Parkinson Disease Detection: Using XGBoost Algorithm to Detect Early Onset Parkinson Disease

Nasif Wasek Fahim
Computer Science and Engineering
East West University
Dhaka, Bangladesh
2017-1-60-037@std.ewubd.edu

Samik Ahmed Eshti
Computer Science and Engineering
East West University
Dhaka, Bangladesh
2017-1-60-058@std.ewubd.edu

Khadiza Akter Nura
Computer Science and Engineering
East West University
Dhaka, Bangladesh
2017-1-60-060@std.ewubd.edu

Md. Jubayer Hossain Abir
Computer Science and Engineering
East West University
Dhaka, Bangladesh
2017-1-60-084@std.ewubd.edu

Mueem Nahid Ibn Mahbub
Computer Science and Engineering
East West University
Dhaka, Bangladesh
2017-1-60-089@std.ewubd.edu

Abstract: **Parkinson's disease is caused by the disruption of the brain cells that produce a substance to allow brain cells to communicate with each other, called dopamine. The cells that produce dopamine in the brain are responsible for the control, adaptation, and fluency of movements. When 60–80% of these cells are lost, then enough dopamine is not produced and Parkinson's motor symptoms appear. It is thought that the disease begins many years before the motor (movement-related) symptoms and therefore, researchers are looking for ways to recognize the non-motor symptoms that appear early in the disease as early as possible, thereby halting the progression of the disease. In this paper, machine learning-based diagnosis of Parkinson's disease is presented. Python machine learning project, using the python libraries scikit-learn, NumPy, pandas, and xgboost, we will build a model using an XGBClassifier. We'll load the data, get the features and labels, scale the features, then split the dataset, build an XGBClassifier, and then calculate the accuracy of our model. 94.87% accuracy was achieved with the least number of voice features for Parkinson's diagnosis.**

Keywords: Parkinson's disease, XGBoost, Scalable, tree, boosting

I. Introduction

Parkinson's disease is a progressive disorder that is caused by the degeneration of nerve cells in the part of the brain called the substantia nigra, which controls movement[1]. These nerve cells die or become impaired, losing the ability to produce an important chemical called dopamine. The symptoms of Parkinson's develop in patients with an 80 percent or greater loss of dopamine-producing cells in the substantia nigra. Normally, dopamine operates in a delicate balance with other neurotransmitters to help coordinate the millions of nerve and muscle cells involved in the movement. Without enough dopamine, this balance is disrupted, resulting in tremor (trembling in the hands, arms, legs, and jaw), rigidity (stiffness of the limbs), slowness of movement, and impaired balance and coordination the hallmark symptoms of Parkinson's[2]. Symptoms generally develop slowly over years. The progression of symptoms is often a bit different from one person to another due to the diversity of the disease. People with Parkinson's Disease may experience Tremor, mainly at rest, and described as a pill-rolling tremor in hands. Other forms of tremor are possible. Also, they may experience Bradykinesia, Limb rigidity Gait, and balance problems[3]. The cause of Parkinson's essentially remains unknown. However, theories involving oxidative damage, environmental toxins, genetic factors, and accelerated aging have been discussed as potential causes for the disease. In 2005, researchers discovered a single mutation in a Parkinson's disease gene (first identified in 1997), which is believed responsible for five percent of inherited cases. Parkinson's disease signs and symptoms can be different for everyone. Early signs may be mild and go unnoticed. Symptoms often begin on one side of your body and usually remain worse on that side, even after symptoms begin to affect both sides. Sometimes it is quite difficult to detect whether there is Parkinson's Disease is present in the patient's body. Our goal is to develop a model well enough so that it can detect the early stage of Parkinson's Disease. In this Python Machine learning project, we will build a model using XGBoost, which we can detect the presence of Parkinson's disease in one's body. XGBoost is a new Machine Learning algorithm designed with speed and performance in mind. XGBoost stands for extreme Gradient Boosting is based on decision trees[4]. In this project using the Python libraries such as numpy, scikit learn, XGBoost we built a model using XGBClassifier. First, we will load the data, get the features and labels, scale the features then split the dataset, built an XGBclassifier, and then calculate the accuracy of our model.

