Predicting Parkinson's Disease for Patients Using Voice Recording

Jennifer Matas, Samantha Harrison, Cesar Pinzon Gomez, Kathryn Shahan & Sandra Ovuegbe

Fundamentals of Data Analytics and Prediction: Group 4

April 14, 2022

Introduction

Parkinson's disease (PD) affects many people, with an estimated prevalence rate of 572 per 100,00 people over the age of 45 having it in North America.[14] It is a progressive neurodegenerative disorder, and often begins its onset around the ages of 65-70.[1] Symptoms include slowness of movement, rigidity, tremors, and speech problems.[1] Since speech problems are one of the earliest symptoms to appear, they have been used in predictive models to determine if a patient has PD.[15] Similar speech recognition studies have tested a variety of machine learning models including Neural networks, logistic regression, Support Vector Models-linear and radial, Decision trees, and Boosting among others[10,11,12]. The majority of studies determine SVM radial to have a higher predictive accuracy of 96.7%[10] up to 99.8%[11] compared to other machine learning techniques, while a few studies propose that advanced Boosting techniques yield better accuracy[12]. In this analysis, we tested Support Vector machines, Boosting, and Random forest algorithms on voice recordings from patients to create a machine-learning model to predict patients as having PD or not. Based on previous studies, we hypothesized that Radial SVM classifiers would yield the highest accuracy for predicting PD.

Methods

In the present study, two datasets containing information from speech signals of patients with suspected Parkinson's disease were received. We received a dataset called 'project_training_set_p' that contained values of 753 features and a binary PD indicator (1=PD/0=no PD) from a total of 528 patients. Since outcome (i.e., PD) data was available in this dataset, we implemented supervised learning strategies to train our ML algorithms.

Statistical Analysis

An initial verification of missing data was carried out and it is confirmed that there were no observations with missing values. Subsequently, we proceeded to evaluate the correlation (r=|>=0.80| of all the predictors.

Initially, this first dataset was divided into a training and test subset. Different split ratios were used in all models, but in the end, a ratio was chosen 60% (n=317) for the training subset and 40% (n=211) for the test set. Support Vector Machines (both Linear and Radial), Random Forest, and Boosting were chosen as the most suitable models of prediction. Two different analytical strategies were employed to assess the predictive ability of each ML strategy. In the first strategy, we used all 753 features to predict PD and in the second strategy we only used the 50 most relevant features to predict PD. The 50 features were selected by the minimum redundancy-maximum relevance (mRMR) feature selection strategy.

Each model was selected with the end goal of binary classification in mind. Support vector machines were selected due to their tendency to function well with high-dimensional data, as they are not as affected by fluctuations in all of the data. We chose to test both linear SVM and SVM with radial kernels, so as to see if the data classifications would best be captured by a linear split, or a radial split. The Random Forest method was selected due to its tendency to work well with large datasets, as it creates multiple decision trees from which to produce its final output. Random forests were also a good fit because the end goal of the project is prediction of Parkinson's patients, not easy interpretation of the model. Finally, boosting was selected because its method of building upon each tree one at a time and combining results along the way works to combat the bias introduced by the imbalanced classes in the training dataset.

While there are many ways of assessing model accuracy, our preferred way of assessing and selecting our final model was accuracy rate. The accuracy rate describes the percentage of observations that the model correctly classifies when it has been fit on the training set, and then run on the test set. The higher the accuracy rate, the better. Because we were predicting a disease outcome, we also chose to use sensitivity and specificity to assess our models. Sensitivity describes the accuracy of a test in correctly classifying a patient as diseased, while specificity describes the accuracy of a test in correctly classifying a patient as not diseased.[9] As sensitivity goes up, specificity goes down, and vice versa. Therefore, it was

important to us to look for a model with a good balance between sensitivity and specificity. Finally, we also used the Receiver Operator Characteristic (ROC) curve, which plots the performance of classification models, for a visual of our model accuracy. The combination of assessing these accuracy measures led to the selection of a Support Vector Machine model with radial kernels and top fifty features used as predictors, as this model had high accuracy and a good balance between sensitivity and specificity, as well as a high curve on the ROC chart.
Using the Caret (Classification and regression training) package, the standardization of the predictors and model training processes for the current classification problem was carried out, as well as repeated cross-validation to select the best tuning parameter for each model. In each method, a ROC curve was plotted for both all the features and the top 50 so that the relationship between sensitivity, specificity, and the area under the curve could be objectively evaluated to choose the best model. Finally, the selected model was applied to the second set of data called "project_test_set_p", to predict and classify each observation as "CASE" or "HEALTHY". These predictions were stored in a .csv archive for final review by the teaching team. All analyses were carried out in R Studio version 4.1.2 (Comprehensive R Archive Network, https://cran.r‑project.org/).

