

“A MACHINE LEARNING APPROACH TO DIAGNOSIS OF PARKINSON’S DISEASE”

PROJECT REPORT

Submitted for the course:
CSE3999 - Technical Answers for Real-world Problems (TARP)

By

RONNIT SAMUEL	15BCE0062
VIGNESH VAIDYANATHAN	15BCE0076
VOLETI RAVI	15BCE0082
SAMUDRA PRATIM BORKAKOTI	15BCE0093
KOMAL BAKSHI	15BCE0231
MANAN BARJATYA	15BCE0319
SAHIL MEHTA	15BCE0431
KUNAL SHARMA	15BCE0622
POORVA SINHA	15BCE0986

Slot: TG1

Name of faculty: **Prof. Srimathi C
(SCOPE)**



VIT[®]
UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)

VELLORE ■ CHENNAI

www.vit.ac.in

April, 2018

CERTIFICATE

This is to certify that the project work entitled “**A MACHINE LEARNING APPROACH TO DIAGNOSIS OF PARKINSON’S DISEASE**” that is being submitted by “**Ronnit Samuel, Vignesh Vaidyanathan, Voleti Ravi, Samudra Pratim Borkakoti, Komal Bakshi, Manan Barjatya, Sahil Mehta, Kunal Sharma and Poorva Sinha**” for Technical Answers for Real-world Problems is a record of bonafide work done under my supervision. The contents of this project work, in full or in parts, have neither been taken from any other source nor have been submitted for any other CAL course.

Place: Vellore

Date: 05.04. 18

Signature of Students:

Ronnit Samuel (15BCE0062)

Vignesh Vaidyanathan (15BCE076)

Voleti Ravi (15BCE0082)

Samudra Pratim Borkakoti (15BCE0093)

Komal Bakshi (15BCE0231)

Manan Barjatya (15BCE0319)

Sahil Mehta (15BCE0431)

Kunal Sharma (15BCE0622)

Poorva Sinha (15BCE0986)

Signature of Faculty: **Prof. Srimathi C**

ACKNOWLEDGEMENTS

We take immense pleasure in thanking **Dr. G. Viswanathan**, our beloved Chancellor, VIT University and respected Dean, **SCOPE**, for having permitted us to carry out the project.

We express gratitude to our guide, **Prof. Srimathi C**, for guidance and suggestions that helped us to complete the project on time. Words are inadequate to express our gratitude to the faculty and staff members who encouraged and supported us during the project. Finally, we would like to thank our ever-loving parents for their blessings and our friends for their timely help and support.

Signature of Students:

Ronnit Samuel (15BCE0062)

Vignesh Vaidyanathan (15BCE076)

Voleti Ravi (15BCE0082)

Samudra Pratim Borkakoti (15BCE0093)

Komal Bakshi (15BCE0231)

Manan Barjatya (15BCE0319)

Sahil Mehta (15BCE0431)

Kunal Sharma (15BCE0622)

Poorva Sinha (15BCE0986)

ABSTRACT

We will investigate applications of machine learning algorithms to medical data, adaptations of differences in data collection, and the use of ensemble techniques. Focusing on the binary classification problem of Parkinson's Disease (PD) diagnosis, we will apply machine learning algorithms to a primary dataset consisting of voice recordings from healthy and PD subjects. Specifically, we will use Artificial Neural Networks, Support Vector Machines, and an Ensemble Learning algorithm. Next, we will adapt a secondary regression dataset of PD recordings and combine it with the primary binary classification dataset, testing various techniques to consolidate the data including treating the regression data as unlabelled data in a semi-supervised learning approach. We will determine the performance of the above algorithms on this consolidated dataset. Performance of algorithms will be evaluated using 10-fold cross validation and results will be analysed in a confusion matrix. Accuracy, precision, recall, and F-score will be calculated. The expands on past related work, which has used either a regression dataset alone to predict a Unified Parkinson's Disease Rating Scale score for PD patients, or a classification dataset to determine healthy or PD diagnosis. In past work, the datasets have not been combined, and the regression set has not been used to contribute to evaluation of healthy subjects.

