Comparative Analysis of Machine Learning Models for Parkinson's Disease with PCA

Jinghua Sun

Background

There has been an increased interest in speech pattern analysis applications of Parkinsonism for building predictive telediagnosis and telemonitoring models. For this purpose, we have collected a wide variety of voice samples, including sustained vowels, words, and sentences compiled from a set of speaking exercises for people with Parkinson's disease. Main issues in learning from such a dataset that consists of multiple speech recordings per subject is that How predictive these various types, e.g., sustained vowels versus words, of voice samples are in Parkinson's disease (PD) diagnosis?

Motivation

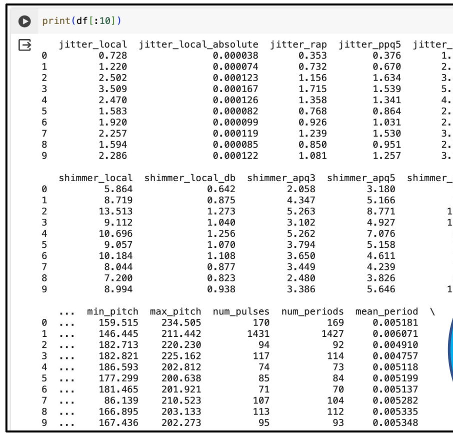
Parkinson's Disease (PD) is a neurodegenerative disorder affecting millions of people worldwide. Early and accurate diagnosis is crucial for effective management and intervention. Machine learning models have shown promise in aiding the diagnosis of PD based on various clinical features. This study aims to explore the impact of dimensionality reduction, specifically through Principal Component Analysis (PCA), on the performance of different machine learning models for predicting Parkinson's Disease.

Surprisingly, out results indicated a decrease in accuracy after PCA.

Reference

*[1]Sakar, B. E., Isenkul, M. E., Sakar, C. O., Sertbas, A., Gurgen, F., Delil, S., Apaydin, H., & Kursun, O. (2013). Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics, 17*(4), 828–834.*

Dataset & Machine Learning Models



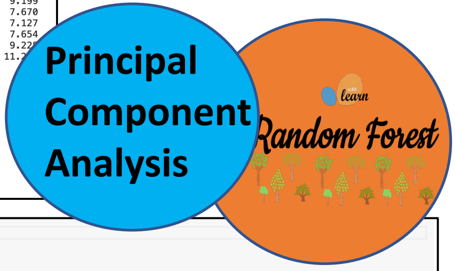
```
print(df[:10])
```

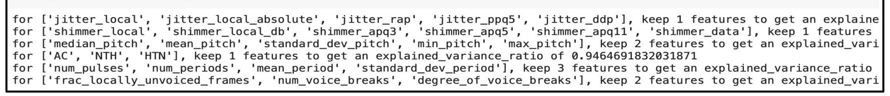
```
   jitter_local  jitter_local_absolute  jitter_rap  jitter_ppq5  jitter_ddp  \
0         0.728               0.000038       0.353        0.376       1.059
1         1.220               0.000074       0.732        0.670       2.196
2         2.502               0.000123       1.156        1.634       3.469
3         3.509               0.000167       1.715        1.539       5.145
4         2.470               0.000126       1.358        1.341       4.075
5         1.583               0.000082       0.768        0.864       2.303
6         1.920               0.000099       0.926        1.031       2.777
7         2.257               0.000119       1.239        1.530       3.716
8         1.594               0.000085       0.850        0.951       2.550
9         2.286               0.000122       1.081        1.257       3.243

   shimmer_local  shimmer_local_db  shimmer_apq3  shimmer_apq5  shimmer_apq11  \
0          5.864             0.642         2.058         3.180          7.194
1          8.719             0.875         4.347         5.166          7.548
2         13.513             1.273         5.263         8.771         16.779
3          9.112             1.040         3.102         4.927         12.823
4         10.696             1.256         5.262         7.076          9.199
5          9.057             1.070         3.794         5.158          7.670
6         10.184             1.108         3.650         4.611          7.127
7          8.044             0.877         3.449         4.239          7.654
8          7.200             0.823         2.480         3.826          9.22
9          8.994             0.938         3.386         5.646          11.2

   ...  min_pitch  max_pitch  num_pulses  num_periods  mean_period  \
0  ...    159.515    234.505         170          169     0.005181
1  ...    146.445    211.442        1431         1427     0.006071
2  ...    182.713    220.230          94           92     0.004910
3  ...    182.821    225.162         117          114     0.004757
4  ...    186.593    202.812          74           73     0.005118
5  ...    177.299    200.638          85           84     0.005199
6  ...    181.465    201.921          71           70     0.005137
7  ...     86.139    210.523         107          104     0.005282
8  ...    166.895    203.133         113          112     0.005335
9  ...    167.436    202.273          95           93     0.005348
```

1.Jitter features measure variations in pitch periods

2.Shimmer features quantify variations in amplitude and the stability of the voice.

