Improving Accuracy in Noninvasive Telemonitoring of Progression of Parkinson's Disease using Two-Step Predictive Model

Shreya Jain[1], Dr. Sujala Shetty[1], Yash Patni[2]

[1]Department of Computer Science Engineering, BITS Pilani Dubai Campus

[2]Department of Mechanical Engineering, BITS Pilani Dubai Campus

***Abstract*— Parkinson's disease has affected over 6.3 million people across the globe. It is estimated that by 2030, the number would rise to 9 million. Almost twenty percent of the people still remain undiagnosed. Parkinson's is the second most common neurodegenerative disease after Alzheimer's. It not only claims the lives of the patients suffering from it but also adversely impacts the lives of their loved ones. A lot of research is being conducted to find modern medical techniques to tackle the ill effects of the disease. Monitoring the progression of the disease plays a vital role in controlling its various symptoms.**

Non-conventional ways of monitoring PD(Parkinson's Disease) provide an edge over the existing techniques as it reduces the financial burden and also limits the number of clinical visits required for it. In this research paper, we aim to build a predictive model that accurately predicts the UPDRS (Unified Parkinson's Disease Rating Scale) of patients using the data collected through noninvasive speech tests. The research hopes to propose a more efficient technique to monitor Parkinson's disease leading to beneficial treatment of the patients.

Keywords—Parkinson's Disease; telemonitoring; speech test; machine learning; classification; regression;

I. INTRODUCTION

A. Parkinson's Disease

Parkinson's disease has affected the lives of many and will continue to affect the lives of many more. It has been estimated that about $27 billion a year are spent annually in medical bills for treatment and management of many of PD's symptoms[1]. PD worsens with time, as it becomes more disabling as it progresses. The patients find it challenging to perform daily activities such as moving across the room, rising from a couch, etc. Sometimes, patients end up on a wheelchair or also bedridden.

A person is believed to be suffering from Parkinson's disease when there is malfunction or death of vital nerve cells in the brain, which produce a chemical called dopamine. When the production of dopamine has reduced or stopped, the signal to move does not get communicated. By the time a person starts to experience motor symptoms of Parkinson's they've already lost approximately 80% of their dopamine producing cells[1]. People may experience non-motor symptoms from loss of other neurotransmitters up to ten years before motor symptoms are noticed. There is currently no cure for Parkinson's. The doctor's goal will be to treat the symptoms to keep your quality of life as high as possible. It is also reported by many that they suffered from stiffness, slowing of movement, your face may show little or no expression, speech impairment, etc. The associated symptoms of PD only deteriorate with time. It has recently discovered that speech impairment is one the first symptoms of PD to show in a patient if observed very carefully.

The medical institutes, hospitals and clinics keeps records of all their cases as they realize the importance and vast applications related to data. To survey these data and to obtain useful results and patterns in relation to disease is one of the objectives of the use of these data. Collected data volume is very high and we must use machine learning algorithms to predict desired results or unknown patterns among massive volume of data. This paper focuses to establish a relationship between speech signal properties and UPDRS. We show that this method leads to clinically useful UPDRS estimation, and demonstrate remote PD monitoring on a weekly basis, tracking UPDRS fluctuations for a six-month period[2]. This can be a useful guide for clinical staff, following the progression of clinical PD symptoms on a regular basis, tracking the UPDRS that would be obtained by a subjective clinical rater.

Machine Learning algorithms tend to simplify the prediction segment. New advanced tools introduced recently have made it possible to collect and process large volumes of medical data possible. This paper presents the potential synergies between machine learning algorithms and progression of PD

II. OBJECTIVES OF THE PAPER

- Summaries the disease – symptoms, risk factors, telemonitoring, rating scale.

- Predictions to be done with more accuracy than the researches have done before this. In order to do that, a two-step algorithm was implemented and its accuracy was compared with prediction using only regression.