Table of Content

1. INTRODUCTION.....	8
2. BACKGROUND.....	9
2.1 Machine Learning Techniques.....	9
2.1.1 Artificial Neural Networks.....	9
2.1.2 Support Vector Machines.....	9
2.1.3 Ensemble Learners.....	10
2.2 Applications to Medical Data.....	10
2.2.1 Self-reported Input.....	10
2.2.2 Clinical Decision Support Systems.....	10
2.2.3 EEG and EKG Data.....	10
2.3 Diagnosing Parkinson's Disease.....	11
3. METHODS.....	12
3.1 Dataset.....	12
3.2 Unbalanced Dataset.....	12
3.3 Neural Networks.....	14
3.4 Support Vector Machines.....	14
3.5 AdaBoost.....	14
3.6 Combined Datasets.....	14
4. EXECUTION.....	16
4.1 Neural Networks.....	17
4.2 Support Vector Machines.....	17
4.3 AdaBoost.....	18
4.4 Combined Datasets.....	18
5. RESULT.....	20
6. CONCLUSION.....	22
7. FUTURE WORK.....	23

List of Figures

1. Fig 1.....	20
2. Fig 2.....	20
3. Fig 3.....	21
4. Fig 4.....	21

List of Tables

3.1	Features from Voice Recordings	13
4.1	Confusion Matrix Sample	16
4.2	Confusion Matrix for Neural Network	17
4.3	Confusion Matrix for Support Vector Machine	17
4.4	Confusion Matrix for AdaBoost	17
4.5	Comparison to Prior Work	18
4.6	Performance for all Classifiers	18
4.7	Confusion Matrix for AdaBoost with Hard Labelling	19
4.8	Confusion Matrix for Semi-Supervised Learning	19
4.9	Performance on Combined Dataset	19

I. INTRODUCTION

Parkinson's Disease (PD) is a degenerative neurological disorder marked by decreased dopamine levels in the brain. It manifests itself through a deterioration of movement, including the presence of tremors and stiffness. There is commonly a marked effect on speech, including dysarthria (difficulty articulating sounds), hypophonia (lowered volume), and monotone (reduced pitch range). Additionally, cognitive impairments and changes in mood can occur, and risk of dementia is increased. PD affects about 1-2% of people over 60.

Traditional diagnosis of Parkinson's Disease involves a clinician taking a neurological history of the patient and observing motor skills in various situations. Since there is no definitive laboratory test to diagnose PD, diagnosis is often difficult, particularly in the early stages when motor effects are not yet severe. Monitoring progression of the disease over time requires repeated clinic visits by the patient. There is no cure, but pharmacological treatment to manage the condition includes dopaminergic drugs.

Diagnosis is clearly a difficulty in PD management, and an effective screening process, particularly one that doesn't require a clinic visit, would be beneficial. Since PD patients exhibit characteristic vocal features, voice recordings are a useful and non-invasive tool for diagnosis. If machine learning algorithms could be applied to a voice recording dataset to accurately diagnosis PD, this would be an effective screening step prior to an appointment with a clinician.

Actually, this sort of binary classification problem is common in many areas of medical diagnosis, and techniques that work well within one domain are likely to be applicable to others. Based on prior work with PD as well as other disorders, support vector machines and artificial neural networks are ideal candidates for this type of problem. Another option for exploration is AdaBoost, an ensemble approach which combines learners in a way that minimizes their weaknesses.

However, one complication that may exist in available medical data is an inconsistency across datasets. For example, multiple datasets for a given disorder often exist, collected from different sources and using slightly different features. Combining them in some effective way into a large, cohesive dataset would result in a more robust and well-trained learner.

Another complication is an occasional lack of labelled examples. If some amount of labelled data exists, a semi-supervised learning algorithm might be applied, which uses a combination of labelled and unlabelled data. One approach is Expectation-Maximization (EM) which, given a model, alternates between making predictions about unlabelled data and recomputing the parameters of the model until convergence.

A successful machine learning approach to binary classification would be applicable to many types of medical diagnosis. Incorporation of semi-supervised learning with EM would result in a more robust system that is relevant to realistic medical datasets. Within the context of PD, this combination of techniques produces a useful screening step for a more frictionless diagnosis experience.

II. BACKGROUND

Many medical decision-making questions can be reduced to binary classification problems, making medical data an ideal domain for several machine learning techniques. A few of the most relevant algorithms, based on past work in this domain, are discussed below. Their applications to medical data in general, reviewed below, share many similarities with the specific case of Parkinson's diagnosis.

