International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-10, August 2019

Machine Learning Techniques for Prediction of Parkinson's Disease using Big Data

S. Kanagaraj, M.S. Hema, M. Nageswara Gupta

Abstract*: The growth of data in the healthcare industry grows exponentially and the annual growth rate is about 40%, managing this amount of data is challenging task. Big Data architecture and frameworks affords the platform for data storage and processing of massive volume of data in healthcare industry. The paper aims to provide Big Data technologies and Machine Learning algorithms to predict Parkinson's Disease (PD). The dataset from PPMI are used in the current study and observe the progression of the Parkinson's Disease. The Movement Disorder Society-Unified Parkinson's Disease (MDS-UPDRS) features are used for the prediction model. The current study focuses on machine learning algorithms from python libraries such as pandas, ski-kit learn, numpy and matplotlib. The important features obtained are tremor, bradykinesia, facial expression is observed as important features for classification. It is observed that logistic regression and multi class classifier performed with accuracy of 99.04% than the other algorithms such as Naïve Bayes, k-Nearest Neighbor, SVM and Neural Network.

Keywords* : Parkinson's Disease, Big Data, Machine Learning, Python, Jupyter Notebook, HDFS, Pandas, Numpy, Sci-kit learn, Seaborn, Matplotlib, Classification.

I. INTRODUCTION

Causes for the second most common neurogenerative disease is unknown and the common causes includes non-genetic factors. The differentiation between the Alzheimer's Disease (AD) and Parkinson's Disease (PD) is the primary identification process, since both AD and PD shares the common features in the early stage [1]. Memory disorder is the main cause of Alzheimer's disease and the movement disorder leads to Parkinson's, where both the diseases are caused by damage in the brain cells which are progressive in nature and leads severe other difficulties of normal life [2] . Modelling tools to study the disease and provide prediction of symptoms, new therapies and control using early diagnosis. Some of the authors suggested quantitative studies for understanding the disease and the other side include models to identify the symptoms [3]. Unprecedented data from various sources such as government, industry, health, social networks and financial sources offers for large data volumes [4]. The smart devices, internet of things with cloud computing and many other technological trends sources for the trends of Big Data revolution. Interesting and important information to be extracted from the available sources and the traditional technologies lacks behind with capability to provide flexibility, scalability and performance issues. The Big Data framework must provide an environment in which many of parameters such as performance, reliability, cost, technological compatibility, support and security are considered, to choose the finest technology for the Big Data Analytics. The potential of Big Data analytics is realized once the challenges are overwhelmed with design of upright data pre-processing system methods and models, also must include management challenges such as privacy, security, governance and ethical aspects [5].

The Big Data framework must build to perform data analytics and decision-making process that makes the managerial decisions to be faster and reduces the jeopardies. [6] The author presented a framework which has a three-layer phases (intelligence, choice and implementation). The decision making in an organization to be spontaneous for the better results and provides advancement to the scientific and technological needs. [7] The survey for Big Data Technologies are carried out to meet the needs of the specific applications requirements. [8] The global view of the Big Data Technology is presented for dynamic visualization of the needs of the end user. The stakeholders must be able to identify the patterns, trends and correlations of the data which are of high- dimensions and the attributes must be analyzed and visualized. The feature selection process must reduce the redundancy and maximize the relevance to the target class labels for classification [9]. Prediction of disease at an earlier stage and precise medicine provides the wellbeing of the humans. [10] The data from multiple health bodies constitute electronic health records (EHR) which gives health practitioners to engage with more of huge volume of information, such difficulties are overwhelmed with the help of proper analytical tools. Priorities are wisely used to implement an analytical platform for the current context and challenges. The need for Big Data in health care used to improve the quality by providing patient centric services, detection of spreading diseases at an earlier stage, improving treatment methods and to screen hospitals quality [11]. Framework for application specific would provide analytical capabilities [12] for data from electronic health records. The tools in the framework to offer analytical avenues for patient centric healthcare system.

