2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Depression Level Prediction in People with Parkinson's Disease during the COVID-19 Pandemic

Hashneet Kaur[1], Patrick Ka-Cheong Poon[1], Sophie Yuefei Wang[1], Diane Myung-kyung Woodbridge[2]

***Abstract*— Many recent studies show that the COVID-19 pandemic has been severely affecting the mental wellness of people with Parkinson's disease. In this study, we propose a machine learning-based approach to predict the level of anxiety and depression among participants with Parkinson's disease using surveys conducted before and during the pandemic in order to provide timely intervention. The proposed method successfully predicts one's depression level using automated machine learning with a root mean square error (RMSE) of 2.841. In addition, we performed model importance and feature importance analysis to reduce the number of features from 5,308 to 4 for maximizing the survey completion rate while minimizing the RMSE and computational complexity.**

***Index Terms*— predictive models, distributed computing, big data applications, COVID-19, AutoML**

I. Introduction

The ongoing COVID-19 pandemic has caused not only physical suffering but also threatened mental health for people across the world [1]. The World Health Organization (WHO) also expressed its concern that worldwide social distancing and isolation requirements might cause an increase in loneliness, anxiety, depression, insomnia, and self-harm or suicidal behavior [1]. This is particularly the case for people who already have chronic illnesses like Parkinson's disease (PD) [2]. First, depression is already a primary concern of people with PD. At any given time, 20 - 40% of people with PD suffer from depression, several times the prevalence in the general population [3]. Second, regular clinical services for PD have been postponed or suspended, causing confusion and anxiety in the PD community [4].

Recent research confirms worsened mental health among the PD population as a global phenomenon. Van der Heide et al. [2] conducted an online survey on over 400 participants with PD in Netherlands and concluded from the responses that COVID-19 increased psychological distress and reduced physical activity among the PD population. In Japan, Kitani-Morii et al. [5] conducted a cross-sectional, hospital-based survey and a phone interview with participants with PD and their family members about neuropsychiatric symptoms. They performed univariate logistic regression analysis on participants and their family members as a control group. The authors found that significantly more participants with PD presented with severe clinical depression compared to the control group. Janiri et al. [6] performed telephone interviews on people with PD at Gemelli University Hospital in Italy and concluded people with PD may be more vulnerable during COVID-19 and might require interventions to reduce irritability and mood instability. The aforementioned research used surveys and/or statistical analysis to confirm the negative impact of COVID-19 on the mental health of people with PD. However, it has not yet been investigated to predict worsened mental health by using data collected before and during the COVID-19 pandemic.

The primary objective of our study is to use a machine learning-based approach to predict an individual's level of anxiety and depression within the PD community based on PD survey and mental health survey data. With machine learning modeling results, healthcare providers would be able to offer timely assistance to prevent worsened mental health in the PD community. Some of the biggest challenges of collecting and analyzing mental health survey results include low completion rate and high-cost [7][8]. If we can identify the most predictive features or strong indicators for worsened mental health, it can expedite the mental health survey data collection process with an improved completion rate. In this study, we investigated which questions and features in the survey could efficiently and accurately identify ones at high risk in mental wellness.

II. Data

This study utilizes the survey data from the Michael J. Fox Foundation's online clinical study, Fox Insight [9]. The survey data is the largest self-reported dataset on the impact of the COVID-19 pandemic on the PD community.

The Fox Insight is an online health study of people with and without self-reported PD consisting of regularly administered questionnaires collected longitudinally over multiple years. For those who do not self-report PD, PD connection (e.g., spouse, relative, caregivers with PD) is captured to understand the participants' experience and environmental or genetic factors. Since in-person trials generally enroll participants who already have access to specialist care and mild symptoms, the online survey includes a more diverse and general population.

The survey was designed in consultation with 9 PD patients and their caregivers. To study the early impact of COVID-19, from April 23, 2020 to May 23, 2020, 7,209 responses were gathered from those with and without PD [10]. In this survey, 77 respondents were tested positive for COVID-19, and out of these 77 respondents, 51 have PD.

