FINAL PROJECT
Classification Analysis of Parkinson Speech Dataset

Elcin Ergin[1], Shu Hayakawa[2], and Timardeep Kaur[3]

[1]260628206-Desautels Faculty of Management, McGill University
[2]260462114-School of Computer Science, McGill University
[3]260617404-Electrical and Computer Engineering, McGill University

***Abstract*—About 90% of patients detected with Parkinson's disease experience changes in their voice or their ability to make speech sounds [1]. So in recent years, there is an increased interest to detect the Parkinson's Disease in patients by using their voice samples. In this project, we are using the dataset from https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with++Multiple+Types+of+Sound+Recordings, which contains 26 speech recordings for each patient which include sustained vowels, words and sentences compiled from a set of speaking exercises for people with Parkinson's Disease. We also attempt to create another dataset with extended feature space containing central tendency and dispersion metrics for all features in the given dataset in order to summarize the data and to reduce the variations between different voice samples of a subject. Popular machine learning classification algorithms like logistic regression, linear and non linear SVM and K-nearest neighbors are applied on the given dataset and the dataset created by authors to classify each patient as being Parkinson positive or Parkinson negative. We further measure the success of each algorithm for their ability to correctly classify the patients into one of these categories. It is observed that SVM performs the best in both the datasets. We also identified three best voice samples, to be used for the purpose of classification, which can be collected in case we have less time to collect voice samples from the patients.**

***Index Terms*—Parkinson's disease, Binary Classification,SVM, KNN, Logistic Regression, Regularization**

I. Introduction and Motivation

Parkinson's disease (PD) is a degenerative neurological disorder that causes partial or complete loss in motor reflexes, speech, behavior, mental processing, and other vital functions [2]. It affects seven to ten million people worldwide, most of them over the age of 60 (Parkinson's Disease Foundation). People with Parkinsonism suffer from speech impairments like dysphonia (defective use of the voice), hypophonia (reduced volume), monotone (reduced pitch range),and dysarthia (difficulty with articulation of sounds or syllables).

Since vocal impairment is prevalent in 70-90% of patients after the onset of disease, we are using this to detect if the patient has Parkinson or not in our project. In addition, it may be one of the earliest indicators of the disease [3] and 29% of patients consider it as one of their greatest hindrances [4]. The main reason behind the popularity of PD diagnosis from speech impairments is that telediagnosis and telemonitoring systems based on speech signals are low in cost and easy to self-use [5]. So these systems reduce the inconvenience and cost of physical visits of patients to the clinics and help in the early diagnosis of the disease [6].

Although medication and surgical intervention can hold back the progression of the disease and alleviate some of the symptoms, there is no available cure. Thus, early diagnosis is critical in order to improve the patient's quality of life and prolong it.

In this study, we mainly replicate the work of Sakar et al. [7] applied on the dataset available on UCI [8]. We also extend the study by including Logistic Resgression and PCA analysis.

In section II, the studies applied on similar datasets are presented. Section III introduce the dataset and features. In section IV, the algorithms and cross validation techniques are explained. Section V gives the analysis of the application in detail. Finally, the discussion of the results and conclusion is given in section VI.

II. Related Work

In literature, different studies can be found which focus on speech measurement for general voice disorders [9], [10], [11] and Parkinson's Disease in particular [12],[13], [14] [5], [15]. Some of the studies use a regression approach to detect the level of PD utilizing the UPDRS (Unified Parkinson's Disease Rating Scale) measurements while other studies approach to problem as a classification problem to detect whether the patient has Parkinson or not.

Little et al. [5] aimed to detect the disease by measuring the dysphonia caused by PD with a dataset containing sustained vowel "a" phonations of 31 subjects. The best performance (91.4%) is obtained by SVM with RBF. Tsanas et.al [15] used linear and nonlinear regression techniques to predict progression of disease (UPDRS level) on a set of 6000 samples of 42 PWP. Tsanas et. al. [14] also studied a classification problem with the extended version of the dataset used in [15]. They also obtained the best results from non-linear SVM with 97% accuracy. Additionally they applied LASSO and Random Forests in their study. The previous works point out the success of SVMs on the classification analysis of the Parkinson speech dataset.

III. Dataset

The training data collected in the context of this study belongs to 20 PD (Parkinson Disease) patients (6 female, 14 male) and 20 healthy individuals (10 female, 10 male) as explained in the original paper [7]. 26 voice samples were collected from each subject, where these voice samples include sustained vowels, numbers, words, and short sentences. Additionally, an independent test set was collected from 28 PD patients who were asked to say only the sustained vowels "a" and "o" three times, respectively which makes a total of 168 recordings. That is why the test results cannot directly reflect the success of models on unhealthy people. It should be noted that this study focuses on the problem of detecting the disease from voice samples of PD patients.