Manuscript published on 30 August 2019.
*****Correspondence Author(s)
S. Kanagaraj, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India, kanagaraj.s.it@kct.ac.in
Dr. M.S. Hema, Department of Computer Science and Engineering, Aurora's Scientific Technological and Research Academy, Hyderabad, India,ghema_shri@yahoo.co.in
Dr. M. Nageswara Gupta, Department of Computer Science and Engineering, Sri Venkateshwara College of Engineering, Bengaluru, India, mnguptha@yahoo.com

Big Data and Parkinson's

The integration of Parkinson's heterogenous data from multiple sources and which are of complex in nature offer opportunities to study the early stages of the neurogenerative of patients, track the progression and quickly provide treatment solutions. The characteristics of Big Data includes large data sets with the heterogenous formats of structured, unstructured and semi-structured data. The complex nature of Big Data requires technologies that are influential and algorithms which are of cutting-edge. The three main characteristics of Big Data called the 3V's include Volume, Velocity and Variety.

(i) Volume: The volume of data generated was mind blowing, since 2.5 quintillion bytes of data through various sources of data which includes smart phones, social networks, sensors, logs and ICTs. About 90% of data generated to the overall available data was last two years. The internet, digital mobiles and internet of things plays important role to drive the voluminous data.

(ii) Velocity: Velocity refers to the speed of data generated over the various devices. The Big Data technologies helps to accept the data explosion and process the data to extract useful information without any bottlenecks. For example, YouTube data videos are processed with prodigious speed to satisfy all the needs of the user.

(iii) Variety: The Big Data are characterized by structured, unstructured and semi structured. The data includes text, images, audio and video. The various formats to be processes and presented in a well needed format.

II. PROBLEM STATEMENT

The goal of this work is to study prediction Parkinson's disease at an early stage from the formerly available public database and find the potential biomarkers for the cause of the disease and find the methods to cure the disease for the pretentious persons.

The research work is to focus on to provide outcome of different queries involved such as,

(i) Identify the different potential biomarkers involved for cause of disease by using Big Data framework and technologies.

(ii) Analyze the different data needs to discover the biomarkers that cause the PD.

(iii) Visualization method that enables the physicians to be easily find the attributes involved in cause of PD.

(iv) Use of machine learning algorithms to find the suitable process and algorithm that guide physician for appropriate clinical decision making.

III. RELATED WORK

Parkinson's Disease are identified with various measures which indicates the stages and severity of the disease. The scales are measure of impairment and disability of the patients. The Unified Parkinson's Disease Rating Scale (UPDRS) and the Hoehn and Yahr (HY) scale are the most commonly used scales to evaluate the severity of Parkinson's [13]. UPDRS provides a comprehensive assessment of infirmity and diminishing by evaluating the most related clinical features of PD [14], whereas the HY scale provides a gross assessment of disease progression through the stages from 0 (no sign of disease) to 5 (severe) [15]. The Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MD-UPDRS) is the revised version of the original UPDRS with several new properties related to non-motor elements of PD and hence it is more comprehensive than the original UPDRS. The properties included are refinement of scoring instructions, emphasis on impairment and disabilities related to minor symptoms and signs of PD. The scales are classified into four parts with 65 items. Part 1 contains 15 items which contains first 7 items of non-motor experience of daily living and 8 items as patient questionnaire. Part II includes 14 items as patient questionnaire which concerns about motor experience of daily living. Part III contains motor examination which includes 34 items and those are examined by specialist. Part IV contains 6 items uses to assess the motor complications like dyskinesias and motor fluctuations related to duration of disease, levodopa dose and duration of levodopa treatment [16].

The original UPDRS have been used as standard to PD evaluation, due to increase in scores over period makes crucial for clinical decision making and does not include non-motor aspects of PD [17,18]. The MDS-UPDRS is more sensitive to changes in scores than the original UPDRS. Wide spectrum of assessment, reliable and subtle instruments are used for estimation of progression and severity in PD.

Motivation and scope of study

The HY scale used to categorize PD into several stages with stage 1 -5 and are grouped as stage 1 & 2 with early stage, stage 3 as moderate and stage 4 & 5 as late stage. Relation between UPDRS and HY done by Scanlon et al. [18, 19] for UPDRS Part III scores and was optimized [19] used genetic algorithm (GA) to refine the parameters. The formula used had two shortcomings with intuitive rules and only Part III items were used instead of all the items in the spectrum.