RESULTS

There was no missingness in the datasets under study. The original training dataset had 528 observations, 267 females and 261 males and 135 and 393 subjects without and with PD, respectively. The test dataset had 228 observations with an unknown diagnosis of PD. Notably, the class distribution is highly imbalanced in the training dataset, an almost 3:1 ratio of subjects with PD to without.

A correlation table indicated that there were many combinations of variables that were highly correlated. Since the training dataset is large the correlation was restricted to r= >=|0.80|. A glimpse of the top 6 positively and negatively correlated features is shown in table 1.

Table 1: Top 6 positively and negatively correlated features

Feature1	Feature 2	coefficient
det_entropy_shannon_10_coef	det_TKEO_mean_10_coef	-0.9994806
det_entropy_shannon_10_coef	det_TKEO_std_10_coef	-0.9991379
det_entropy_shannon_8_coef	det_TKEO_std_8_coef	-0.9989710
det_entropy_shannon_8_coef	det_TKEO_mean_8_coef	-0.9985239
det_entropy_shannon_1_coef	det_TKEO_mean_1_coef	-0.9978357
tqwt_minValue_dec_26	tqwt_maxValue_dec_26	-0.9975420
app_entropy_shannon_8_coef	app_entropy_shannon_9_coef	0.9999994
app_LT_entropy_log_9_coef	app_LT_entropy_log_10_coef	0.9999994
app_entropy_log_8_coef	app_entropy_log_9_coef	0.9999996
app_LT_entropy_shannon_8_coef	app_LT_entropy_shannon_9_coef	0.9999998
app_LT_entropy_log_8_coef	app_LT_entropy_log_9_coef	1.0000000
apq3Shimmer	ddaShimmer	1.0000000

Model building

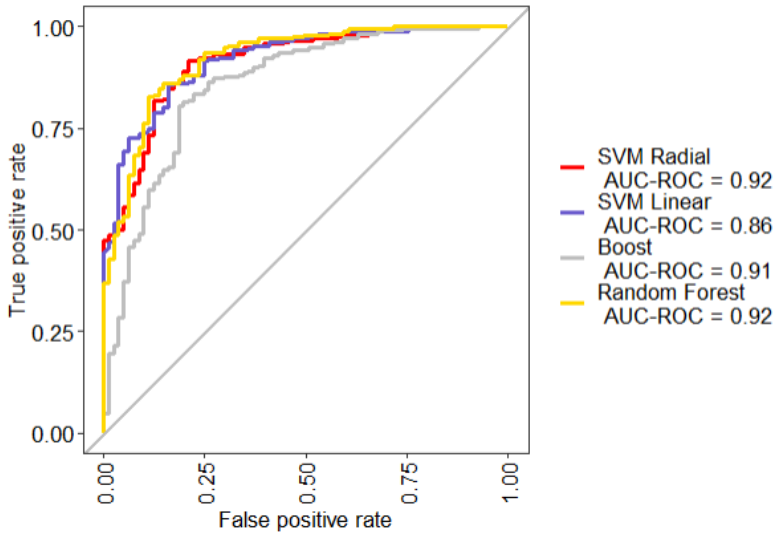
The top 50 features selected with mRMR are shown in table 2. For the top 50 features: 5 were from the baseline feature subset, 2 from vocal fold feature, 10 from MFCC, 4 from WT, and 29 from TQWT. Accuracy, sensitivity and specificity for all ML models on all features and top 50 features is shown in

table 3. Overall, ML models on top 50 features produced a higher accuracy than on all features, with smaller discrepancy between sensitivity and specificity. For the top 50 features a ROC plot shown in figure 1, gives the highest ROC value for both SVM - radial and random forest at ROC = 0.92. Because SVM with radial kernels using the top 50 features yielded high test accuracy = 0.863 and little discrepancy between sensitivity = 0.685 and specificity = 0.924 it was selected as the ML strategy to employ to predict PD on the final test dataset.

Table 3: Accuracy, Senstivity, and Specificity of ML models on all features (m=753) and top 50 features selected by mRMR.