II. Related Work

Parkinson's disease is a universal public health problem. This disease affects about 1% of the world population over the age of 55 [5]. After the invention of machine learning techniques, these kinds of diseases are easy to detect. Using ML techniques in the medical field, diagnosis, and outcome prediction are being benefited [6]. Many works have been done also on this Parkinson disease. Some worked on to present a comprehensive review of the prediction of PD by using ML-based approaches [7]. Some suggested an expert PD diagnosis system based on a genetic algorithm [8]. Others analyzed the possibilities of diagnosing a PD affected subjects from a normal subject [9]. Functional neuroimaging of ML holds the promise of improved diagnosis and allows assessment in early disease [10]. The NNs and adaptive neuro-fuzzy classifier with linguistic hedges are investigated for automatic diagnosis of PD [11]. The performance of the probabilistic neural network (PNN) for automatic diagnosis of PD is evaluated [12]. SMV classifier is also investigated for PD [13]. Daryl Chang suggested the importance of ML in increasing the efficiency and accuracy of diagnosing the PD [14]. Sachin Shetty has successfully classified the PD from another neurodegenerative disease by applying a Gaussian RBF based kernel with an SMV classifier [15]. Shyam V.Perumal has used wearable sensors to develop a gait monitoring system for patients with PD [16]. Frenkel Toledo used stride-to-stride variability of gait timing to diagnose PD [17].

III. Methodology

For this project, we used different python libraries such as sciKit-learn, NumPy, pandas, and xgboost to build a model by using XGBClassifier. We completed this project following these steps and going with this flow:

1. We first loaded the required data
2. We got all the features and labels
3. We scaled the features
4. After that split the dataset
5. We built an XGBClassifier
6. Finally, we calculated the accuracy of the model

A. XGBoost

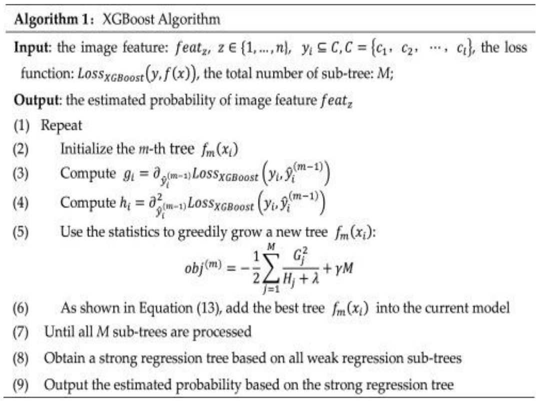
To solve the Parkinson's disease detection problem, we used the XGBoost algorithm. It is very useful if the data set is large. It is a perfect combination of software and hardware optimization techniques to yield superior results using less computing resources in the shortest amount of time. It is basically an extreme gradient boosting method introduced by data scientist Tianqi Chen in 2014. Since then this algorithm has gained big popularity and compared to other machine learning techniques it is quite simple. It’s accessibility and advanced features make it a versatile algorithm for Windows, Linux, and OS X. To implement XGBoosting gradient boosting is required. This is a machine learning algorithm for regression, classification, and ranking. Gradient boosting requires a loss function (a function that maps an event or values of one or more variables onto a real number), a weak learner, and an additive model (a nonparametric regression method) to calculate loss and modify decision trees to minimize the error. The XGBoost algorithm supercharges gradient boosting tasks. When applied, this open-source software library untangles regression, classification, and ranking issues. It’s very fast, accurate, and accessible, so it’s no wonder that is has been adopted by numerous companies, from Google to start-ups. XGBoost is designed within the framework of the decision tree algorithm of gradient boosting. It is scalable. The algorithm can produce billions of outcomes quickly. Along with the tree learning algorithms, XGBoost can solve linear models, too. That means it can handle parallel computations on one machine.

B. XGBoost Accessibility

The XGBoost software library is accessible from many common interfaces, including, but not limited to:

- Command Line Interface
- C++
- Scikit-learn / Python
- Caret package / R
- Julia
- Java and JVM languages
- Apache Spark, Hadoop, and Flink

C. XGBoost Algorithm



```
Algorithm 1: XGBoost Algorithm
Input: the image feature: $feat_z$, $z \in \{1, \dots, n\}$, $y_i \subseteq C, C = \{c_1, c_2, \cdots, c_l\}$, the loss
  function: $Loss_{XGBoost}(y, f(x))$, the total number of sub-tree: $M$;
Output: the estimated probability of image feature $feat_z$
(1) Repeat
(2)     Initialize the $m$-th tree $f_m(x_i)$
(3)     Compute $g_i = \partial_{\hat{y}_i^{(m-1)}} Loss_{XGBoost}\left(y_i, \hat{y}_i^{(m-1)}\right)$
(4)     Compute $h_i = \partial^2_{\hat{y}_i^{(m-1)}} Loss_{XGBoost}\left(y_i, \hat{y}_i^{(m-1)}\right)$
(5)     Use the statistics to greedily grow a new tree $f_m(x_i)$:
            $obj^{(m)} = -\frac{1}{2}\sum_{j=1}^{M}\frac{G_j^2}{H_j + \lambda} + \gamma M$
(6)     As shown in Equation (13), add the best tree $f_m(x_i)$ into the current model
(7) Until all $M$ sub-trees are processed
(8) Obtain a strong regression tree based on all weak regression sub-trees
(9) Output the estimated probability based on the strong regression tree
```