In this work the MDS-UPDRS, HY scale and machine learning algorithms are studied to analyze the performance of algorithms on PPMI database.

IV. METHODOLOGY

A. The Data

The Data used for work was collected from Parkinson's Progression Marker Initiative (PPMI) funded by The Michael J. Fox foundation for Parkinson's Research (MJFF) to identify the biomarker's involved in PD progression and to develop a better treatment method. The open access clinical and imaging data can be accessed with rights and user agreements.

B. About the Database

The database consists of 161 files and data were available in the csv format. The files are categorized as study docs, subject characteristics, biospecimen, curated data cuts, enrolment, imaging, internal, medical history, motor assessments, non-motor assessments and remote data collection.

C. Data pre-processing

The data source of PPMI which includes MDS-UPDRS which includes the severity of PD and scores are recorded with the dimension between 0-4. The pre-processing of data includes calculating the total scores of MDS-UPDRS.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The important aspect of Machine learning is to use and benefit from the build model. A sequence of steps is involved in data processing components is called a data pipeline. The PPMI data set of motor MDS-UPDRS with total of 313 attributes are used for predictions of PD. The 715 number of instances are used in this machine learning process which are of numerical and data files are of csv format. If there are missing features with values most of the machine learning algorithms cannot work appropriately. To overcome the above issues the attributes with missing values are to be fixed. The three types of process involved are (i) get rid of the corresponding rows, (ii) get rid of the whole attribute and (iii) set values to zero, mean, median, mode, etc. Option 3 is of good choose to fill the missing values and in Scikit-Learn the Imputer class takes care of the missing value with the median value.

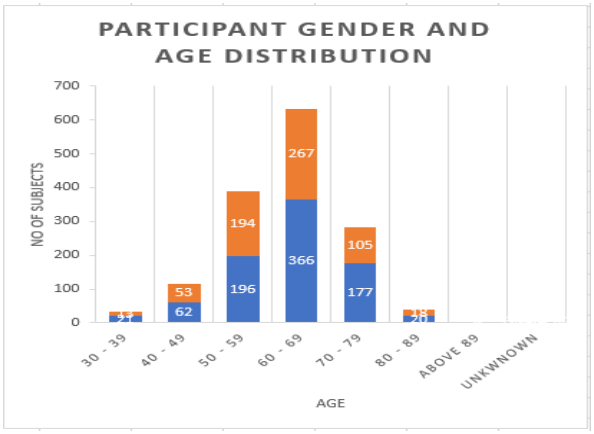

Fig. 1. PD Patient age and gender distribution

D. Details of files in PPMI database

Table-I Details of files in PPMI database

S. No	File Details	No of files
1	Total no of csv files	161
2	Administrative data files	41
3	Clinical and questionnaire files	120

E. Attributes details of PPMI database

Table-II Attribute details of PPMI database

S. No	Attribute Details	No of attributes
1	Total attributes	3281
2	No of unique attributes	1726
3	Numerical attributes	968
4	Categorical attributes	1571
5	Date	564
6	Time	52
7	Attribute without type	126

The total number of records includes 715 patients, out of which 465 are male and 250 are female patients. Healthy control includes 213 and the Parkinson's Disease patients are 421 in count and SWEDD (Scans without evidence of dopaminergic deficit) are 81 in number.

F. Patient details

S. No	Attribute Details	No of attributes
1	Healthy Control - HC	213
2	Parkinson's Disease – PD	421
3	Scans without evidence of dopaminergic deficit - SWEDD	81