3.'UPDRS': Unified Parkinson's Disease Rating Scale

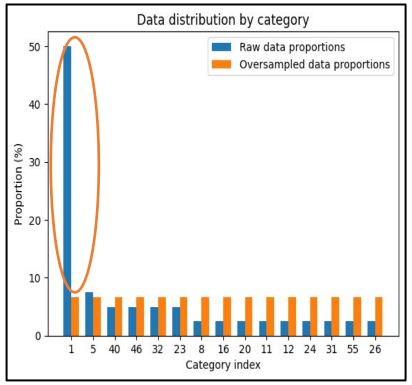




```
dfs_pca = []
for gp in gps:
  dfs_pca.append(pca(df[gp]).copy())
```

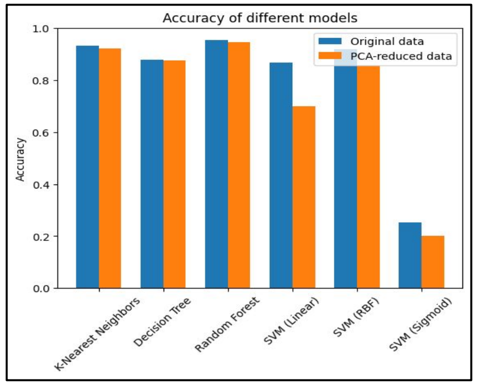
```
for ['jitter_local', 'jitter_local_absolute', 'jitter_rap', 'jitter_ppq5', 'jitter_ddp'], keep 1 features to get an explaine
for ['shimmer_local', 'shimmer_local_db', 'shimmer_apq3', 'shimmer_apq5', 'shimmer_apq11', 'shimmer_data'], keep 1 features
for ['median_pitch', 'mean_pitch', 'standard_dev_pitch', 'min_pitch', 'max_pitch'], keep 2 features to get an explained_vari
for ['AC', 'NTH', 'HTN'], keep 1 features to get an explained_variance_ratio of 0.9464691832031871
for ['num_pulses', 'num_periods', 'mean_period', 'standard_dev_period'], keep 3 features to get an explained_variance_ratio
for ['frac_locally_unvoiced_frames', 'num_voice_breaks', 'degree_of_voice_breaks'], keep 2 features to get an explained_vari
```

Analysis Results



Original dataset is imbalanced, the proportion of 1 took nearly **50%** which is much more than other levels

Applied PCA Algo to select the **most important 10 features** which distributed evenly into each category



1.Overall accuracy **decreased** after PCA

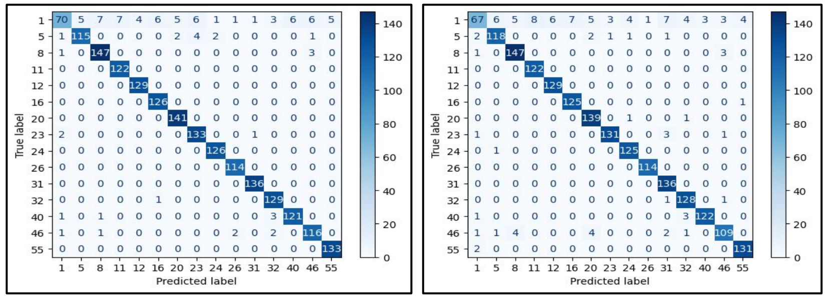
2.**Random forest** predict most accurately both with and without PCA