2.1 Machine Learning Techniques

2.1.1 Artificial Neural Networks

Artificial neurons were first proposed in 1943. Motivated by biological neurons, the artificial neuron received several weighted inputs and produced an output, based on some threshold. The perceptron model built on this early work, adding a learning rule to improve the performance of the neural network. However, the perceptron model was severely limited, unable to solve non-linearly separable functions such as XOR. Eventually, multilayer perceptrons were developed to address the original perceptron model's shortcomings.

Artificial neural nets have been widely used across a myriad of applications. For example, a checkers-playing program used neural networks to train a player. The board was encoded as a vector of available board positions, with values assigned based on whether the square was empty, taken by a regular piece, or taken by a King. The neural network consisted of three hidden layers. The first hidden layer completed spatial pre-processing, representing each sub-square of the board as a node, for a total of 91 nodes. The second and third hidden layers had 40 and 10 nodes, respectively. The network outputted a value between -1 and 1, representing the goodness of the board from the current player's perspective. The weights for the network were initially specified through a uniform sample, and several networks played against one another. The winners were declared 'parents', and they generated 'o spring networks' with weights varied by a parameter vector. The process was repeated for many generations to produce an ideal neural network.

2.1.2 Support Vector Machines

Support vector machines are binary classifiers that can be applied to linearly separable datasets. They separate data into classes using a hyper-plane. SVMs can also be used non-linearly by mapping the data to a higher-dimensional space, thus making the data separable. This mapping is done by a kernel function. SVMs perform well with large feature spaces, as long as the data is separable with a wide margin. They also do well with sparse datasets, as in text classification.

In the absence of large amounts of labelled data, pool-based active learning can be utilized with SVMs. The learning algorithm has access to a pool of unlabelled data, and is able to choose a subset of that pool to use as training data. The learner chooses pool data to use such that the data minimizes the size of the learner's set of hypotheses, and brings it closer to a single hyper plane. This approach allows for using less labelled data.

2.1.3 Ensemble Learners

Ensemble learners combine different machine learning algorithms. There is no one algorithm that always performs well on all domains, and ensemble learners are a way of combining the advantages of different learners. A good ensemble algorithm will be made up of diverse base learners that have varied strengths. The different learners can be combined in a number of ways. They can work in parallel on all of the inputs, and their outputs can be combined in some way. Alternatively, a multistage combination will train the base learners on different subsets of the input data. For example, the AdaBoost algorithm first trains an initial learner, and then trains subsequent learners on data that the first learner misclassifies. This way, the weaknesses of each base-learner are made up for by the next learner.

2.2 Applications to Medical Data

Medical diagnosis presents an ideal domain for machine learning algorithms. A large part of diagnosis falls under pattern recognition, based on large amounts of data, and ML algorithms are well-suited to this task. For an algorithm to be effective in this domain, it needs to be able to handle noisy and missing data, rely on relatively few medical tests, and complement the role of physicians [Kon01]. Machine learning algorithms have been applied to a variety of medical data, some examples of which are outlined below.

2.2.1 Self-reported Input

Some diagnoses rely largely on patient-reported information, rather than biological tests. A prime example of this is diagnosis of mental disorders, which is based on how a patient's symptoms compare to criteria outlined in the Diagnostic and Statistical Manual of Mental Disorders. Symptoms are determined through consultation with a mental health professional, and are largely reported by the patient. Automated systems have been proposed which will produce a diagnosis based on user-reported information [YC96].

2.2.2 Clinical Decision Support Systems

Clinical decision support systems help healthcare professionals make diagnosis decisions based on patient data. These systems can be rule-based, in which case they are created with a knowledge base and a set of rules. Alternatively, they can utilize machine learning to learn from past data and recognize patterns. Several such systems have been proposed, including a statistical approach to diagnosing digestive disorders based on an electronically-administered interview of the patient [SK84]. However, the use is not yet widespread, in part due to lack of data availability and to limited adoption of uniform computer systems [Gre07].

2.2.3 EEG and EKG data

Recordings of electrical activity in the body can be used to diagnose a variety of disorders. Electroencephalograms (EEGs) are recordings from the brain and contain a wealth of features that can be used by machine learning algorithms. A classification algorithm using EEGs was able to diagnose Alzheimer's Disease with 86.05% accuracy [Pod12].