[1]Hashneet Kaur, Patrick Ka-Cheong Poon and Sophie Yuefei Wang are students in the MS in Data Science Program, University of San Francisco and contributed equally. hkaur25,pkpoon,ywang471@usfca.edu

[2]Diane Myung-kyung Woodbridge, Ph.D. is an Assistant Professor at the MS in Data Science Program, University of San Francisco. dwoodbridge@usfca.edu

Questions
Traveled to areas that had or later received a travel advisory because of a high number of COVID-19 cases?
I have not gone out in public since the COVID-19 pandemic began.
Had COVID-19 symptoms?
Have you been tested for COVID-19?
What was your COVID-19 test result?
Have you been hospitalized due to COVID-19?
Difficulty paying rent or mortgage.
Difficulty paying other bills.
Difficulty sleeping (insomnia) - trouble getting to sleep or staying asleep through the night.

TABLE I: Example questions related to COVID-19

The original data includes 5,308 features, including symptoms, health, and lifestyle, from 50,849 unique participants with/without PD. Among those features, 4,877 features were questionnaires, where 3,543 questionnaires have multiple-choice options. This includes 209 questions specific to symptoms and experiences of COVID-19 in Table I. In addition, only 2,803 participants with PD have survey records both prior to and during the COVID-19 outbreak[1].

III. Algorithm

In order to process data with a significantly large number of features, we utilized distributed computing and machine learning frameworks to process data efficiently. We stored data from Fox Insight in Amazon Web Service (AWS, [11]) Simple Storage Service (S3) and preprocess the data using Apache Spark [12] on Elastic Map Reduce (EMR). For building regression models to predict one's depression level during the pandemic, we installed H2O AutoML [13] on the EMR cluster to build various models and their ensembles. Our previous studies [14][15] show that distributed computing enhances the data processing and model building time considerably.

A. Data Preprocessing

Data preprocessing primarily involves two steps: data filtering and mood score computation based on the 15-item Geriatric Depression Scale (GDS-15) [16] (Table II). The goal of data filtering was to identify PD patients who responded to the COVID-19 survey and to extract all the most recent survey responses prior to COVID-19.

To compute the target labels for the predictive regression model, a measuring scale was required for evaluating the depressive severity during the COVID-19 pandemic. We utilized the GDS-15 score as a primary indicator for depression, as the GDS-15 was shown to serve as a well-performed screening instrument and distinguish depressed patients from non-depressed patients in PD [17]. A score of 12-15 indicates severe depression, 9-11 indicates moderate depression, 5-8 indicates mild depression and 0-4 indicates normal [16]. Algorithm 2 is the pseudocode for calculating the mood score.

[1]Data used in the preparation of this article were obtained from the Fox Insight database (https://foxden.michaeljfox.org/insight/explore/insight.jsp) on 02/12/2021.

Algorithm 1 Data filtering to extract participants in prior to and during the COVID-19 pandemic

1: **select** patients **from** COVID-19 Survey **where** 1) is diagnosed with Parkinson's Disease 2) have responded to the GDS-15 surveys ;
2: **select** * **from** other survey and participant data **left join** participant in line 1 **on** participant with data collected before COVID-19 ;
3: **for** each participant in line 2 **do**
4: duration = COVID-19 survey age - other survey age
5: **if** participant has duration < 0 for each survey **then**
6: get the most recent pre COVID-19 feature among multiple records
7: **end if**
8: **end for**
9: **return** patient data and survey results before COVID-19

TABLE II: GDS-15 questions and corresponding answers indicating depression.

Question	Answer
Are you basically satisfied with your life?	No
Have you dropped many of your activities and interests?	Yes
Do you feel that your life is empty?	Yes
Do you often get bored?	Yes
Are you in good spirits most of the time?	No
Are you afraid that something bad is going to happen to you?	Yes
Do you feel happy most of the time?	No
Do you often feel helpless?	Yes
Do you prefer to stay at home, rather than going out and doing new things?	Yes
Do you feel you have more problems with memory than most?	Yes
Do you think it is wonderful to be alive now?	No
Do you feel pretty worthless the way you are now?	Yes
Do you feel full of energy?	No
Do you feel that your situation is hopeless?	Yes
Do you think that most people are better off than you are?	Yes

The mood score plotted in Figure 1 shows a right-skewed distribution, with the majority scores at approximately 0-4. The preprocessed data is stored in a parquet format with a columnar storage for storage and performance optimization.

