1)

[19] Jiawei Han and Micheline Kamber, (2001) "Data Mining Concepts and Techniques", Morgan Kauffman Publishers.