Electrocardiograms (EKGs) are often used to detect arrhythmia, which is any abnormality of the heartbeat. They can be indicative of heart disease and other conditions. In 1989, a model was derived from the Cleveland Clinic heart disease data set and compared it to CADENZA, a Bayesian algorithm. Both models were found to over-predict heart disease, though this occurred more with CADENZA.

After that, a novel machine learning approach to diagnosing and classifying cardiac arrhythmia was presented, called the VF15 algorithm. It used a genetic algorithm to learn feature weights. Then, each feature voted on a class prediction. The algorithm had a 62% accuracy on this task and was found to outperform Naive Bayes.

Another study collected data on ischemic heart disease, including signs and symptoms, EKG, and scintigraphy. Several algorithms were applied, including Naive Bayes, neural networks, k-nearest neighbours, and two decision tree algorithms. These were compared to clinicians' diagnoses. Naive Bayes had the best sensitivity/recall, whereas clinicians, followed by neural nets, had the highest specificity [KKG+99].

More recently, a group compared various machine learning algorithms for arrhythmia diagnosis based on EKG data, with an emphasis on minimizing false positives and dealing with noisy data. They used the UCI Machine Learning Repository Arrhythmia dataset, and highlighted the need to improve on VF15s 62% accuracy. They evaluated a Bayesian artificial neural network classifier as compared to Naive Bayes, decision trees, logistic regression, and neural networks [GMCL05].

2.3 Diagnosing Parkinson's Disease

Speech tests can be used for monitoring Parkinson's Disease, due to vocal impairment being a common symptom and early indicator. Using an at-home recording device, such as one developed by Intel for PD telemonitoring, can conveniently allow PD patients' health to be monitored remotely. Specified voice recordings can be passed through signal processing algorithms and a classification and regression tree to predict a rating on the unified PD rating scale.

Another study described a weakly supervised multiple instance learning approach to detecting symptoms of Parkinson's Disease. This approach addressed the issue of self-reporting resulting in inaccurate or incomplete data. Their algorithm learned to localize symptoms to approximate, rather than exact, time ranges, and making it suitable for the sparse data that may result from incomplete reporting. Gil and Johnson used a multilayer network with one hidden layer and an output layer that output healthy or PD. The inputs were passed through a sigmoidal activation function, and gradient descent back-propagation was used to modify the weights. They achieved a classification accuracy of 92.31%. They also trained an SVM using the sequential minimal optimization (SMO) algorithm. SMO speeds up training of SVMs, particularly those with non-linear kernel functions (Platt 1998), using a divide and conquer approach. Gil and Johnson used a linear kernel with 91.79% accuracy, and a Pearson VII function kernel, with accuracy of 93.33% [GM09]. Mandal and Sairam also used a neural network with a sigmoidal activation function. They modified weights using back-propagation with dynamic learning rate and momentum, and achieved an accuracy of 97.6471%. They also used SVM with a linear kernel and obtained an accuracy of 97.6471% [MS12].

III. METHODS

3.1 Dataset

The primary dataset used in this project is from the UCI Machine Learning Repository [BL13], and contains data from voice recordings of 23 subjects with Parkinson's disease and 8 control subjects. There are a total of 195 recordings, from which 22 different voice measure features have been extracted. The features are listed in Table 3.1. Each example also includes a subject identifier and a binary classification attribute which indicates whether or not the subject has PD. There is also a secondary dataset, described in 3.6.

3.2 Unbalanced Data

A number of strategies were explored for dealing with an unbalanced dataset. In the oversampling approach, samples from the minority (healthy) class were replicated to create a dataset with equal proportions of both classes. In the under-sampling approach, fewer samples were taken from the majority (PD) class, leading to a smaller but balanced dataset. The final alternative, applicable only for support vector machines, was to adjust the cost-factor of the model, such that training errors on negative examples outweighed errors on positive examples. Since the negative (healthy) class was the minority class, this penalized false positives more than it did false negatives. The evaluation section includes results from the most effective of these approaches for each model. Table 3.1 below shows the features that have been extracted from subjects' voice recordings.