The list of features used in this study are given in Table I. In the paper the features are extracted considering previous studies [5], [16]. The first 5 features listed in table represent frequency parameters, 6 to 10 represent pulse parameters, 11 to 15 represent amplitude parameters, 16 to 18 are voicing parameters, 19 to 23 are pitch parameters and remaining are harmonicity parameters.

TABLE I: List of features

	Features
1	Jitter(local)
2	Jitter(local,absolute)
3	Jitter(rap)
4	Jitter(ppq5)
5	Jitter(ddp)
6	Number of pulses
7	Number of periods
8	Mean period
9	Standard deviation of period
10	Shimmer(local)
11	Shimmer(local,dB)
12	Shimmer(apq3)
13	Shimmer(apq5)
14	Shimmer(apq11)
15	Shimmer(dda)
16	Fraction of locally unvoiced frames
17	Number of voice breaks
18	Degree of voice breaks
19	Median pitch
20	Mean pitch
21	Standard deviation
22	Minimum pitch
23	Maximum pitch
24	Autocorrelation
25	Noise-to-harmonic
26	Harmonic-to-noise

When we examined the correlations between features as it is seen in figure 1, we see that there are strong positive and negative correlations between the features in same groups. For example, features 6 to 10 (pulse parameters) show strong correlation as they are in the same group. Even though we cannot include or exclude features only using the cross correlation values, applying regularization could improve the results.

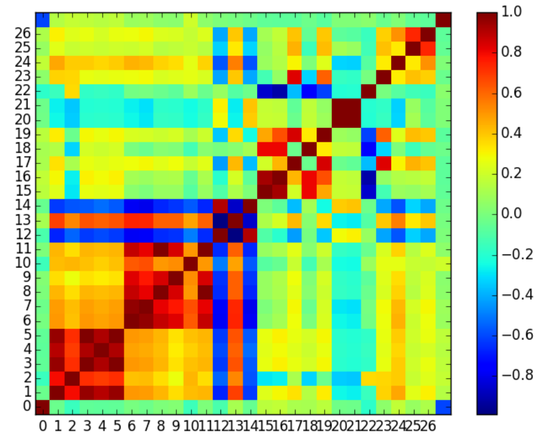

Fig. 1: Cross Correlation Between Features

IV. Methodology

A. Algorithms

In this study we applied three machine learning algorithms which are logistic regression, support vector machines(SVM) and k-nearest neighbours(KNN). SVM and KNN are the methods applied in the original paper. We extend it by including logistic regression to get base results. In this section, the theory is explained briefly. The implementation of the models and the results are presented in Section V.

1) Logistic Regression: Logistic Regression is a discriminative learning algorithm that tries to directly estimate $P(y|X)$. In the two-class case where $y \in \{0, 1\}$, the bayesian probability of $y = 1|X$ becomes:

$$p(y = 1|X) = \sigma(\mathbf{w}^T X)$$

where:

$$\sigma(z) = \frac{1}{1 + exp(-z)}$$

Under the i.i.d. assumption the likelihood function of the weights $P(X_1, y_1, ..., X_n, y_n|\mathbf{w})$ is:

$$\prod_{i=1}^{n} \sigma(\mathbf{w}^T X^{(i)})^{y^{(i)}} (1 - \sigma \mathbf{w}^T X^{(i)}))^{(1-y^{(i)})}$$

When L2-regularization (Ridge) term is added with regularization parameter λ, the following log-likelihood function is obtained:

$$\sum_{i=1}^{n} [y_i log(\sigma(\mathbf{w}^T X^{(i)})) + log(1 - \sigma(\mathbf{w}^T X^{(i)}))] + \frac{\lambda}{2m} \sum_{j=1}^{m} w_j^2$$

Instead of adding squared terms of weights, adding absolute values of the weights would yield the L1-regularization (LASSO).

In this study, both ridge regression and LASSO are utilized.

2) Support Vector Machines: Introduced by Vapnik based on his statistical learning theory [17], SVMs are powerful models commonly used for classification and regression problems. In a broad sense, the standard linear SVM constructs a hyper plane to classify data using the input variables. SVM solves a quadratic programming problem with linear constraints.