III. Overview of Parkinson's Disease

Parkinson's disease is a progressive, neurological disease that mainly affects movement but can also affect cognition. Its symptoms continue and worsen over time. Different parts of the brain work together by sending signals to each other to coordinate all of our thoughts, movements, emotions, and senses. Parkinson's primarily affects neurons in an area of the brain called the substantia nigra. Some of these dying neurons produce dopamine, a chemical that sends messages to the part of the brain that controls movement and coordination [2]. As PD progresses, the amount of dopamine produced in the brain decreases, leaving a person unable to control movement normally.

A. Symptoms of Parkinson's Disease

Among the various symptoms of PD, the most common ones experience by the patients are –

- Tremor
- Rigidity
- Slowness of Movement
- Postural Instability
- Speech Impairment

B. Risk Factors

Non-Preventable Risks

- **Age** – Over 95% of the cases that are reported in PD, the recorded age is above 60. Very rarely cases have been reported in the early ages. Some researchers believe that the neural damage associated with Parkinson's worsens with age[5].
- **Family History** – Cause of the disease is not known as of now. Still it has been observed that the patients with family history of the disease or any other neurodegenerative disease are more susceptible to suffer from it.
- **Genetic Factor** – Even though the researchers have not been able to pin point the exact gene that triggers the disease, it is evident that genetic mutations is one of the reasons for the disease.
- **Gender** – Males are more likely to get Parkinson's than females. Possible reasons for this may be that males have greater exposure to other risk factors such as toxin exposure or head trauma. It is statistically observed that males are more prone to the disease than females. It is due to males have greater exposure to other risk factors such as toxin exposure or head trauma.

Preventable Risks

- **Toxic Exposure** – Ongoing exposure to herbicides and pesticides may put one at a slightly higher risk of getting Parkinson's disease.
- **Some Medicines** – Certain medications such as antipsychotics used to treat severe paranoia and schizophrenia can cause Parkinsonism (symptoms that resemble Parkinson's disease)

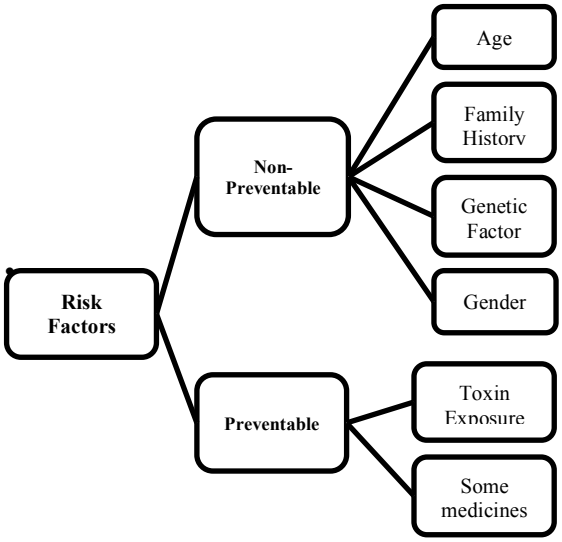

Figure 1: Risks of Parkinson's Disease

C. Noninvasive Telemonitoring

Noninvasive telemonitioring has emerged as an ideal option over the convention mode of monitoring PD. It brings down the cost of the overall treatment associated with the disease substantially[3]. At the same time it eliminates the need of frequent and inconvenient visits to the clinic. Eventually the adoption of the noninvasive speech tests would lead to relieving workload on healthcare professionals and overall increase in clinical evaluation of the patients.

D. Rating Scale

In almost all patients of Parkinson's disease the symptoms unfold in a different way and the severity of the symptoms differs significantly[4]. In order to uniformly monitor the progression of the disease, a universal rating scale is needed. After a series of physical examinations by trained healthcare professionals, the test observations are mapped to a metric specifically designed to follow disease progress, known as the Unified Parkinson's Disease Rating Scale (UPDRS), which reflects the presence and severity of symptoms. Other scales used for the same purpose are Yahr Scale, UPDRS, Schwab and England ADL Scale, etc. In UPDRS, patients are assigned a value in the range of 0-176, where 0 denotes healthy individual and 176 total disability. This method has a lot of potential due to its cost effectiveness, reducing the frequent and inconvenient clinic visits.