MDVP:F0(Hz)	Average vocal fundamental frequency
MDVP:Fhi(Hz)	Maximum vocal fundamental frequency
MDVP:Flo(Hz)	Minimum vocal fundamental frequency
MDVP:Jitter(%)	Average absolute difference of pitch between consecutive periods, divided by the average pitch
MDVP:Jitter(Abs)	Average absolute difference of pitch between consecutive periods
MDVP:RAP	Average absolute pitch difference between a period and the average of it and its neighbours, divided by the average pitch
MDVP:PPQ	Average absolute pitch difference between a period and the average of it and its four closest neighbours, divided by the average pitch

Jitter:DDP	Average absolute pitch difference between consecutive differences between consecutive periods, divided by the average
MDVP:Shimmer	Average absolute difference between amplitudes of consecutive periods, divided by average amplitude
MDVP:Shimmer(dB)	Measure of period-to-period variability of amplitude, in decibels
Shimmer:APQ3	Average absolute difference between amplitude of a period and the average of it and its neighbours' amplitudes, divided by the average amplitude
Shimmer:APQ5	Average absolute difference between amplitude of a period and the average of it and its four closest neighbours' amplitudes, divided by the average amplitude
MDVP:APQ	Measure of variation in amplitude
Shimmer:DDA	Measure of variation in amplitude
NHR	Noise to harmonic ratio
HNR	Harmonic to noise ratio
RPDE	Recurrence period density entropy, a measure of periodicity
D2	Nonlinear dynamical complexity measure
DFA	A measure of the stochastic self-similarity of noise in the speech signal
spread1	Nonlinear measure of fundamental frequency variation
spread2	Nonlinear measure of fundamental frequency variation
PPE	Nonlinear measure of fundamental frequency variation

Table 3.1: Features that have been extracted from subjects' voice recordings.

3.3 Neural Networks

The primary dataset was prepared for analysis by scaling all features to values between 0 and 1 and converting the set to a classification dataset format. The neural network library PyBrain [SBW+10] was used to create a feed-forward neural network with a bias node. The input layer of the network consisted of 22 neurons, one for each of the scaled features. There was one linear hidden layer, with 13 neurons. The output layer was a single classification neuron with a sigmoid activation function. This design was based on previous work [GM09]. A back-propagation trainer was created with a learning rate and momentum that were varied from 0.01 to 1 in increments of 0.01 to determine optimal parameter values.

3.4 Support Vector Machines

Three versions of the primary dataset were produced: one with all the features, one with a 10-feature subset, and another with a 4-feature subset. Subsets were determined based on prior work [LMH+09], although the paper being replicated ([MS12]) does not explicitly enumerate which features were used.

The support vector machine was created using the SVM-light library [Joa99]. A variety of kernels were used, including linear, 2-, 3-, and 4-degree polynomial, and sigmoid tanh. The cost-factor was determined based on whether or not the dataset was balanced; it was set to 1 for the balanced datasets and to lower values for the imbalanced set. The C parameter, which is a trade-off between training error and SVM margin, was varied from 0.01 to 10000 at factor of 10 intervals.

3.5 AdaBoost

An ensemble boosting approach was implemented using the scikit-learn library [PVG+11]. Base learners were created with both SVMs and Decision Tree Classifiers. When the weak learner was an SVM, a linear kernel was used and the C parameter was varied from 0.01 to 10000 at factor of 10 intervals. The number of learners were varied from 1 to 1000000 to determine optimal performance.

3.6 Combining Datasets

In addition to the primary classification dataset, which has been used thus far in this work, there is a secondary regression dataset of voice recordings available from [BL13]. The secondary dataset contains a total of 5875 recordings from 42 subjects, with 16 vocal attributes each. The primary dataset includes these 16 attributes along with 6 additional voice recording features. The secondary dataset's output is a score on the Unified Parkinson's Disease Rating Scale (UPDRS).

In determining whether there is a way to utilize this secondary dataset, it was useful to create a combined dataset containing the additional points.

Since there isn't a perfect overlap of attributes, there are two ways of doing this. A combined dataset was created with a reduced feature space, which only included the intersection of features across the primary and secondary datasets. Additionally, a sparse dataset was created, including the union of features across both datasets, and holding a null value when a feature was not available. Both of these approaches were evaluated.

To continue treating this as a classification problem, two methods of handling the secondary dataset's regression outputs were attempted. One was to set a hard label on each data-point in the secondary dataset, by rounding the UPDRS score to either healthy or PD. This effectively transformed the secondary dataset into a classification set, and left us with a standard binary classification problem. The highest performing algorithm, based on evaluation of the models described above, was used to test this approach.