B. Machine Learning

Automated Machine Learning (AutoML) enables automation of time, and cost-intensive model building processes, including model and feature selection and hyperparameter optimization [18][19]. H2O Auto ML created models including random forest [20], generalized linear model [21], gradient boosting machine [22], XGBoost [23], deep learning [24], and stacked ensemble [25] with various hyper parameters. H2O AutoML trains Stacked Ensemble models using the best-trained from each family of models and all-trained models to produce a highly predictive ensemble model [26]. We trained regression models using H2O, validated and ranked models based on root mean square error (RMSE, (1))

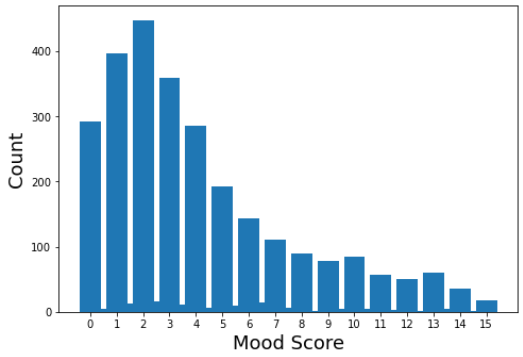

Fig. 1: Mood score distribution after preprocessing

Algorithm 2 Mood score calculation

```
for participant filtered from Algorithm 1 do
    mood score = 0
    for each post COVID-19 GDS-15 question do
        if the answer indicates depression then mood score += 1
        else if participant did not wish to answer then mood score += 0.5
        end if
    end for
end for
return mood score
```

using the validation set.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2} \quad (1)$$

In order to build machine learning models and validate the created models, we split the preprocessed data into 80%, and 20% ratio as training and validation set accordingly, where we train models using 10-fold cross-validation.

IV. Results

In this study, we used a 3-node EMR cluster where each node has 4 vCPU and 16 GB memory. For distributed computing and AutoML, we installed Spark 3.0 and H2O.

Table III shows the 10 best models among 80 models with all of the 5,308 features. The best model with the lowest RMSE value was a stacked ensemble using 78 models, while the second-best model used the best model of each family.

TABLE III: Top 10 machine learning models and their RMSE values of the validation set using all the features - each feature and corresponding ranking were used for experiments

Model	RMSE
Stacked Ensemble - All Models	2.841
Stacked Ensemble - Best of Family	2.851
XGBoost #1	2.879
XGBoost #2	2.880
XGBoost #3	2.906
XGBoost #4	2.916
XGBoost #5	2.922
XGBoost #6	2.922
XGBoost #7	2.929
Deep Learning #1	2.930

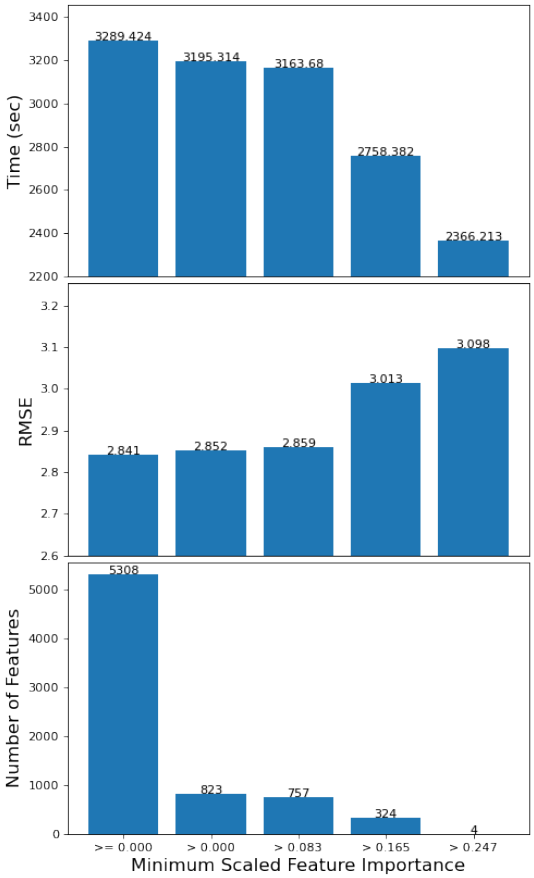

Fig. 2: AutoML outcome with different number of features

Using the models and their importance to create 'Stacked Ensemble Model - All Models' in Table III and the importance of each feature contributed to an individual model, we calculated weighted feature importance ($\bar{v}$) in Equation 2.