IV. Overview of Machine Learning Algorithms

A. Classification

C4.5 is a decision tree algorithm that can be used to classify whether a PD patient is in Stage 1 or Stage 2. The algorithm is based on the information gain or entropy, which helps decide which attribute of a given instance will best classify the instances in the dataset. It determines information gain for attribute to determine the best attribute to be used a split point. This technique is repeated until a leaf node is created.

Entropy = $\sum - p_i \log_2 p_i$
Where, p_i is the probability of class i

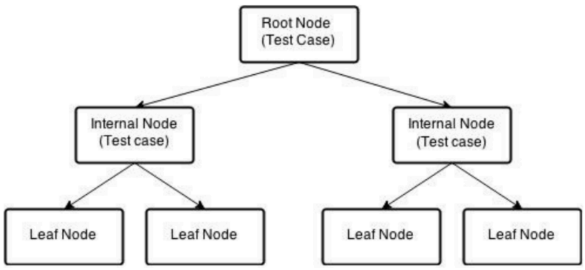

Figure 2: Layout of a Decision Tree

B. Regression

i. Multivariate Linear Regression

Linear regression is an approach that attempts to model the relationship between dependent variable and one or more explanatory independent variables. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. Linear regression is one of the first methods developed to work out regression problems. Hence it can be used to fit a predictive model to training dataset of y and X values. On a test set, then this fitted model can be used to estimate the value of y from the existing value of X.

A linear regression line has an equation of the form

$$\boldsymbol{Y = a + bX}$$

where $\boldsymbol{X}$ represents the vector of explanatory variables and $\boldsymbol{Y}$ is the dependent variable. The slope of the line is $\boldsymbol{b}$, and $\boldsymbol{a}$ is the intercept (the value of $\boldsymbol{y}$ when $\boldsymbol{x} = 0$).

ii. Regression Tree

A regression tree is built using information gain and prunes it using reduced-error pruning (with backfitting). This fast decision tree learner is built in the same manner as the C4.5 but here regression is done, instead of classification. Thus the leaf node of this tree helps assign a numeric value to the UPDRS.

iii. K- Nearest Neighbor

K nearest neighbors is an algorithm that predicts the numerical target based on a similarity measure (e.g., distance functions). A simple implementation of KNN regression is to calculate the average of the numerical target of the K nearest neighbors. KNN regression uses the same distance functions as KNN classification.

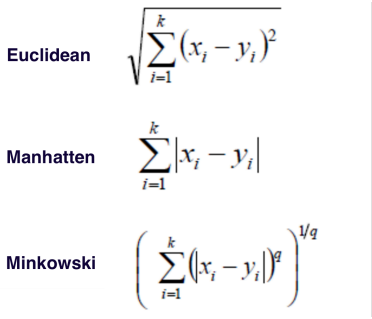

Figure 3: Distance Functions

iv. Multilayer Perceptron Neural Networks

Multilayer Perceptron Neural Networks is an ANN that uses back propagation for training the network. This network takes a set of input values that it maps to appropriate outputs. An MLP contains multiple hidden layers of nodes in a directed and each layer is fully connected to the next one. Each node except the input nodes, is a neuron (processing element) with a non linear activation energy associated with it.