The second method was to treat the additional data-points as unlabelled data. Since we have labelled data from the primary dataset, the addition of unlabelled data made this problem an ideal candidate for a semi-supervised learning approach. Expectation maximization, using the scikit-learn library [PVG+11], was used to train a model on the labelled data, use that model to assign soft labels to the unlabelled data, and train a new model using both datasets. This process continued until convergence.

IV. EXECUTION

All learners were evaluated using 10-fold cross validation. Data was shuffled then divided into 10 sections, and the learner was trained on 9 of those sections and tested on the 10th. The test section was rotated such that this process occurred 10 times. The learner's performance on the test set was compared against the expected values, and a confusion matrix was generated, consisting of the numbers of true positives, true negatives, false positives, and false negatives. An example confusion matrix is shown in Table 4.1. These values were used to calculate accuracy, recall, precision, and F-score, as shown in Equations 4.1 through 4.4.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4.1)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4.2)$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4.3)$$

$$\text{F-score} = 2 \frac{\text{precision recall}}{\text{precision} + \text{recall}} \quad (4.4)$$

Results for each of the learners, along with the parameters selected to obtain that result, are described below. The results here are based on using all the features, but one of the papers being reproduced used an undefined subset of the feature space. Without knowing which features were used, it is difficult to replicate those results exactly. This is an area for improvement in future work. A comparison to prior work is shown in Table 4.5.

	Truth: PD	Truth: Healthy
Predicted: PD	True positives	False positives
Predicted: Healthy	False negatives	True negatives

Table 4.1: A sample confusion matrix.

	Truth: PD	Truth: Healthy
Predicted: PD	147	30
Predicted: Healthy	0	18

Table 4.2: A confusion matrix for the neural network classifier

4.1 Neural Network

The best performance was achieved using the primary dataset with all 22 attributes, scaling each feature to a value between 0 and 1. Learning rate was 0.05, momentum was 0.01, and weight decay was 0.01. The trainer ran for 1000 epochs and resulting in the confusion matrix shown in Table 4.2 and the performance shown in Table 4.6.

4.2 Support Vector Machine

The full primary dataset was used with a reduced feature space consisting of 10 features. The kernel was a 2nd degree polynomial, cost factor was set to 0.85, and the C parameter was 1000. This was the best performing configuration out of all of those that were tried. This resulted in the confusion matrix seen in Table 4.3 and performance in Table 4.6.

	Truth: PD	Truth: Healthy
Predicted: PD	136	16
Predicted: Healthy	11	32

Table 4.3: A confusion matrix for the support vector machine classifier

	Truth: PD	Truth: Healthy
Predicted: PD	143	9
Predicted: Healthy	4	39

Table 4.4: A confusion matrix for the AdaBoost classifier

		Gil and Johnson 2009	Mandal and Sairam 2012
Neural Network	0.908	0.948	0.86
Support Vector Machine	0.910	0.948	0.93
AdaBoost	0.957	-	0.95

Table 4.5: Comparison to prior work. F-scores from this work, compared to [GM09] and [MS12].

4.3 AdaBoost

A decision tree classifier was used as the base learner, and the full primary dataset was used for training and testing. AdaBoost was used to combine 1000 base learners, and resulted in the confusion matrix in Table 4.4. The performance of this ensemble approach can be seen in Table 4.6.

4.4 Combined Dataset

For both hard labelling and expectation maximization, two versions of the combined dataset were used: a sparse dataset and a set with a reduced feature space. AdaBoost had the highest performance of the three approaches outlined above, so it was used in evaluating these datasets. Various approaches for dealing with unbalanced data were used, and oversampling the minority set had the highest performance. Under-sampling the majority set was not used because doing so would not take advantage of the additional data available through the secondary dataset. The models were tested on samples from the primary dataset. A balanced subset of samples was retained for testing, and 3-fold cross validation was used for evaluation.

	Accuracy	Recall	Precision	F-Score
Neural Network	0.846	1.000	0.831	0.908
Support Vector Machine	0.862	0.925	0.895	0.910
AdaBoost	0.933	0.973	0.941	0.957

Table 4.6: Performance metrics for all classifiers, trained and tested on the primary dataset.