The optimization problem for SVM can be represented in the following way [18].

$$min\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$s.t.:$$

$$\xi_i \geq 0$$

$$y_i(x_i^T w) \geq 1 - \xi_i \qquad \forall i$$

where ξ_i is the proportional amount by which the prediction is on the wrong side of its margin. Using a "cost" parameter C the objective function tries to bound the total proportional amount by which predictions are misclassified.

The Lagrange function is:

$$L_P = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i - \sum_i \alpha_i(y_i(w^T x_i) - (1-\xi_i)) - \sum_{i=1}^{m}\mu_i\xi_i$$

The derivatives of L_P w.r.t. w is taken and set to 0 to obtain w.

$$w^* = \sum_i \alpha_i y_i x_i$$

The support vector classifier described so far can be extended using kernels to obtain non linear boundaries by mapping input features via $h(x_i)$. Considering that $h(x_i)$ is seen as its inner products in the Lagrange form; without specifying $h(x_i)$ itself, knowledge of the kernel function would be enough. In this study, for non-linear SVM, radial basis function (RBF) is used as kernel function considering the earlier studies applied on similar datasets [14], [5].

3) KNN: Nearest neighbour is a non parametric method which uses the observations in training set closest in the input space to x to produce predictions. Specifically, the k-nearest neighbor fit for $\hat{Y}$ is defined in the following way [18]:

$$\hat{Y} = \frac{1}{k}\sum_{x_i \in N_k(x)}$$

where $N_k(x)$ is the neighborhood of x defined by the k closest points x_i in the training sample. In this study we used Euclidean distance as the metric for closeness as in the original paper [7].

B. Cross-Validation Techniques

1) Leave one Subject Out (LOSO): Since the dataset contains multiple speech recordings per subject it may cause a bias and variance problems in the application. To overcome these problems, the first suggestion in the paper [7] was to use leave-one-subject-out (LOSO) validation scheme in which all the voice samples of one individual is left out to be used for validation as if it is an unseen individual, and the rest of the samples is used for training. According to the LOSO validation scheme, if the majority of the voice samples of test individual are classified as PWP, then the individual is classified as positive and otherwise negative.

2) Summarized Leave One Out (s-LOO): The second proposition in paper [7] was to use a Summarized-Leave-One-Out validation scheme, where feature values of 26 voice samples from each subject was summarized using central tendency and dispersion metrics. Their aim is to decrease the variation effect between different voice samples of one subject.

C. Performance Criteria

To evaluate the models different types of evaluation criteria such as accuracy, sensitivity, specificity and Matthew correlated coefficient (MCC) are incorporated in this study. Unfortunately, there is no negative labeled example is in the test set as it is mentioned in III, which makes Specificity directly 0.

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Sprecificity = \frac{TN}{TN + FP}$$

MCC, with the below formula, gets the value of +1 when all of the prediction values are correct, -1 when the predictions and actual values totally disagree, and 0 when the classification is no better than a random prediction.

$$MCC = \frac{TP * TN - FPxFN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

V. IMPLEMENTATION AND RESULTS

A. Implementation with LOSO

We applied LOSO cross validation technique discussed in section IV-B using the logistic regression, KNN, linear and non-linear SVM using the "Scikit" package available in Python [19].

1) Hyperparameter Selection for Algorithms: We used LOSO CV scheme to select best parameter sets for each method applied on the original training set which includes 26 voice samples from each of the 40 subjects. For each example, we also replicated the LOSO CV scheme to decrease the variation in the result on validation set.

Logistic Regression

We implemented logistic regression using stochastic gradient descent algorithm. The parameters to be tuned for logistic regression include regularization coefficient (alpha), penalty