	Accuracy	Sensitivity	Specificity
SVM Radial- All features	0.848	0.741	0.885
SVM Radial- Top 50 features	0.863	0.685	0.924
SVM Linear- All features	0.806	0.519	0.904
SVM Linear- Top 50 features	0.844	0.611	0.924
Boosting- All features	0.863	0.611	0.949
Boosting- Top 50 features	0.863	0.741	0.904
Random Forest- All features	0.867	0.537	0.981
Random Forest- Top 50 features	0.882	0.648	0.962

Figure 1. ROC plot using top 50 features



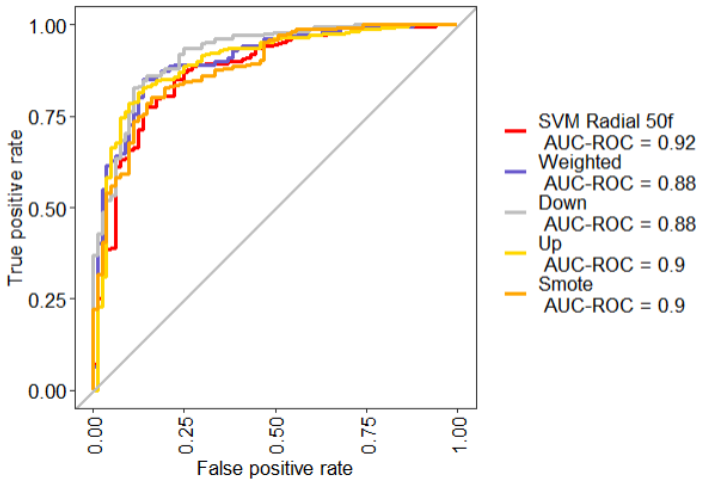
Discussion

In this study, we have tested the accuracy of several different machine learning models in predicting Parkinson's disease from vocal samples of patients. This presented a binary classification problem, as the goal was to identify each patient as either having Parkinson's disease or not having it. Each model was tested using all provided features, as well as tested using the top fifty features as selected by mRMR. The initial dataset had many predictors and few observations.A model with many more predictors than observations runs the risk of overfitting, in which the model fits too well to the training

data instead of capturing general trends, and then fails to predict the test data as accurately as possible. This also results in a model with high variance, as it is fitted too close to small fluctuations in the training data. The final predictive model chosen was the Support Vector Machine model with radial kernels using the top fifty features as predictors, since this model had both a high test accuracy and high sensitivity and specificity.

The results of our study have several implications for future research in machine learning and predicting Parkinson's disease. First, our results suggest that support vector machines may be a good model to continue experimenting with in order to predict Parkinson's disease, as the model had a high accuracy rate. This finding is consistent with similar literature that found SVM to be most accurate compared to boosting and decision trees. Second, the results show the importance of dealing with high-dimensional data in a manner that reduces the possibility of overfitting and high variance in the model. This is shown by the model accuracy consistently being the same as or higher than the original results when the model is run only using the top fifty features. Feature selection is common in similar literature, however, none of them show the stark contrast between accuracies derived from all features vs accuracies from the top selected features as our study does. Finally, the results show a step forward in Parkinson's research by creating another accurate model that can help identify Parkinson's patients earlier from their vocal sample and contribute to earlier medical intervention. Further research should be done using this model and others to further test the possibility of classifying Parkinson's Disease using machine learning. Our model also has the potential to be applied to the prediction of other neurological diseases like Alzheimer's that impact speech.

Figure 2: ROC Curves after applying methods to improve performance for imbalanced data



We had three main study limitations. First, because the dataset was so large, we noticed the predictors were highly correlated. Multicollinearity causes an overfitting problem because a change in one predictor will cause changes in another correlated predictor[13]. Second, after fitting our SVM radial model, we observed a class imbalance because the number of patients diagnosed with Parkinson's outweighed the number of patients without Parkinson's (393 vs 135). A class imbalance can lower the predictive performance of our models. We tried to address this in 4 ways as seen in figure 2. First, by imposing heavier costs when errors are made on the minority class (class weights)[8], second, by randomly removing instances in the majority class (down-sampling)[8], third by randomly replicating instances in the minority class (upsampling)[8], and fourth by downsampling the majority class and synthesizing new minority instances (SMOTE)[8]. The techniques, however, did not prove futile since the original SVM model outperformed the 4 modeling techniques. Lastly, the original training dataset had a small sample size of 528 with 753 predictors. Fitting our models on this dataset would result in overfitting since we had more predictors than observations. To address this, we performed mRMR feature selection to include only the

top 50 features in our models. Feature selection was found to be appropriate in this analysis since the model fit with only 50 features outperformed those fit with all the features.