	Reduced feature space		Sparse dataset	
	Truth: PD	Truth: Healthy	Truth: PD	Truth: Healthy
Predicted: PD	45	15	47	26
Predicted: Healthy	3	33	1	22

Table 4.7: A confusion matrix for AdaBoost using hard-labelled combined datasets.

	Reduced feature space		Sparse dataset	
	Truth: PD	Truth: Healthy	Truth: PD	Truth: Healthy
Predicted: PD	30	0	33	0
Predicted: Healthy	18	48	15	48

Table 4.8: A confusion matrix for AdaBoost using expectation maximization to label combined datasets.

Confusion matrices for hard labelling and expectation maximization are presented in Tables 4.7 and 4.8, respectively. F-scores for both approaches are in Table 4.9.

	Reduced feature space	Sparse dataset
Hard labelled	0.714	0.777
Expectation Maximization	0.769	0.815

Table 4.9: F-scores for classifiers trained and tested on the combined dataset.

IV. RESULT

The screenshots below show the working of the prototype application that we have proposed.

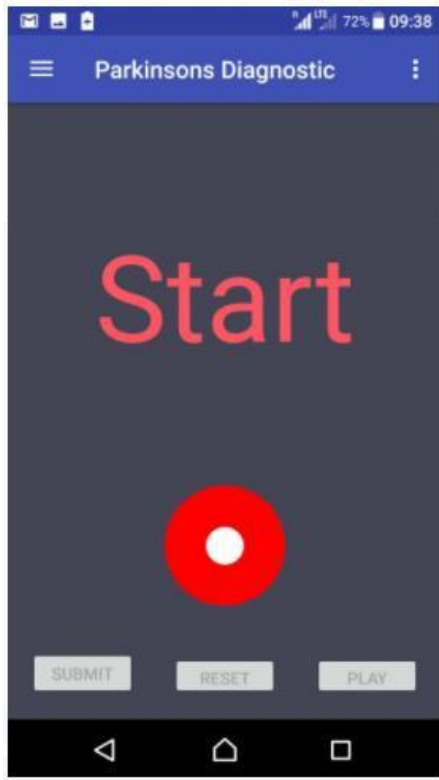


Fig1: Starting the diagnosis

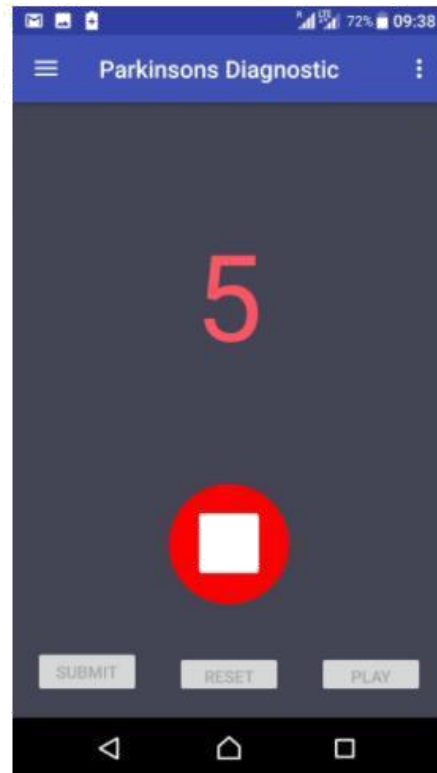


Fig2: 10 seconds recoding of patient

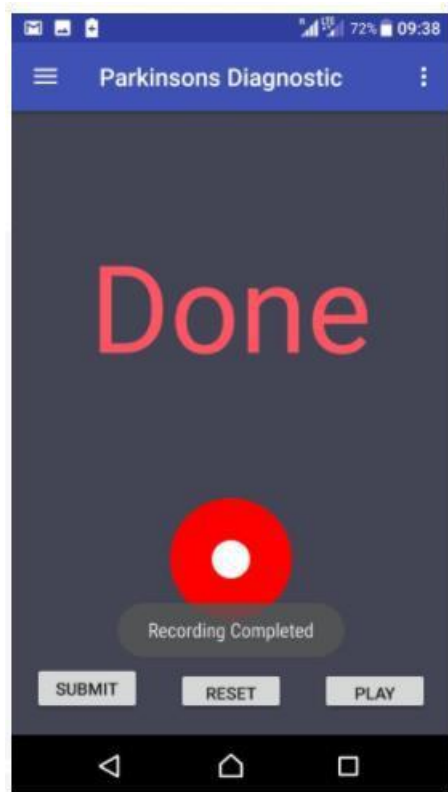


Fig3: Recording submission

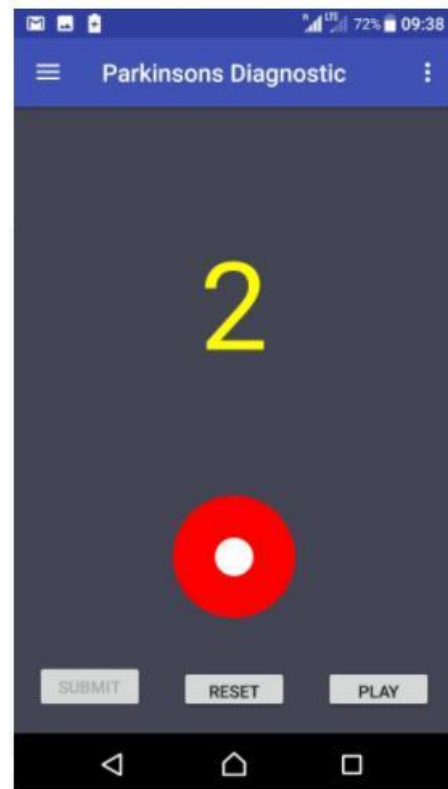


Fig4: Degree of Certainty

Fig1 shows the initiation of diagnosis. Fig2 depicts the recording done to the voice modulation of the patient to judge his/her health. Fig3 is about the Countdown for the patient to speak for 10 seconds which will help the application to give an output. Consequently, Fig4 is the result in terms of Degree of Certainty of Parkinson's Disease in that patient.

V. CONCLUSION

Parkinson's Disease is often difficult to diagnosis, but even at early stages, small vocal differences may be machine-detectable. Using this information, it becomes possible to predict PD using voice recordings from potential patients. Several machine learning approaches are effective for this task, with F-Scores above 0.90. The most effective approach is ensemble learning using AdaBoost, with an F-Score of 0.957. The results for all three learners were similar to those in the literature, with some expected difference. Since prior work used a reduced feature space which was not explicitly defined, this work was done with the understanding that past results would not be exactly replicated.