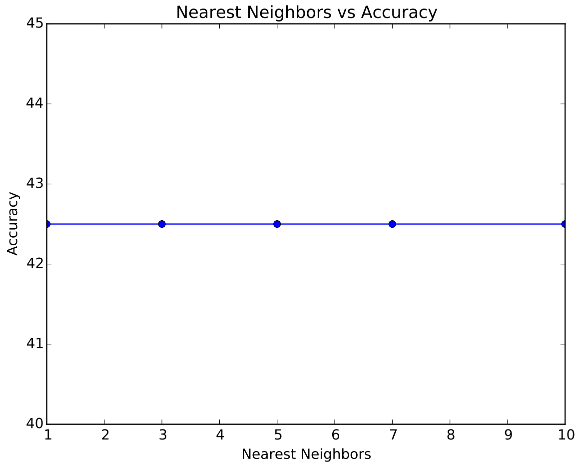

Fig. 2: Cross Validation Results of KNN for LOSO

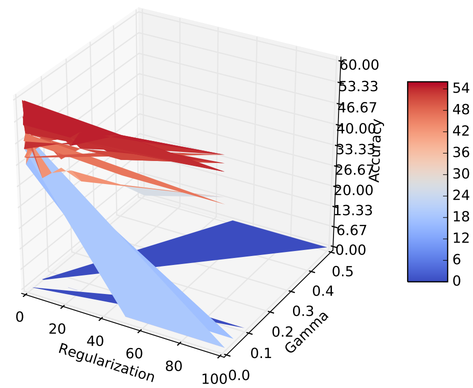

Fig. 3: Cross Validation Results of non linear SVM for LOSO

(type of regularization-L1, L2 and elasticnet) and number of iterations (epochs). We performed the exhaustive grid search over the large range of values to find the optimal values of these parameters using 10 replications of LOSO CV. We selected the best parameters according to accuracy. The best parameters found are: penalty as L1, number of iterations as 30 and alpha as 0.006. The figure showing the cross validation can be found in the appendix as figure 9.

KNN

We are only finding optimal value for number of nearest neighbours. We tried with different number of neighbors varying between 1 and 10; however, all of them gave the same accuracy as it is seen in the figure 2. So, we have selected the k=5 randomly.

Linear SVM

For linear SVM, the only parameter to be tuned was the regularization coefficient (C) and the best value for it is found as 5. The corresponding plot can be found in appendix as figure 10.

Non Linear SVM

We used RBF kernel as it is mentioned in section IV-A. The parameters tuned are regularization coefficient (C) and kernel width γ . C is found as 1 and γ as 0.00005. The variation in accuracy with respect to parameters can be seen in a 3-D plot in figure 3. The dark red area show the area that gives the best accuracy.

2) Results for Validation and Test Set: In this section we present the results of LOSO implementation for both validation and test set. Since we have both positive and negative labels in the validation set, we provide all the performance metrics for validation set in table II. As we have only positive labeled (PWP) subjects in the test set, specificity and sensitivity measures are not meaningful for the test results. In table III, only accuracies for test result is given.

According to the results, logistic regression, KNN and linear SVM did not perform well according to the negative MCC values which means that they are worse than random guess. However, KNN performed well on the test set with 60.7% accuracy.

Non-Linear SVM with positive MCC performed slightly better than the other methods. It also has the best accuracy value for the test set which is 60.7%. This result was expected since the best results in the earlier studies were obtained via non-linear SVM.

TABLE II: Results for Validation Set using LOSO

	Accuracy	**Specifity**	**Sensitivity**	**MCC**
Logistic Regression	47.5%	47%	55%	-0.05
KNN	42.5%	43%	40%	-0.15
Linear SVM	47.5	48%	45%	-0.05
Non Linear SVM	60%	62%	70%	0.2

TABLE III: Results for Test Set using LOSO

	Accuracy
Logistic Regression	25%
KNN	60.7%
Linear SVM	32.1%
Non Linear SVM	60.7%

B. Implementation with s-LOO

As it is mentioned in section IV-B, the multiple voice samples from each subject could create a variance problem. In this implementation, we created a new dataset with extended feature space and reduced example space. For each subject, we took 26 voice samples and calculated mean, median, trimmed mean as central tendency metrics and standard deviation, interquartile range and mean absolute deviation as dispersion metrics. In the new dataset we have 40 examples which is the number of subjects and 183 features. Since we have a huge number of features compared to the number of examples, regularization and choosing the best features are the important aspect of this implementation.