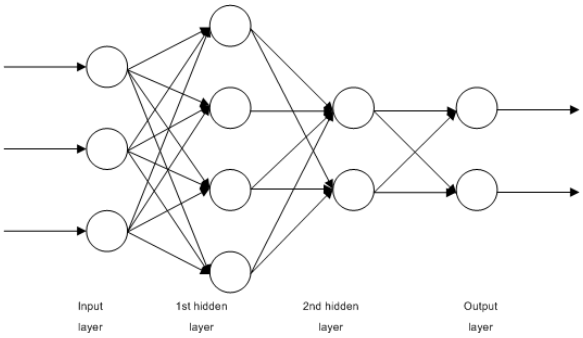

Figure 4: Layout of a Neural Network

V. Methodologies

A. Dataset Description

The dataset of telemonitoring of Parkinson's disease used for the research was obtained from UCI machine learning archive[5]. It was created by Athanasios Tsanas and Max Little of the University of Oxford, in collaboration with 6 medical centers in the US and Intel Corporation which developed the AHTD telemonitoring device to record the speech signals. The dataset is a compilation of recordings

from 42 patients (28 men and rest women) diagnosed with early stage PD. These patients were made to record speech of patients while they sustain the vowel sound. These attributes include subject number, subject age, subject gender, time interval from baseline recruitment data, motor-UPDRS, total-UPDRS, and 16 biomedical voice measures (vocal features)[7].

B. Pre-processing of the Dataset

When data is gathered, the process often leads to collection of some faulty data or some missing values. Some times data collected has out of the range values (e.g., Income: -100) and even impossible data combinations (e.g. Sex: Male, Pregnant: Yes), missing values and many more. Source of these errors could be human error, or mismanagements of the database. Thus, data collected must be filtered for such errors before any analysis is done on it [8]. In the dataset used there were no missing values, many attributes though needed normalization and conversion into the right data type etc. But whenever there are missing values, the following ways can be adopted to remove them. In case the class attribute is missing the record must be removed. In case other attributes are missing either it is replaced using global estimation or local estimation.

Table 1: Dataset description

Dataset	**No. of Attributes**	**No. of Instances**
Parkinsons Telemonitoring Dataset	22	5875

The attribute types sex and age by default were numerical as statistics are being calculated for it[6]. So these attributes must be converted to nominal type first. Also the subject attribute is removed to generalize the result for various patients. The numeric attributes were normalized using the min-max normalization to get the new value from 0 to 1.

Table 2: Description of the Attributes

Attributes	Description
Subject	Integer that uniquely identifies each subject
Age	Subject age
Gender	Subject Sex '0' – Male,'1' - Female
Test_time	Time since in trial
Jitter	Measure of variation in fundamental frequency - Speech
Shimmer	Measure of variation in Amplitude - Speech
NHR	Noise to Harmonics Ratio
HNR	Harmonics to Noise Ratio
RPDE	Recurrence Period Density Entropy
DFA	Detrended Fluctuation Analysis
PPE	Pitch Period Entropy

C. Visualisation Using Scatterplot

Visualizing tools like scatterplot can be a helpful tool in determining the strength of the relationship between two variables. A scatterplot successfully plots all the variables relationships with each other. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. Thus, a scatter plot is a good way of feature selection for the predictive models to be applied.

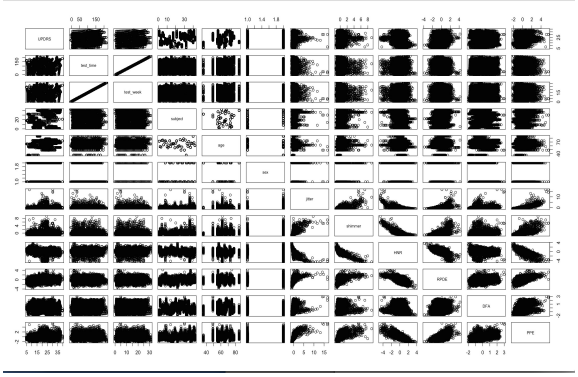

Figure 5: Scatter Plot

D. Creating a Classification

Before proceeding with regression, we first tried to classify each patient's stage in Parkinson's disease. The stages of the disease are defined using the UPDRS.
The stages of the disease are -
Stage 1 : 0-34
Stage 2 : 35-69
Stage 3 : 70-104
Stage 4 : 105-139
Stage 5 : 140 and above

The dataset used was classified in only two stages as the dataset did not cover the entire range of UPDRS. This class attribute was made easily using the above mentioned ranges. The class attribute was then converted to nominal. And to understand the relationships each attribute was plotted against the class attribute. As the dataset had only early stage PD patients, we only had Stage 1 and Stage 2 in the created class attribute.