D. XGBoost Time Complexity

Training with XGBoost takes O(tdx***log***n)**,** where **t** is the number of trees, **d** is the height of the trees, and **x** is the number of non-missing entries in the training data. Prediction for a new sample takes O(td**).**

E. Secret Behind XGBoost Great Performance

XGBoost and Gradient Boosting Machines (GBMs) are both ensemble tree methods that apply the principle of boosting weak learners using gradient descent architecture. However, XGBoost improves upon the base GBM framework through systems optimization and algorithmic enhancements.

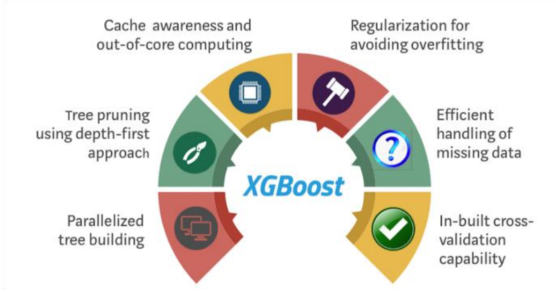

Cache awareness and out-of-core computing
Regularization for avoiding overfitting
Tree pruning using depth-first approach
Efficient handling of missing data
XGBoost
Parallelized tree building
In-built cross-validation capability

Fig. Beneficial Feature of XGBoost

F. Unique Features of XGBoost

- **Regularization:** XGBoost has an option to penalize complex models through both L1 and L2 regularization. Regularization helps in preventing overfitting.
- **Handling sparse data:** Missing values or data processing steps like one-hot encoding make data sparse. XGBoost incorporates a sparsity-aware split finding algorithm to handle different types of sparsity patterns in the data.
- **Weighted quantile sketch:** Most existing tree-based algorithms can find the split points when the data points are of equal weights (using quantile sketch algorithm). However, they are not equipped to handle weighted data. XGBoost has a distributed weighted quantile sketch algorithm to effectively handle weighted data.
- **Block structure for parallel learning:** For faster computing, XGBoost can make use of multiple cores on the CPU. This is possible because of a block structure in its system design. Data is sorted and stored in in-memory units called blocks. Unlike other algorithms, this enables the data layout to be reused by subsequent iterations, instead of computing it again.
- **Cache awareness:** In XGBoost, non-continuous memory access is required to get the gradient statistics by row index. Hence, XGBoost has been designed to make optimal use of hardware. This is done by allocating internal buffers in each thread, where the gradient statistics can be stored.
- **Out-of-core computing:** This feature optimizes the available disk space and maximizes its usage when handling huge datasets that do not fit into memory.

IV. Result

In this project, we detected the presence of Parkinson's Disease in people using various factors. For this, we used an XGBClassifier and made use of the sciKit-learn library to prepare the dataset. Through this process, we achieved an accuracy of 94.87%, which is more than decent considering the number of lines of code in this project.

V. Analysis

We divided the dataset into 80 percent training data and 20 percent testing data. After then we initialized the XGBClassifier was used and the model had been trained under the Ensemble Learning group in ML. After that, we found out that,

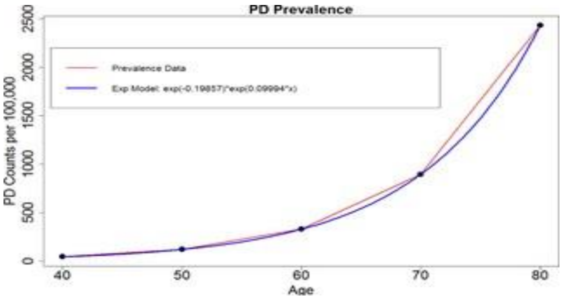


Fig. Increasing PD patient count with age

Finally, we generated y_pred (predicted values for x_test) and calculated the accuracy of our model. Comparing XGBooster with other algorithms the accuracy, precision, recall, etc. is very praiseworthy. XGBooster is not only able to keep up with all those other algorithms but exceeds them in performance which led us to pick this algorithm as the base of our project which deals with detecting Parkinson's disease.