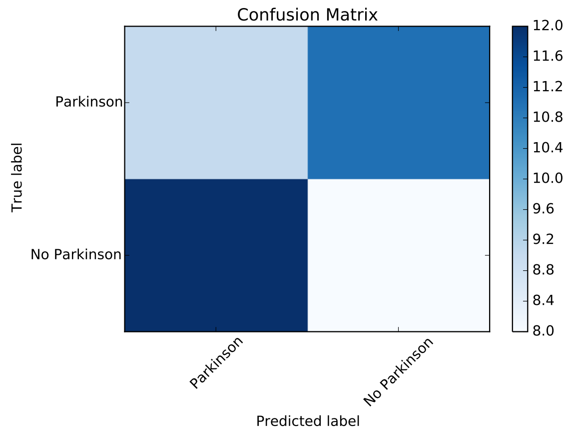

Fig. 4: Confusion Matrix for KNN for LOSO

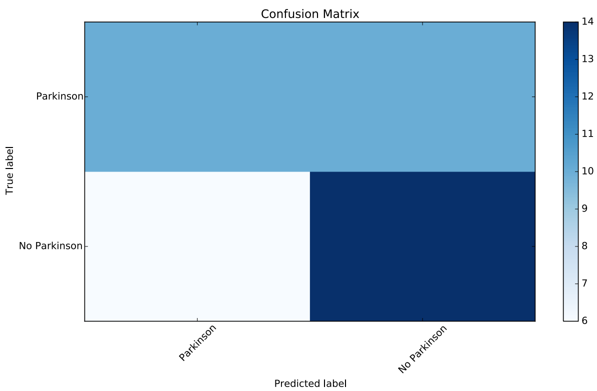

Fig. 5: Confusion Matrix for Non Linear SVM for LOSO

1) Hyperparameter Selection for Algorithms:

Logistic Regression

Similar to selection of hyper-parameters in LOSO CV implementation, we searched for the best parameters for s-LOO using grid search. Again we incorporated type of regularization as another parameter to be tuned in this stage. At the end we obtained penalty type as L2 and regularization coefficient as 0.5.

KNN

In KNN, in order to control number of features included in the model, we also applied PCA (Principal Component Analysis). Number of components to be included in PCA is decided via leave-one-out cross validation. Both number of neighbours and number of components in PCA is tuned via grid search. The optimal number of components is found as 10 and number of neighbours is found as 8. As it is seen in figure 6, number of neighbours highly affects the accuracy.

Linear SVM

Similar to application of KNN, PCA is also included in the study to decrease the number of features in the model. Best number of component for PCA is found as 29. Even though

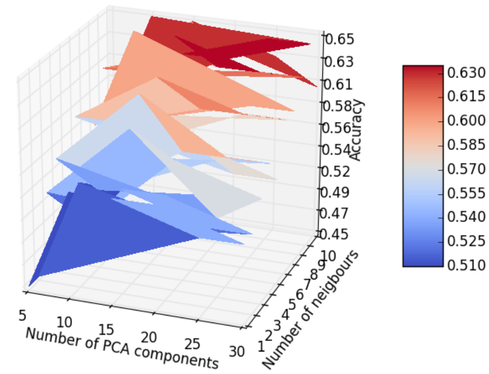

Fig. 6: Cross Validation Results of KNN with PCA for s-LOO

it was 10 in KNN, when the figure 6 was examined, it is seen that KNN with 29 components in PCA also gives a close result. The best penalty coefficient is found as 10000 using grid search.

Non-Linear SVM

The best number of components for PCA, penalty coefficient and γ are found as 29, 100 and 1e-06, respectively.

2) Results for Validation and Test Set: Similar to the result in the first implementation, all the performance metrics are given for validation set in table IV and accuracies are given for test set in table V. With summarized features, the results are improved significantly. All methods gave positive MCC. The accuracy values for test set are higher than 90% except linear SVM application. Among them, non-linear SVM performed best by predicting all the values correctly.

TABLE IV: Results for Validation Set using s-LOO

	Accuracy	Specifity	Sensitivity	MCC
Logistic Regression	72.5%	75%	70%	0.45
KNN	65%	65%	60%	0.25
Linear SVM	77.5%	70%	85%	0.55
Non Linear SVM	75%	70%	72.5%	0.45