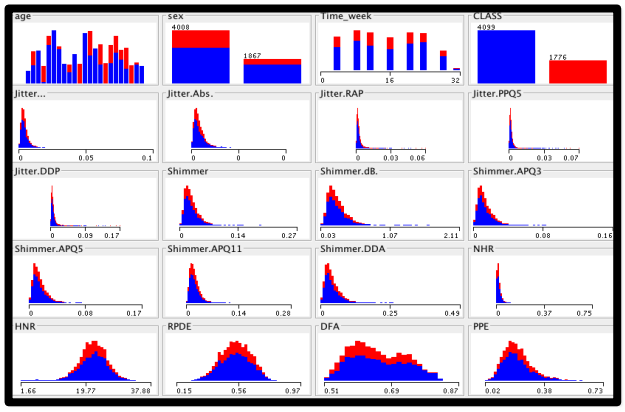

Figure 3: Attributes Vs. Created Class

VI. Proposed Model

In this paper, we propose a two-step predictive model that uses the dataset and first classifies the stage of the disease then use the predicted class as an attribute to predict the UPDRS. This paper further compares the UPDRS directly using regression as well as the proposed two-step algorithm.

INPUT : Parkinson's Telemonitoring Dataset

OUTPUT : The algorithm to predict the UPDRS of patients.

PROCEDURE:

1. The dataset is pre-processed using WEKA tools. Following operations are performed on the dataset
    a. Replacing missing values
    b. Normalization of values
    c. Conversion Test_time by dividing it by 7 to give the Test_week
    d. Conversion of attribute type
2. Processed dataset is passed through feature selection where attributes with less significance are deleted.
3. From UPDRS values, a class is created which gives stage of PD. After the pre-processing is done the dataset is then uploaded in WEKA.
4. The dataset first employs Linear Regression, Regression Tree, MLP and KNN Classifier to directly give the total UPDRS.
5. Then dataset is used then used to apply C4.5 to obtain stage. Using stage, another regression model is made using the four above algorithms.

10- Fold Cross validation technique is used to in all the applied algorithms. The results are compared on the basis of correctly classify instances, to figure out which methodology 4 or 5 is giving better results. The results also help us to compare the various regression techniques to figure out the most effective model [6].