$$\bar{v} = \frac{\sum_{i=1}^{n} m_i \cdot v_i}{\sum_{i=1}^{n} m_i} \quad (2)$$

, where m is model importance, and v is scaled feature importance in m. The distribution of weighted feature importance ranged between 0.001 and 0.492, and the 10 features with the highest weighted feature importance are in Table IV.

Limiting the features with different thresholds of weighted feature importance, we re-trained and validated AutoML outcomes. Figure 2 shows the number of features qualifying the scaled feature importance, the minimum RMSE, and time to train and validate the models. Having fewer features helped reduce the execution time by 2.861 to 28.066%, while RMSE increased by 0.387 to 9.046%. As the number of features decreases, execution time decreases while RMSE increases.

V. Conclusion

While recent studies have concluded that people with PD suffer from extreme depression and anxiety during the

TABLE IV: 10 features with the highest weighted feature importance in 'Stacked Ensemble Model - All Models' in Table III

Feature	Weighted Feature Importance
Have you experienced loss of interest in what is happening around you or in doing things in the last month?	0.492
Age	0.365
Due to having Parkinson's disease, how often during the last month have felt depressed?	0.320
Family Neurological History Version - Grandchild	0.249
What's Bothering You - Severity	0.241
Do you experience walking difficulties due to Parkinson's disease?	0.230
Your Current Health	0.226
Your Annual Acute Health Conditions	0.225
Remember back to when your Parkinson's symptoms began. On which side of your body did your symptoms start?	0.224
Your Health History - Eye surgery	0.223

COVID-19 pandemic, research on predicting and providing early intervention is lacking. In this study, we designed a distributed data pipeline that can preprocess data and apply machine learning algorithms to a dataset with a large number of features efficiently. We applied H2O AutoML to predict the depression level score based on GDS-15 and achieved a low RMSE value of 2.841 using data from 2,803 participants with PD. While the original data set includes 5,308 features, we were able to narrow it down to 4 features in Table IV that have the strongest predictive power using both model importance and feature importance analysis, with an RMSE of 3.098.

We will expand our study to investigate the best survey cycle and questionnaire types and contents to predict participants' depression levels more accurately for our future research. This will help determine which questionnaires be asked based on the previous responses to dynamically identify the risk factors and provide timely medical assistance to prevent any adverse events.

ACKNOWLEDGMENT

The Fox Insight Study (FI) is funded by The Michael J. Fox Foundation for Parkinson's Research. We would like to thank the Parkinson's community for participating in this study to make this research possible.

REFERENCES

[1] A. Kumar and K. R. Nayar, "Covid 19 and its mental health consequences," *Journal of Mental Health*, vol. 180, no. 6, pp. 817–8, 2020.

[2] A. van der Heide, M. J. Meinders, B. R. Bloem, and R. C. Helmich, "The impact of the covid-19 pandemic on psychological distress, physical activity, and symptom severity in parkinson's disease," *Journal of Parkinson's disease*, no. Preprint, pp. 1–10, 2020.

[3] A. Lieberman, "Depression in parkinson's disease–a review," *Acta Neurologica Scandinavica*, vol. 113, no. 1, pp. 1–8, 2006.

[4] T. Schirinzi, R. Cerroni, G. Di Lazzaro, C. Liguori, S. Scalise, R. Bovenzi, M. Conti, E. Garasto, N. B. Mercuri, M. Pierantozzi *et al.*, "Self-reported needs of patients with parkinson's disease during covid-19 emergency in italy," *Neurological Sciences*, vol. 41, no. 6, pp. 1373–1375, 2020.

[5] F. Kitani-Morii, T. Kasai, G. Horiguchi, S. Teramukai, T. Ohmichi, M. Shinomoto, Y. Fujino, and T. Mizuno, "Risk factors for neuropsychiatric symptoms in patients with parkinson's disease during covid-19 pandemic in japan," *PloS one*, vol. 16, no. 1, p. e0245864, 2021.