Table-III Patient details

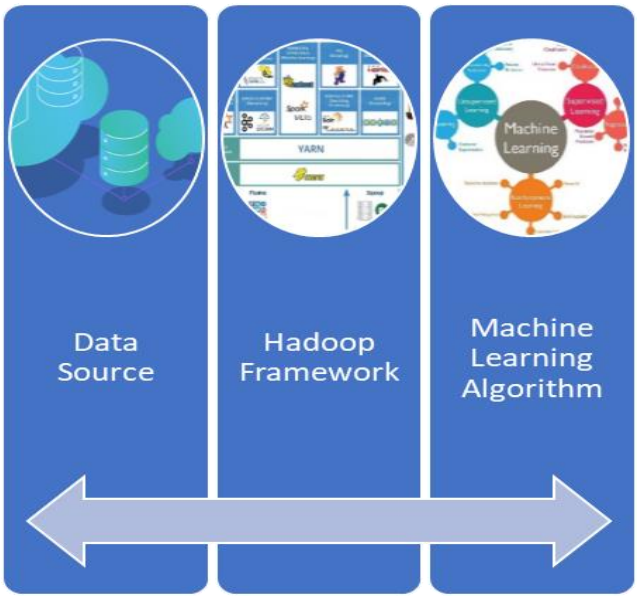

Fig. 2. Data prediction flow process

Fig. 3. The Machine Learning Process

Hardware & Software details

The hardware used for the work consists of Intel i7 processor with 32 GB RAM. The machines are connected as clusters with one master and three slave nodes. The software packages include Anaconda 3 with packages of Pandas, Numpy, Sci-kit learn, Seaborn and Matplotlib.

V. EXPERIMENT, RESULTS AND DISCUSSION

The experiment is carried out in Hadoop framework environment to store, process and analyze the data set. The data from the PPMI repository is injected to HDFS storage and required items are used in Jupyter Notebook for analytics. The total instances are divided into 80-20 with 572 for training the data sets and 115 instances for test data sets. The non-linear machine learning algorithms such as Naïve Bayes, Decision Tree, k-Nearest Neighbour,

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Support Vector Machines (SVM) and Neural Network are used for prediction of PD. The ensemble machine learning algorithm Random Forest was used for prediction process for more robust predictions with multiple models. The classification accuracy for above algorithms are compared to show the performance of machine learning algorithms.

Table-III Comparison cluster algorithms with performance measure

Algorithm	Percentage of correctly classified instances	TP Rate	FP Rate	Precision	F-Measure
Naïve Bayes	93.28	0.933	0.018	0.950	0.936
k-Nearest Neighbour	95.14	0.951	0.158	0.951	0.50
Support Vector Machines	97.80	0.978	0.039	0.978	0.978
Neural Network	98.99	0.990	0.019	0.990	0.990
Multiclass Classifier	**99.04**	**0.990**	**0.021**	**0.990**	**0.990**
Logistic Regression	**99.04**	**0.990**	**0.021**	**0.990**	**0.990**

VI. CONCLUSION

In this paper, the comparative study of various machine learning algorithms is carried out. For analysis and prediction of Parkinson's PPMI data sets and six different classification algorithms are used. The results show that the multiclass classifier and logistic regression better performed than the other algorithms for the data sets. In future, more number of biomarker features are to be included for the prediction of progression of PD. Advanced stages of PD are estimated based on the visits from baseline t0 visit 12. The rating scales such as MDS-UPDRS and HY Scale with classifiers will provide effective method for estimation of severity of PD.

VII. ACKNOWLEDGMENT

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up -to-date information on the study, visitwww.ppmi-info.org.

PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including [list the full names of all of the PPMI funding partners found at www.ppmi-info.org/fundingpartners].