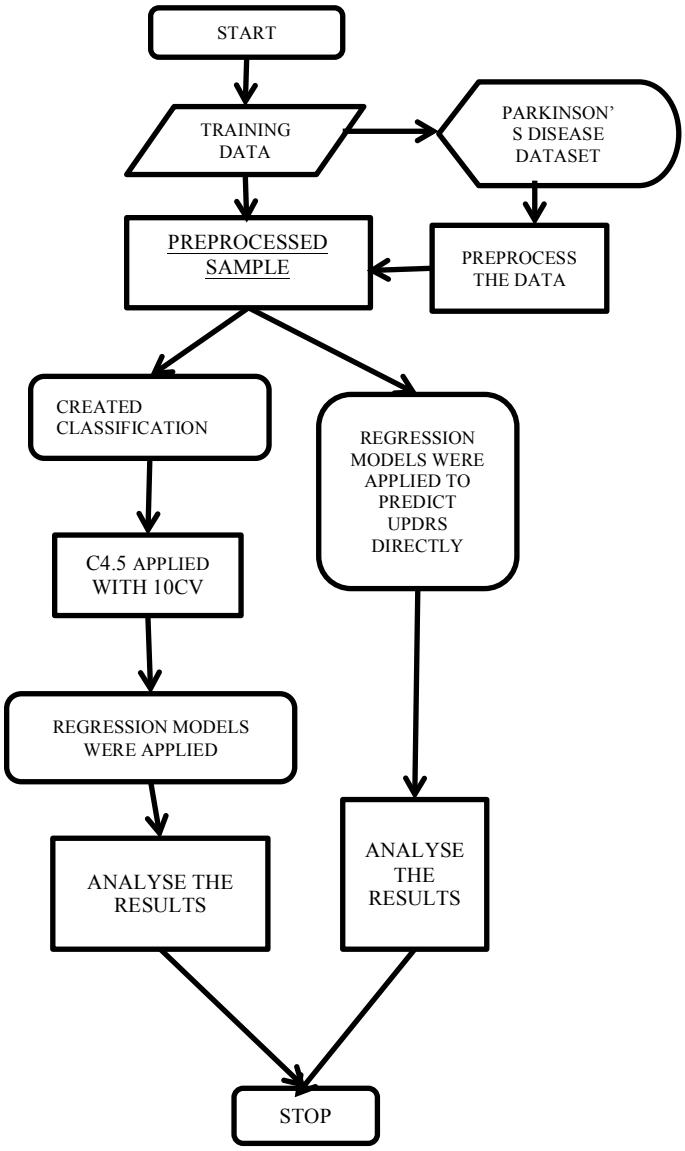

Figure 6: Flowchart of Algorithm

VII. Results and Analysis

A. *Results*

i. Classification

Decision Tree C4.5 is used for classification of the stage of Parkinson's disease. It is called J48 in the WEKA toolkit. It is used with a 10-fold cross validation technique.

Table 2: Performance results from Decision Tree C4.5

	No. of Instances	**Percentage**
Correctly Classified Instances	5837	99.3532%
Incorrectly Classified Instances	38	0.6468%

Table 3: Other results from Decision Tree C4.5

Kappa Statistic	0.9846
Mean Absolute Error	0.0077
Root Mean Squared Error	0.0788
Relative Absolute Error	1.8146%

Table 4: Confusion Matrix for C4.5

	A - tested_positive	B - tested_positive
A - tested_positive	4088 (i)	11(ii)
B - tested_positive	27(iii)	1749(iv)

In the above confusion matrix, the values represent following:
i. : Number of correct predictions that the instance tested positive
ii. : Number of incorrect predictions that the instance tested negative
iii. : Number of incorrect predictions that the instance tested positive
iv. : Number of correct predictions that the instance tested negative

ii. Regression Model

Table 5: Performance results for Regression Models

	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error
Linear Regression	4.2138	5.4468	48.6733 %	50.901 %
	2.8501	3.6817	32.9221 %	34.405%
Regression Tree	0.6623	1.8531	7.6502 %	17.3172 %
	0.514	1.4128	6.1384 %	13.2031%
Multilayer Preceptron	1.5571	2.283	17.9861 %	21.3344 %
Neural Network	1.2844	1.8589	14.8356 %	17.3715 %
K- Nearest	0.5806	2.0727	6.7061 %	19.37 %
Neighbors	0.4311	1.1997	4.9791 %	11.211%

Parameters in Test Statistic

- Kappa Statistic: It is a measure that compares Observed Accuracy with Expected Accuracy
- Mean Absolute Error: Average of the absolute error between the original and predicted value.
- Root Mean Squared Error: Measure of the differences between value predicted by a model or an estimator and the original values.
- Relative Absolute Error: Ratio of the absolute error of the measurement to the accepted measurement

B. Analysis

The C4.5 classification technique is used to classify the stage. It predicts the stage correctly with an accuracy of about 99.35%. Kappa Statistic is close to 1 indicating that actual and expected stage are mostly the same in the test dataset.