3. **SVM(RBF)** predicted second most accurately both with and without PCA

4.The least accuracy is predicted by **SVM(SIG)** both with and without PCA

Interpretation of Results

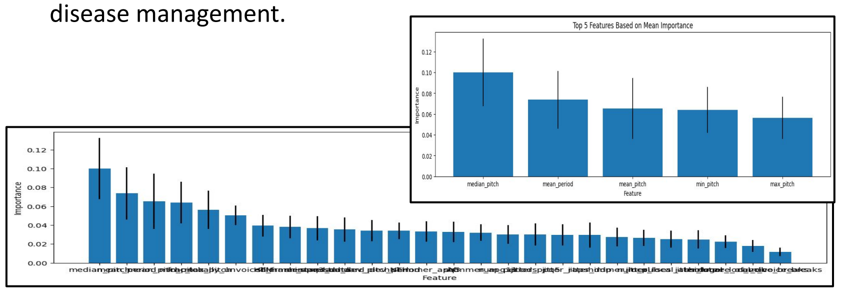
Random Forest



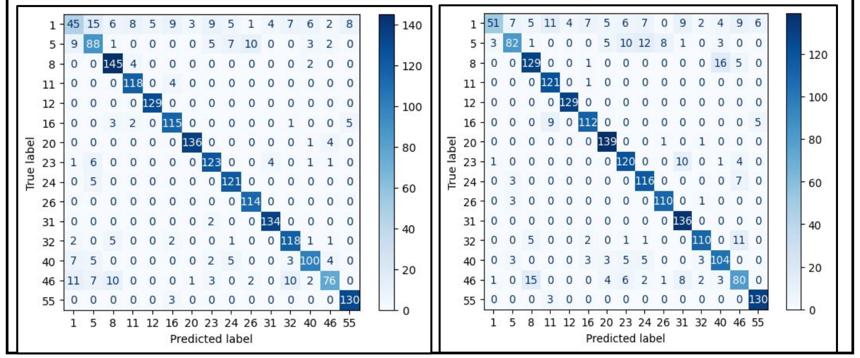
TOP 5 Importance Features

MEDIAN_PITCH,MEAN_PERIOD,MEAN_PITCH,MIN_PITCH,MAX_PITCH

- **MEDIAN_PITCH :** Could be linked to vocal fold dynamics, muscle control, or other factors associated with Parkinson's disease.
- **MEAN_PERIOD :** Related to voice quality and potential motor control issues associated with Parkinson's disease.
- **Healthcare Professionals:** Decision-making & Treatment planning.
- **Researchers:** Speech as a biomarker for neurological conditions.
- **Patients:** A convenient and regular monitoring method, improving disease management.



SVM-RBF


True label
Predicted label
1: 45 15 6 8 5 9 3 9 5 1 4 7 6 2 8
5: 9 88 1 0 0 0 0 5 7 10 0 0 3 2 0
8: 0 0 145 4 0 0 0 0 0 0 0 0 2 0 0
11: 0 0 0 118 0 4 0 0 0 0 0 0 0 0 0
12: 0 0 0 0 129 0 0 0 0 0 0 0 0 0 0
16: 0 0 3 2 0 115 0 0 0 0 0 1 0 0 5
20: 0 0 0 0 0 0 136 0 0 0 0 0 1 4 0
23: 1 6 0 0 0 0 0 123 0 0 4 0 1 1 0
24: 0 5 0 0 0 0 0 0 121 0 0 0 0 0 0
26: 0 0 0 0 0 0 0 0 0 114 0 0 0 0 0
31: 0 0 0 0 0 0 0 2 0 0 134 0 0 0 0
32: 2 0 5 0 0 2 0 0 1 0 0 118 1 1 0
40: 7 5 0 0 0 0 0 2 5 0 0 3 100 4 0
46: 11 7 10 0 0 0 1 3 0 2 0 10 2 76 0
55: 0 0 0 0 0 3 0 0 0 0 0 0 0 0 130
True label
Predicted label
1: 51 7 5 11 4 7 5 6 7 0 9 2 4 9 6
5: 3 82 1 0 0 0 5 10 12 8 1 0 3 0 0
8: 0 0 129 0 0 1 0 0 0 0 0 0 16 5 0
11: 0 0 0 121 0 1 0 0 0 0 0 0 0 0 0
12: 0 0 0 0 129 0 0 0 0 0 0 0 0 0 0
16: 0 0 0 9 0 112 0 0 0 0 0 0 0 0 5
20: 0 0 0 0 0 0 139 0 0 1 0 1 0 0 0
23: 1 0 0 0 0 0 0 120 0 0 10 0 1 4 0
24: 0 3 0 0 0 0 0 0 116 0 0 0 0 7 0
26: 0 3 0 0 0 0 0 0 0 110 0 1 0 0 0
31: 0 0 0 0 0 0 0 0 0 0 136 0 0 0 0
32: 0 0 5 0 0 2 0 1 1 0 0 110 0 11 0
40: 0 3 0 0 0 3 3 5 5 0 0 3 104 0 0
46: 1 0 15 0 0 0 4 6 2 1 8 2 3 80 0
55: 0 0 0 3 0 0 0 0 0 0 0 0 0 0 130