

**A DATA MINING METHODOLOGY FOR PREDICTING EARLY STAGE
PARKINSON'S DISEASE USING
NON-INVASIVE, HIGH DIMENSIONAL GAIT SENSOR DATA**

ABSTRACT

Parkinson's disease (PD) is the second most common neurological disorder after Alzheimer's disease. Key clinical features of PD are motor-related and are typically assessed by healthcare providers based on qualitative visual inspection of a patient's movement/gait/posture. More advanced diagnostic techniques such as computed tomography scans that measure brain function, can be cost prohibitive and may expose patients to radiation and other harmful effects. To mitigate these challenges, and open a pathway to remote patient-physician assessment, the authors of this work propose a data mining driven methodology that uses low cost, non-invasive sensors to model and predict the presence (or lack therefore) of PD movement abnormalities and model clinical subtypes. The study presented here evaluates the discriminative ability of non-invasive hardware and data mining algorithms to classify PD cases and controls. A 10-fold cross validation approach is used to compare several data mining algorithms in order to determine that which provides the most consistent results when varying the subject gait data. Next, the predictive accuracy of the data mining model is quantified by testing it against unseen data captured from a test pool of subjects. The proposed methodology demonstrates the feasibility of using non-invasive, low cost, hardware and data mining models to monitor the progression of gait features outside of the traditional healthcare facility, which may ultimately lead to earlier diagnosis of emerging neurological diseases.

Keywords: *Parkinson's disease (PD), gait, non-invasive, sensor, data mining, healthcare, prediction, neurological disease, privacy, telehealth*

1. INTRODUCTION

Parkinson's disease (PD) is the second most common neurodegenerative disorder after Alzheimer's disease, affecting more than one million people in North America and about 2% of the population over the age of 65 years (Gil and Manuel 2009), (Patel et al. 2009), (Lang and Lozano 1998). The key abnormality

in PD is the deficiency of dopamine in the substantia nigra, a region in the midbrain (Hughes, Ben-Shlomo, et al. 1992). Although dopaminergic therapy offers substantial relief from the debilitating motor symptoms associated with PD, the lack of PD features for the dopamine cell loss makes it difficult to determine the optimal time to administer such medication (Lawrence, Evans, and Lees 2003). In addition, greater than 50% of the dopaminergic neurons are lost by the time that a patient is clinically diagnosed with PD, making early-stage detection challenging and critically important for neuroprotective efforts (Gil and Manuel 2009). The progression of the disease is formally described as the Hoehn and Yahr Staging of PD (from stage I to stage IV) (S. J. G. Lewis et al. 2005). PD motor symptoms typically include tremor at rest, rigidity, bradykinesia, and gait-/posture-related dysfunction (Hoff et al. 2001; Fahn 2003). Recent studies have demonstrated that some gait abnormalities may appear particularly early in PD, although these early features remain to be elucidated fully (Huang et al. 2012; Lewek et al. 2010). Non-motor symptoms of PD also exist such as hypophonia, sleep disturbances, depression, and dementia (Olanow 2004). Physicians and scientific investigators presently use published criteria to establish a PD diagnosis (as discussed further in the next section) but these approaches can vary among raters. In addition, PD may evade early diagnosis due to the lack of routine screening, hindering efforts of targeting early intervention.

In order to mitigate these challenges, the proposed methodology aims to quantitatively evaluate PD by using low-cost, non-invasive and privacy preserving hardware to capture patient data and mine it for potentially novel PD features that can be used in PD diagnosis. Furthermore, the proposed methodology may enable patients to perform PD diagnosis in non-traditional settings, such as in the comfort of their home, as the hardware and resulting machine learning algorithms are not specific to a clinical location. The proposed methodology aims to serve as a decision support system for physicians so that patients at risk of PD can be flagged at the early stages. Related approaches to PD diagnosis employ artificial intelligence (AI) techniques. For example, a biometric “smart pen” has been proposed to capture fine motor skills by measuring acceleration, finger grip force, and refill force (Barth et al. 2012). Other similar data collection techniques were employed by Patel et al. (Patel et al. 2009). However, the main limitation of these existing

techniques is that the PD features are pre-determined and may omit latent controller variables reflecting neural organization, particularly at the early stages of PD. Gait impairment, conversely, may involve distinctive abnormalities in inter-limb dynamics that occur relatively early in PD (Huang et al. 2012; Lewek et al. 2010; Bartels et al. 2003). Moreover, detection of PD using gait-related features may be well suited for the application of machine learning, since limb dynamics can be captured easily and inexpensively by remote sensing technologies. In addition to n -fold cross validation techniques that will be employed to evaluate the veracity of the proposed data mining models, the results from this study will also be compared against previous works by (Huang et al. 2012; Lewek et al. 2010) that have identified notable gait related features (e.g., variations in arm swing as a discriminating feature of PD patients).