Further, it can be clearly seen from the results that the regression model is performing better after classification has been applied. The root mean squared error has reduced for all regression models after the classification step. Linear regression it reduced from 5.44 to 3.68. This has the maximum error. It is also observed that K- Nearest Neighbor gives the least error among regression models. KNN Classifier has Mean Absolute Error as 0.4311, RMSE as 1.19 and the Relative Absolute Error as 11.2%.

VIII. Conclusion

The telemonitoring of progression of PD is an important medical problem that needs to be addressed at the earliest [9]. Constant progression of the disease can help us in controlling its many unmanageable symptoms. The noninvasive speech tests help to make it less of a financial burden on the family as well as reduce the number of clinical visits. This paper shows the relationship between speech properties and UPDRS as an issue worthy of further investigation. Here we present a method that first computes the stage of the disease using classification algorithm, which act as a feature for statistical regression techniques. We show that this method leads to clinically useful UPDRS estimation, and demonstrate remote PD monitoring on a weekly basis, tracking UPDRS fluctuations for a six-month period.

In future it is planned to collect more data from other medical institutes to test the proposed model. The scope of the research was limited due to access to public datasets. Thus the paper can be extended with more data of patients with later stage PD as well. Further collection of data could probably help us identify new prognostic elements to be incorporated.

References

[1] About Parkinson's Disease." Facts about Parkinson's Disease Neurological Movement Disorder. Web. 17 May 2014.

[2] Little, M.A., McSharry, P.E., Hunter, E.J., Spielman, J., and Ramig, L.O. (2009), Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease in /emphIEEE Transac- tions on Biomedical Engineering, 56(4):1015-1022.

[3] Tsanas, A. Little, M.A., McSharry, P.E., Hunter, E.J., Spielman, J., and Ramig, L.O. (2012), Novel speech signal processing algorithms for high-accuracy classification of Parkinson's Dis- ease *IEEE Transactions on Biomedical Engineering*, 59(5):1264- 1271.

[4] Tsanas, A. Little, M.A., McSharry, and Ramig, L.O. (2010). Accurate Telemonitoring of Parkinsons Disease Progression by Noninvasive

Speech Tests in *IEEE Transactions on Biomedical Engineering*, 47(4):884-893.

[5] A. Elbaz, J. H. Bower, D. M. Maraganore, S. K. McDonnell, B. J. Peterson, J. E. Ahlskog, D. J. Schaid, and W. A. Rocca, "Risk tables for parkinsonism and Parkinson's disease," *J. Clin. Epidemiol.*, vol. 55, pp. 25–31, 2002.

[6] Ömer Eskidere, Figen Ertaş, Cemal Hanilçi, A comparison of regression methods for remote tracking of Parkinson's disease progression, Expert Systems with Applications, Volume 39, Issue 5, April 2012, Pages 5523-5528, ISSN 0957-4174.

[7] M. C. de Rijk, L. J. Launer, K. Berger, M. M. Breteler, J. F. Dartigues, M. Baldereschi, L. Fratiglioni, A. Lobo, J. Martinez-Lage, C. Trenkwalder, and A. Hofman, "Prevalence of Parkinson's disease in Europe: A collabo- rative study of population-based cohorts," *Neurology*, vol. 54, pp. 21–23, 2000.

[8] M. Rajput, A. Rajput, and A. H. Rajput, "Epidemiology," in *Handbook of Parkinson's Disease,* 4th ed., R. Pahwa and K. E. Lyons, Eds. New York: Informa Healthcare, 2007, ch. 2.

[9] S. K. Van Den Eeden, C. M. Tanner, A. L. Bernstein, R. D. Fross, A. Leimpeter, D. A. Bloch, and L. M. Nelson, "Incidence of Parkinson's disease: Variation by age, gender, and race/ethnicity," *Amer. J. Epidemiol.*, vol. 157, pp. 1015–1022, 2003.