Utilizing information from additional datasets is not as helpful as predicted. This may be due to a variety of factors. When there isn't a perfect overlap between features in different datasets, combining them requires discarding some of the features, or having a sparse set with many missing values. Neither of these approaches are optimal in creating a robust model. Additionally, both the primary and secondary datasets in this case were unbalanced, and combining them simply increased the overall level of imbalance. There are a few ways of dealing with this, but none are as optimal as using a balanced dataset. Under-sampling requires throwing away data, which leaved fewer data points with which to train and test a model. Oversampling requires duplication of some data points, possibly leading to over fitting. Adjusting cost factor can help, but there is a limit to the level of imbalance it can tolerate.

VI. FUTURE WORK

There are several areas for improvement in future work on this topic. Feature analysis can be utilized to determine the ideal subset of the feature space to use for the highest performance. Since some of the attributes are related, this lack of independence may cause these attributes to have a disproportionate influence on the results. Additionally, it's possible that some attributes do not contribute to the results, and reducing the feature space in some way may lead to better performance. The number of possible feature sets for these 22 attributes is the power set of the attributes, which has a cardinality of 4,194,304. Some heuristic approach to limit the number of possible subsets would make this problem more feasible to solve.

Additionally, the imbalance of the dataset likely had a detrimental effect on performance. Collecting voice recordings from healthy subjects seems a manageable task, particularly in comparison to the more difficult task of finding PD subjects from whom to collect recordings. Obtaining more data to create a balanced dataset would likely lead to a significant performance improvement for the various learners.

Finally, the secondary regression dataset might be used in ways other than augmenting the primary dataset. For example, regression data collected over time may be useful in monitoring the progression of PD, reducing the need for constant and inconvenient clinical appointments for PD patients.

All of the approaches used here, as well as the suggested modifications, may also be applicable to other binary classification problems within the medical domain. PD is a useful starting point due to the availability of noninvasively-collected data, but the findings here are relevant to a variety of conditions.