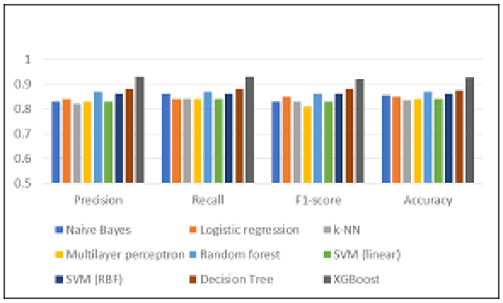


Fig. Comparison between algorithm

VI. Conclusion

For thousands of years, Parkinson's disease has been affecting the mankind. Humans who suffer from this disease suffer a lot from its effects were treated with varying results by a variety of plant-based treatments, some of which are still in use today. The concerning matter is that still there is still no diagnostic test to determine the disease. This paper presented the prediction of Parkinson's disease by using machine learning-based approaches. We have learned about XGBoost, a scalable tree boosting system, and provides state-of-the-art results. We can use this algorithm to solve other machine learning systems related problem as well. A sparsity aware algorithm is used to hand the sparse data and a justice weighted quantile sketch for approximate learning. XGBoost can solve real-world scale problems using a minimal amount of resources.

VII. References

[1] B. S. Carsten Lundby, Paul Robach, "Blood Doping -- Effects and Detection," *Br. J. Pharmacol.*, pp. 1–25, 2012, doi: 10.1111/j.1476.

[2] M. Ali *et al.*, "Parkinson ' s Disease Read the Story of a Patient Diagnosed with Disease," no. 2267, p. 60008, 2020.

[3] J. S. Hawley, "What is Parkinson ' s disease ?," *Park. Dis. Improv. Patient Care*, vol. 501, no. c, pp. 1–6, 2014.

[4] J. Browniee, "A Gentle Introduction to XGBoost for Applied Machine Learning," *Mach. Learn. Mastery*, pp. 1–20, 2016, [Online]. Available: https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/.

[5] R. Betarbet, T. B. Sherer, and J. T. Greenamyre, "Animal models of Parkinson's disease. [Review] [89 refs]," *Bioessays*, 2002.

[6] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," *BMC Med. Res. Methodol.*, vol. 19, no. 1, pp. 1–18, 2019, doi: 10.1186/s12874-019-0681-4.

[7] S. Bind, A. K. Tiwari, and A. K. Sahani, "A Survey of Machine Learning Based Approaches for Parkinson Disease Prediction," *Int. J. Comput. Sci. Inf. Technol.*, 2015.

[8] D. Avci and A. Dogantekin, "An Expert Diagnosis System for Parkinson Disease Based on Genetic Algorithm-Wavelet Kernel-Extreme Learning Machine," *Parkinsons. Dis.*, 2016, doi: 10.1155/2016/5264743.

[9] E. Abdulhay, N. Arunkumar, K. Narasimhan, E. Vellaiappan, and V. Venkatraman, "Gait and tremor investigation using machine learning techniques for the diagnosis of Parkinson disease," *Futur. Gener. Comput. Syst.*, 2018, doi: 10.1016/j.future.2018.02.009.

[10] P. Piccini and A. Whone, "Functional brain imaging in the differential diagnosis of Parkinson's disease," *Lancet Neurology*. 2004, doi: 10.1016/S1474-4422(04)00736-7.

[11] "AUTOMATIC RECOGNITION OF PARKINSON'S DISEASE FROM SUSTAINED PHONATION TESTS USING ANN AND ADAPTIVE NEURO-FUZZY CLASSIFIER," *Mühendislik Bilim. ve Tasarım Derg.*, 2010, doi: 10.21923/mbtd.12375.

[12] M. Ene, "Neural network-based approach to discriminate healthy people from those with Parkinson's disease," *Ann. Univ. Craiova, Math. Comp. Sci. Ser*, 2008.

[13] D. Gil and M. Johnsson, "Diagnosing Parkinson by using artificial neural networks and support vector machines," *Glob. J. Comput. Sci. Technol.*, 2009.

[14] D. Chang, M. Alban-hidalgo, and K. Hsu, "Diagnosing Parkinson ' s Disease From Gait," *Stanford*, 2015.

[15] S. Shetty and Y. S. Rao, "SVM based machine learning approach to identify Parkinson's disease using gait analysis," 2016, doi: 10.1109/INVENTIVE.2016.7824836.

[16] S. V. Perumal and R. Sankar, "Gait monitoring system for patients with Parkinson's disease using wearable sensors," 2016, doi: 10.1109/HIC.2016.7797687.

[17] F.-T. S., G. N., P. C., H. T., G. L., and H. J.M., "Effect of gait speed on gait rhythmicity in Parkinson's disease: Variability of stride time and swing time respond differently," *J. Neuroeng. Rehabil.*, 2005.