The organization of this paper is as follows. This section gives a brief introduction and background about the problem. Section 2 describes some previous work related to the current research and compares them to the methodology proposed in this paper. Section 3 introduces the data mining driven methodology for early stage PD detection. A case study involving PD patients and controls is presented in Section 4, along with the results and discussion. Section 5 concludes the paper and outlines future work.

2. LITERATURE REVIEW

2.1 Clinical PD diagnosis

Currently, clinical PD diagnosis is primarily based on clinical history and physical examinations. For example, a patient is first evaluated by a general practitioner, followed by a referral to a neurologist if patient abnormalities are suspected. A neurologist would then interview the patient, followed by a symptom questionnaire. Finally, neurology movement disorder specialist would examine information based on published diagnostic PD criteria to finally determine if someone is a PD patient or not (Schrag, Ben-Shlomo, and Quinn 2002), (Hughes, Ben-Shlomo, et al. 1992), (Rajput et al. 1991). Typically, a patient's history and medication records would first be screened for preliminary PD related evidence, followed by several physical examinations such as blood and imaging tests, depending on the physician's observations. Physicians may give a trial of dopaminergic drugs to observe patient response and changes in health

conditions (Freed, Curt R. 1992). Despite the ability of current clinical based criteria to correctly diagnose PD, they require each patient to visit the clinic for a series of invasive and non-invasive tests. Furthermore, different physicians may give results based on variable rating scales such as Unified Parkinson's Disease Rating Scale (UPDRS) and Intermediate Scale for Assessment of PD (ISAPD). These rating scales are subjective and qualitative, sometimes making interpretation or comparison difficult (Martinez-Martin et al. 1994). Other techniques used to diagnose PD include neurologically based techniques such Single-photon emission computed tomography (SPECT) scans that measure dopaminergic transport levels in the brain. These techniques can be used to assist the diagnosis process and differentiate PD from essential tremor and other neurodegenerative disorders that may exhibit similar characteristics of PD patients (M. M. Lewis et al. 2007), (Dastgheib, Lithgow, and Moussavi 2011), (Huang et al. 2006). While these diagnosis techniques have been proven to be effective at accurately diagnosing PD, greater than 50% of the dopaminergic neurons are lost by the time that a patient is clinically diagnosed with PD, making early-stage detection challenging and critically important for neuroprotective efforts (Gil and Manuel 2009). Furthermore, such techniques (e.g., SPECT scans) are cost prohibitive and require pre-domain knowledge about the factors that affect PD. These challenges are further exacerbated by the variations in PD symptoms and motor complications (referred to as motor fluctuations) among different people, hereby potentially reducing the quality of the resulting diagnosis.

2.2 Phonation based PD diagnosis

Researchers have also shown that voice may be a useful signal for differentiating PD patients from controls since the vast majority of PD patients typically exhibit some form of vocal disorder such as reduced loudness, increased vocal tremor, and noise (Ho et al. 1999), (Little et al. 2009). For example, the phonation impairments between PD patients and control groups have been measured based on 132 dysphonia features using Support Vector Machines (SVM) algorithms. Researchers were able to detect PD cases using only 10 of the 132 dysphonia features with about 99% accuracy (Tsanas et al. 2012). Similar studies can be found in (Tsanas 2010), (Tsanas et al. 2010). Even though these phonation based methodologies were able

to distinguish PD patients from controls, phonation data collection could be difficult and complicated (e.g., due to variations in accents, speech patterns, etc.), which may increase the testing requirements for the subjects (Tsanas et al. 2012). In addition, even though phonation defects are common evidence in PD, they are typically good indicators for detecting advanced PD stages (Stewart et al. 1995), (Darley, Aronson, and Brown 1975), (Critchley 1981), as opposed to early stage detection.