TABLE V: Results for Test Set using s-LOO

	Accuracy
Logistic Regression	92.5%
KNN	92.8%
Non Linear SVM	100%

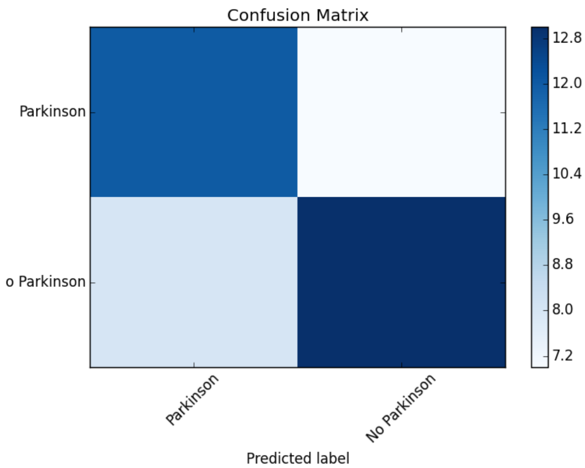

Fig. 7: Confusion Matrix for KNN for s-LOO

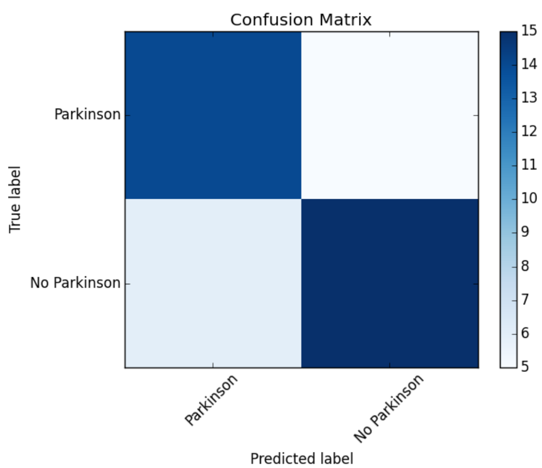

Fig. 8: Confusion Matrix for Non Linear SVM for s-LOO

C. Selection of Best Voice Samples

As already discussed, the voice samples collected in training dataset contains sustained vowel "a", "o", "u", numbers from 1 to 10, short sentences and words. We are also interested in finding what the best voice sample to collect should be when we have limited time to collect voice samples. In the table VI, we present the best three voice sample which gives us highest accuracy with non linear SVM that can only be used instead of all 26 voice samples to predict the prevalence of Parkinson's Disease in patients. Considering the voice samples are collected in Turkish and in Turkish "four" and "ten" includes "o", this result shows the importance of vowel "o" for the analysis.

TABLE VI: Results for Test Set using s-LOO

	Accuracy
Sustained "o"	62.5%
four	62.5%
ten	57.5%

VI. Discussion and Conclusion

In this study, we explored the different classification algorithms which can be used to detect Parkinson's disease using voice sample of subjects. First, we used the dataset available on UCI [8], then we created a summarized version of the dataset to decrease the variations between the samples collected from the same subject. We observed that we obtained higher accuracy with the summarized dataset as presented in the original paper [7]. Additionally, other results that we obtained conforms to the results presented in the paper. We also observed that non linear SVM performed best in both implementations (LOSO and s-LOO).

The authors of the original paper did not test their models on the test set even though they provided it. It might be because of data being biased towards only one class. So, we have performed all our main analysis on the validation set using leave-one-out cross validation technique. Additionally we test the performance of the best model on the test set, and it turns out that our best classifier i.e. non-linear SVM can predict all the examples correctly. Further, logistic regression and KNN gives us 92 % accuracy on test Set. It is also worth noting that the same two patients are misclassified with logistic regression and KNN. It is probably due to noise in the sample recordings of these two patients. But SVM was still robust to classify even those two misclassified examples correctly.

Statement Of Contribution

We hereby state that all the work presented in this report is that of the authors.

References

[1] "Speech and communication," http://www.pdf.org/en/speech_problems_pd, accessed: 2015-12-10.

[2] J. Jankovic, "Parkinson?s disease: clinical features and diagnosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[3] B. Harel, M. Cannizzaro, and P. J. Snyder, "Variability in fundamental frequency during speech in prodromal and incipient parkinson's disease: A longitudinal case study," *Brain and cognition*, vol. 56, no. 1, pp. 24–29, 2004.

[4] L. Hartelius and P. Svensson, "Speech and swallowing symptoms associated with parkinson?s disease and multiple sclerosis: a survey," *Folia Phoniatrica et Logopaedica*, vol. 46, no. 1, pp. 9–17, 1994.

[5] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," *Biomedical Engineering, IEEE Transactions on*, vol. 56, no. 4, pp. 1015–1022, 2009.

[6] A. Tsanas, M. A. Little, P. E. McSharry, B. K. Scanlon, and S. Papapetropoulos, "Statistical analysis and mapping of the unified parkinson?s disease rating scale to hoehn and yahr staging," *Parkinsonism & related disorders*, vol. 18, no. 5, pp. 697–699, 2012.

[7] B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun, "Collection and analysis of a parkinson speech dataset with multiple types of sound recordings," *Biomedical and Health Informatics, IEEE Journal of*, vol. 17, no. 4, pp. 828–834, 2013.