References

1. Tysnes OB, Storstein A. Epidemiology of Parkinson's disease. *J Neural Transm (Vienna)*. 2017;124(8):901-905. doi:10.1007/s00702-017-1686-y

2. De Jay Nicolas, Papillon-Cavanagh Simon, Olsen Catharina, Bontempi Gianluca, Haibe-Kains Benjamin. mRMRe: an R package for parallelized mRMR ensemble feature selection. September 3, 2021; https://cran.r-project.org/web/packages/mRMRe/vignettes/mRMRe.pdf. Accessed April 02, 2022.
3. Liu L, Chen L, Zhang Y-H, et al. Analysis and prediction of drug–drug interaction by minimum redundancy maximum relevance and incremental feature selection. *Journal of Biomolecular Structure and Dynamics*. 2017;35(2):312-329.
4. Radovic M, Ghalwash M, Filipovic N, Obradovic Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics*. 2017;18(1):9.
5. Toğaçar M, Ergen B, Cömert Z. Detection of lung cancer on chest CT images using minimum redundancy maximum relevance feature selection method with convolutional neural networks. *Biocybernetics and Biomedical Engineering*. 2020;40(1):23-39.
6. Kuhn M. Classification and Regression Training. March 11, 2022; https://cran.r-project.org/web/packages/caret/caret.pdf. Accessed April 10, 2022.
7. Raghuwanshi, B. S., & Shukla, S. (2018). Class-specific extreme learning machine for handling binary class imbalance problem. *Neural networks : the official journal of the International Neural Network Society*, *105*, 206–217. https://doi.org/10.1016/j.neunet.2018.05.011
8. Martin, Dan. Handling class imbalance with R and Caret - An introduction. December 10, 2016; https://dpmartin42.github.io/posts/r/imbalanced-classes-part-1. Accessed April 12, 2022.
9. Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, *56*(1), 45–50. https://doi.org/10.4103/0301-4738.37595
10. Reid J, Parmar P, Lund T, Aalto DK, Jeffery CC. Development of a machine-learning based Voice Disorder Screening Tool. *American Journal of Otolaryngology*. 2022;43(2):103327. doi:10.1016/j.amjoto.2021.103327
11. Singh J, Rajnish R, Singh DK. Designing a machine learning model to predict parkinson's disease from Voice Recordings. *Second International Conference on Sustainable Technologies for Computational Intelligence*. 2021:95-103. doi:10.1007/978-981-16-4641-6_9
12. Nissar I, Rizvi D, Masood S, Mir A. Voice-based detection of parkinson's disease through Ensemble Machine Learning Approach: A performance study. *EAI Endorsed Transactions on Pervasive Health and Technology*. 2019;5(19):162806. doi:10.4108/eai.13-7-2018.162806
13. Wu S. Multi-collinearity in regression. Medium. https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea. Published June 5, 2021. Accessed April 13, 2022.
14. Marras, C., Beck, J.C., Bower, J.H. *et al.* Prevalence of Parkinson's disease across North America. *npj Parkinson's Disease* 4, 21 (2018). https://doi.org/10.1038/s41531-018-0058-0
15. Erdogdu Sakar, B., Serbes, G., & Sakar, C. O. (2017). Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease. *PloS one*, *12*(8), e0182428. https://doi.org/10.1371/journal.pone.0182428

Appendix

Table 2: Top 50 features selected by minimum redundancy – maximum relevance (mRMR)

tqwt_energy_dec_26	std_6th_delta_delta
tqwt_TKEO_mean_dec_12	Ea2
tqwt_kurtosisValue_dec_18	GQ_prc5_95
ppq5Jitter	mean_MFCC_2nd_coef
tqwt_energy_dec_25	std_delta_log_energy
tqwt_TKEO_std_dec_7	tqwt_kurtosisValue_dec_20
tqwt_entropy_log_dec_12	tqwt_kurtosisValue_dec_12
std_9th_delta_delta	tqwt_energy_dec_27
tqwt_kurtosisValue_dec_34	tqwt_kurtosisValue_dec_26
tqwt_TKEO_mean_dec_11	det_LT_entropy_shannon_3_coef
std_7th_delta_delta	tqwt_energy_dec_12
tqwt_entropy_shannon_dec_34	DFA
std_11th_delta_delta	tqwt_TKEO_mean_dec_13
tqwt_entropy_log_dec_34	tqwt_entropy_log_dec_26
std_delta_delta_log_energy	tqwt_entropy_shannon_dec_33
ddpJitter	tqwt_entropy_shannon_dec_8
Ed2_3_coef	std_8th_delta_delta
tqwt_maxValue_dec_12	tqwt_TKEO_mean_dec_26
tqwt_energy_dec_28	tqwt_kurtosisValue_dec_36
IMF_SNR_SEO	stdDevPeriodPulses
mean_delta_log_energy	std_10th_delta_delta
tqwt_entropy_shannon_dec_12	tqwt_TKEO_mean_dec_36
tqwt_kurtosisValue_dec_27	tqwt_energy_dec_21
tqwt_kurtosisValue_dec_16	tqwt_entropy_shannon_dec_36
meanPeriodPulses	det_LT_entropy_shannon_4_coef