2.3 Gait based PD diagnosis

Another approach to detecting PD is based on gait impairment evaluations such as reduced stride length, reduced arm swing, reduced gait speed and stride freezing. For example, abnormal movement/gait in PD patients could be recognized by movement disorder specialist and measured through accelerometers (ACC) and surface electromyographic (EMG) signals (Bonato et al. 2004). Another example is to have sensor-based systems placed on the upper and lower limbs of each subject so as to measure the hand motor and gait functions (Barth et al. 2012). However, the main limitation among these methodologies is that the applied markers may not be the most predictive indicators of PD. Furthermore, these data collection techniques require physical contact with the patient, which may affect the manner in which the patient responds.

3. METHODOLOGY

To overcome the limitations of PD diagnosis discussed in the previous section, the proposed data-mining driven methodology employs machine learning techniques that reduce subjective biases (resulting from human physical examination) and identifies latent PD features that may be relevant in detecting early stage PD. The authors hypothesize that: *non-invasive sensors that capture 3D gait data, and subsequent machine learning algorithms employed on the acquired data, detect gait features relevant to PD.* Figure 1 presents the overall methodology which is based on extracting relevant PD motor features and applying these features to detect PD gait patterns. The methodology is partitioned into four steps. Step 1 outlines the manner in which patient gait data is acquired in a non-invasive manner and stored for subsequent data mining/knowledge discovery. Step 2 involves data cleaning and feature dimension reduction. The process

of data mining and knowledge discovery is introduced in step 3, where multiple machine learning algorithms are investigated and benchmarked against one another. Finally, step 4 evaluates the model performance of different machine learning algorithms and outlines the potential for the methodology to be integrated into a healthcare decision support system.

The data captured by each patient at each time interval (33ms) is considered an independent sample, so as to minimize the assumptions made about the gait features, their correlations with each subsequent data point captured in time, and their correlations with the class variable (i.e., whether a subject has PD or not). For example, it may be determined that whether a subject has PD or not has more to do with the maximum position his/her arm is able to swing upwards, rather than where that arm position was in the previous time stamp. Researchers in the data mining community have proposed algorithms that augment existing data mining classification algorithms (such as those employed in this work) to factor data streams with concept drift (C. S. Tucker and Kim 2011; Wang et al. 2003; Richards et al. 2011; Domingos and Hulten 2000; Hulten, Spencer, and Domingos 2001). As new incoming patient data is acquired, the models generated by the data mining algorithms employed in this work can be iteratively updated, once their predictive accuracies fall below a given threshold (determined by the healthcare decision makers). Using non-invasive sensors, the authors have demonstrated the feasibility of modeling body movement data in this manner to predict emerging human behavior patterns (for security applications) and emotional states (for engineering education applications) with relatively high accuracy (Manohar and Tucker 2013; Behoora and Tucker 2014; Behoora and Tucker 2015). The authors have also demonstrated that these non-invasive sensors are capable of modeling and predicting medication adherence in patients with neurologically induced movement disorders (C. Tucker et al. 2015). The methodology presented in this work aims to advance the field of neurological disease detection in the following ways:

1. The ability to use low cost, non-invasive data acquisition techniques that concurrently maintain patient privacy by only collecting the three-dimensional (3D) geospatial data pertaining to a patient's node location.

2. The ability to extract relevant gait features in order to model and predict early stage PD.
3. The ability to quantify PD movement abnormalities, independent of clinic location (e.g., from a patient's home), hereby serving as an early warning system that can be communicated to physicians and healthcare decision makers.