[8] “Parkinson speech dataset,” https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with++Multiple+Types+of+Sound+Recordings, accessed: 2015-11-24.
[9] J. B. Alonso, J. De Leon, I. Alonso, and M. A. Ferrer, “Automatic detection of pathologies in the voice by hos based parameters,” *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 275–284, 2001.
[10] M. Little, P. McSharry, I. Moroz, and S. Roberts, “Nonlinear, biophysically-informed speech pathology detection,” in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 2. IEEE, 2006, pp. II–II.
[11] B. Boyanov and S. Hadjitodorov, “Acoustic analysis of pathological voices. a voice analysis system for the screening of laryngeal diseases,” *Engineering in Medicine and Biology Magazine, IEEE*, vol. 16, no. 4, pp. 74–82, 1997.
[12] L. Cnockaert, J. Schoentgen, P. Auzou, C. Ozsancak, L. Defebvre, and F. Grenez, “Low-frequency vocal modulations in vowels produced by parkinsonian subjects,” *Speech communication*, vol. 50, no. 4, pp. 288–300, 2008.
[13] D. A. Rahn, M. Chou, J. J. Jiang, and Y. Zhang, “Phonatory impairment in parkinson’s disease: evidence from nonlinear dynamic analysis and perturbation analysis,” *Journal of Voice*, vol. 21, no. 1, pp. 64–71, 2007.
[14] A. Tsanas, M. Little, P. E. McSharry, J. Spielman, L. O. Ramig *et al.*, “Novel speech signal processing algorithms for high-accuracy classification of parkinson’s disease,” *Biomedical Engineering, IEEE Transactions on*, vol. 59, no. 5, pp. 1264–1271, 2012.
[15] A. Tsanas, M. Little, P. E. McSharry, L. O. Ramig *et al.*, “Accurate telemonitoring of parkinson’s disease progression by noninvasive speech tests,” *Biomedical Engineering, IEEE Transactions on*, vol. 57, no. 4, pp. 884–893, 2010.
[16] S. O’Sullivan and T. Schmitz, “Parkinson disease,” *Physical Rehabilitation,*, pp. 856–894, 2007.
[17] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
[18] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, “The elements of statistical learning: data mining, inference and prediction,” *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2005.
[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