[6] D. Janiri, M. Petracca, L. Moccia, L. Tricoli, C. Piano, F. Bove, I. Imbimbo, A. Simonetti, M. Di Nicola, G. Sani *et al.*, "Covid-19 pandemic and psychiatric symptoms: The impact on parkinson's disease in the elderly," *Frontiers in Psychiatry*, vol. 11, p. 1306, 2020.

[7] P. J. Batterham, "Recruitment of mental health survey participants using internet advertising: content, characteristics and cost effectiveness," *International journal of methods in psychiatric research*, vol. 23, no. 2, pp. 184–191, 2014.

[8] M.-k. Suh, T. Moin, J. Woodbridge, M. Lan, H. Ghasemzadeh, A. Bui, S. Ahmadi, and M. Sarrafzadeh, "Dynamic self-adaptive remote health monitoring system for diabetics," in *2012 Annual international conference of the ieee engineering in medicine and biology society*. IEEE, 2012, pp. 2223–2226.

[9] L. Smolensky, N. Amondikar, K. Crawford, S. Neu, C. M. Kopil, M. Daeschler, L. Riley, E. Brown, A. W. Toga, and C. Tanner, "Fox insight collects online, longitudinal patient-reported outcomes and genetic data on parkinson's disease," *Scientific data*, vol. 7, no. 1, pp. 1–9, 2020.

[10] E. G. Brown, L. M. Chahine, S. M. Goldman, M. Korell, E. Mann, D. R. Kinel, V. Arnedo, K. L. Marek, and C. M. Tanner, "The effect of the covid-19 pandemic on people with parkinson's disease," *Journal of Parkinson's disease*, no. Preprint, pp. 1–13, 2020.

[11] Amazon Web Service. (2021) Amazon. [Online]. Available: https://aws.amazon.com

[12] Apache Spark, "Apache spark: Lightning-fast cluster computing," 2021. [Online]. Available: http://spark.apache.org

[13] E. LeDell and S. Poirier, "H2O AutoML: Scalable automatic machine learning," *7th ICML Workshop on Automated Machine Learning (AutoML)*, July 2020. [Online]. Available: https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf

[14] D. Fozoonmayeh, H. V. Le, E. Wittfoth, C. Geng, N. Ha, J. Wang, M. Vasilenko, Y. Ahn, and D. M.-k. Woodbridge, "A scalable smartwatch-based medication intake detection system using distributed machine learning," *Journal of Medical Systems*, vol. 44, no. 4, pp. 1–14, 2020.

[15] D. M.-k. Woodbridge and K. Wong, *Big Data in Psychiatry and Neurology*. Elsevier, 2021, ch. A Scalable Medication Intake Monitoring System.

[16] S. A. Greenberg, "The geriatric depression scale (gds)," *Best Practices in Nursing Care to Older Adults*, vol. 4, no. 1, pp. 1–2, 2012.

[17] D. Weintraub, K. A. Oehlberg, I. R. Katz, and M. B. Stern, "Test characteristics of the 15-item geriatric depression scale and hamilton depression rating scale in parkinson disease," *The American journal of geriatric psychiatry*, vol. 14, no. 2, pp. 169–175, 2006.

[18] P. Gijsbers, E. LeDell, J. Thomas, S. Poirier, B. Bischl, and J. Vanschoren, "An open source automl benchmark," *arXiv preprint arXiv:1907.00909*, 2019.

[19] M.-A. Zöller and M. F. Huber, "Survey on automated machine learning," *arXiv preprint arXiv:1904.12054*, vol. 9, 2019.

[20] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[21] N. E. Breslow, "Generalized linear models: checking assumptions and strengthening conclusions," *Statistica Applicata*, vol. 8, no. 1, pp. 23–41, 1996.

[22] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[23] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[24] A. Candel, V. Parmar, E. LeDell, and A. Arora, "Deep learning with h2o," *H2O. ai Inc*, 2016.

[25] M. J. Van der Laan, E. C. Polley, and A. E. Hubbard, "Super learner," *Statistical applications in genetics and molecular biology*, vol. 6, no. 1, 2007.

[26] h2o.ai, "Automl: Automatic machine learning," 2020. [Online]. Available: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html