REFERENCES

1. Gunjan Pahuja and TN Nagabhushan "Statistical Approach towards Parkinson ' s Disease Progression," vol. 3, no. 2,2016.
2. L. V Kalia, A. E. Lang, and G. Shulman, "Parkinson ' s disease," *Lancet*, vol. 386, no. 9996, pp. 896–912, 2015.
3. Y. Sarbaz and H. Pourakbari, "A review of presented mathematical models in Parkinson's disease: black- and gray-box models," *Med. Biol. Eng. Comput.*, vol. 54, no. 6, pp. 855–868, 2016.
4. A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, "Big Data technologies: A survey," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 30, no. 4, pp. 431–448, 2018.
5. U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods," *J. Bus. Res.*, vol. 70, pp. 263–286, 2017.
6. N. Elgendy and A. Elragal, "Big Data Analytics in Support of the Decision Making Process," *Procedia Comput. Sci.*, vol. 100, pp. 1071–1084, 2016.
7. A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, 2015.
8. A. Genender-Feltheimer, "Visualizing High Dimensional and Big Data," *Procedia Comput. Sci.*, vol. 140, pp. 112–121, 2018.
9. D. R. Leff and G.-Z. Yang, "Big Data for Precision Medicine," *Engineering*, vol. 1, no. 3, pp. 277–279, 2015.
10. F. Khennou, Y. I. Khamlichi, and N. E. H. Chaoui, "Improving the use of big data analytics within electronic health records: A case study baseD OpenEHR," *Procedia Comput. Sci.*, vol. 127, pp. 60–68, 2018.
11. J. Archenaa and E. A. M. Anita, "A survey of big data analytics in healthcare and government," *Procedia Comput. Sci.*, vol. 50, pp. 408–413, 2015.
12. V. Palanisamy and R. Thirunavukarasu, "Implications of big data analytics in developing healthcare frameworks - A review," *J. King Saud Univ. - Comput. Inf. Sci.*, 2017.
13. I. D. Dinov *et al.*, "Predictive big data analytics: A study of Parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations," *PLoS One*, vol. 11, no. 8, pp. 1–28, 2016.
14. S. Grover, S. Bhartia, Akshama, A. Yadav, and K. R. Seeja, "Predicting Severity of Parkinson's Disease Using Deep Learning," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1788–1794, 2018.
15. V. Anantharam, A. Kanthasamy, A. A Willette, S. Nilakanta, M. SenthilarumugamVeilukandammal, and B. Ganapathysubramanian, "Big Data and Parkinson's Disease: Exploration, Analyses, and Data Challenges.," *Proc. 51st Hawaii Int. Conf. Syst. Sci.*, vol. 9, pp. 2778–2783, 2018.
16. A. Blochberger and S. Jones, "Parkinson's disease clinical features and diagnosis," *Clin. Pharm.*, vol. 3, no. 11, pp. 361–366, 2011.
17. D. Nyenhuis *et al.*, "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results," *Mov. Disord.*, vol. 23, no. 15, pp. 2129–2170, 2008.
18. A. Schrag and N. Quinn, "Dyskinesias and motor fluctuations in Parkinson's disease," *Brain*, vol. 123, no. 11, pp. 2297–2305, 2002.
19. S. Kanagaraj, M.S. Hema, M. Nageswara Gupta, "Environmental Risk Factors and Parkinson‟s Disease – A Study Report", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-4S2, December 2018

AUTHORS PROFILE

S. Kanagaraj is a Assistant Professor in the department of Information Technology at Kumaraguru College of Technology, Coimbatore. He received B.E(CSE) degree from Periyar University, M.E (CSE) degree in Computer Science and Engineering degree from Anna University, Chennai. His research areas are health care analytics, machine learning and data science. He is a member of IEEE professional society and college student branch counsellor since 2015. He was co-principal investigator for funded project from IEEE foundation grants program for the project titled "Smart agriculture for sustainable food production" and "Rehabilitation for sustainable future of Delta farmers". He has published around 10 research paper in international, national and conferences.

Dr. M. S. Hema is a Professor in the department of Computer Science and Engineering at Aurora's Scientific, Technological an Research Academy, Hyderabad. She received B.E(CSE) degree from Bharathiar university, M.E (CSE) and Ph.D degree in Computer Science and Engineering degree from Anna University, Chennai. Her research areas are health care analytics, machine learning and data science. She received young Scientist award from Vision Group of Science and Technology (VGST), Karnataka in the academic year 2016-17. She has published around 40 research paper in national and international journals and conferences respectively. She has permanent membership in ISTE and ACCS.

Dr. M. Nageswara Guptha is a Professor in the department of Computer Science and Engineering and vice principal at Sri Venkateshwara College of Engineering, Bangalore. He received B.E(CSE) degree from Bharathiar university, M.E (CSE) and Ph.D degree in Computer Science and Engineering degree from Anna University, Chennai. Her research areas are service oriented architecture, ICT solution and Data Mining. He has published patents and around 40 research paper in national and international journals and conferences respectively. He has permanent membership in ISTE and ACCS.