APPENDIX A
EXTRA FIGURES

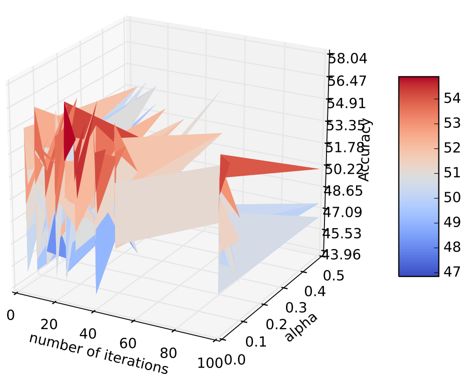

Fig. 9: Cross Validation Results of Logistic Regression for LOSO

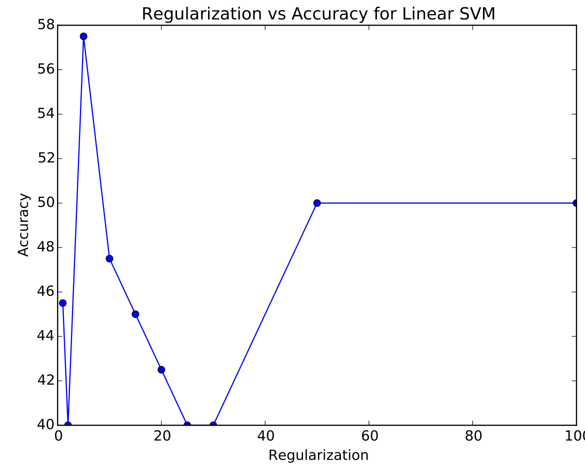

Fig. 10: Cross Validation Results of linear SVM for LOSO

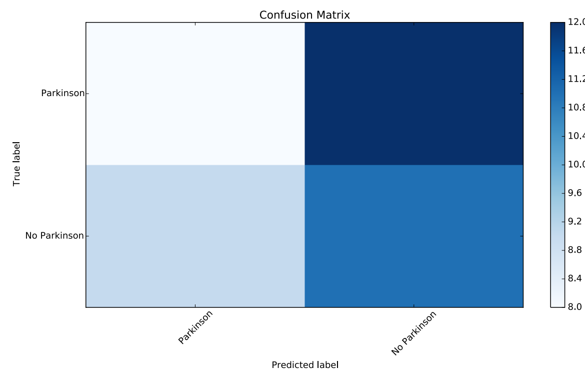

Fig. 11: Confusion Matrix for Logistic Regression for LOSO

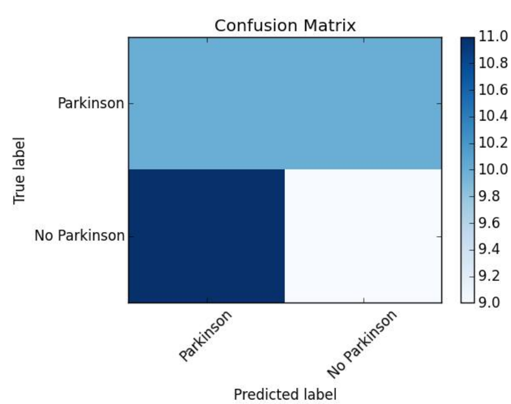

Fig. 12: Confusion Matrix for Linear SVM for LOSO

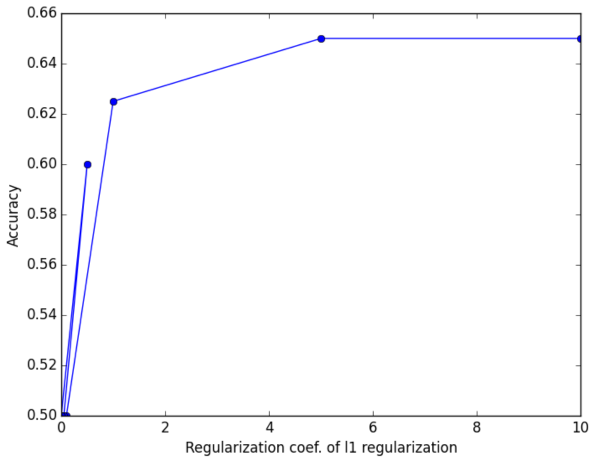

Fig. 13: Cross Validation Results of Logistic Regression with L1 Regularization for s-LOO

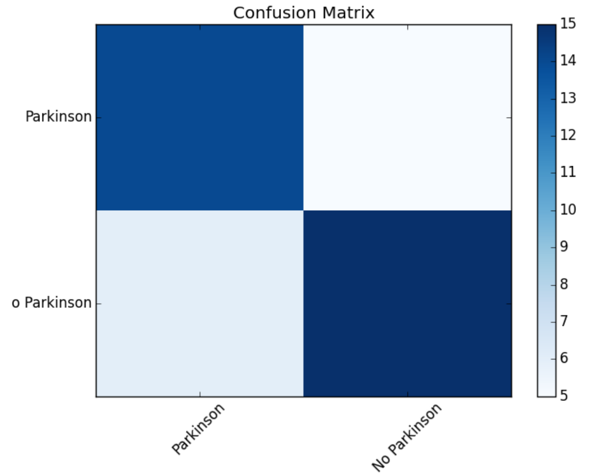

Fig. 15: Confusion Matrix for Logistic Regression for s-LOO

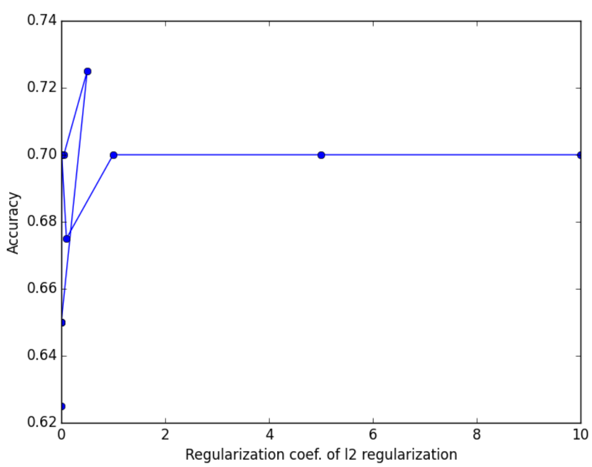

Fig. 14: Cross Validation Results of Logistic Regression with L2 Regularization for s-LOO

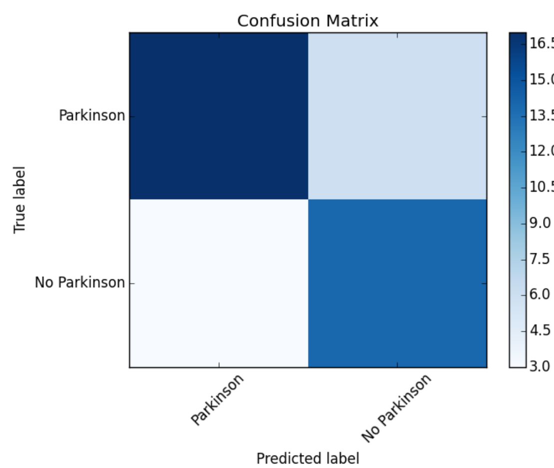

Fig. 16: Confusion Matrix for Non Linear SVM for s-LOO