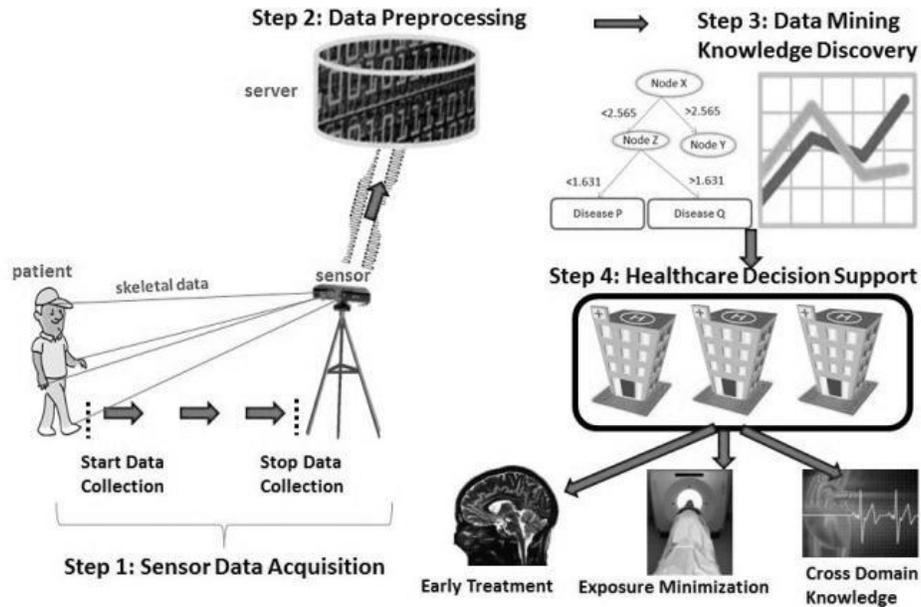


Figure 1: Overview of proposed data mining methodology

Step 1: Sensor Data Acquisition

Step 1 of the proposed methodology utilizes low cost, off the shelf hardware sensors (e.g., Microsoft Kinect) to capture human gait data. While multimodal sensors such as the Kinect are capable of capturing multiple data streams (VGA, depth using infrared sensors), only the 3D coordinate skeleton data (XYZ coordinate data of each node) is required for the proposed methodology as shown in Figure 2. For example, the Kinect can non-invasively approximate full body human gait by using its depth sensors to map 20 nodes on a human body as seen in Figure 2, starting with the XYZ coordinate data of the *Head* node, down to the XYZ coordinate data of the *Foot Right* and *Foot Left* nodes. As a patient walks through space, the coordinate data for each of these nodes is captured and stored in a structured table similar to that shown in Figure 2.

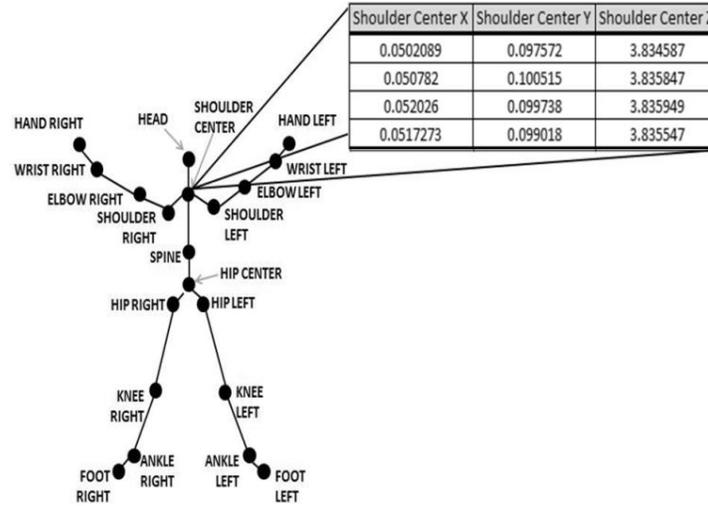


Figure 2: Skeleton image of a person with 20 nodes and example data from Shoulder Center node

In addition, since human gait is approximated using only the 3D unidentifiable skeleton data (XYZ coordinate points of each node) as seen in Figure 2, patients’ privacy would be preserved as well. The Microsoft Kinect sensor is employed in the case study to collect human gait due to its wide availability, relatively low cost and high fidelity (Kamel Boulos 2012). This hardware enables researchers to capture data samples (initially containing XYZ node locations for each of the 20 nodes) in a timely and efficient manner. Each row of the table in Figure 2 represents a data sample that is captured every 33ms, representing position data from each of the 20 nodes. The hardware is capable of adjusting the 20 node locations, based on varying degrees of human characteristics (size, shape, height, etc.). While the Kinect is one of the most popular off the shelf hardware for skeletal and depth sensing capabilities, other alternatives such as the Asus Xtion Live (“Xtion PRO LIVE - Multimedia - ASUS” 2013) and PrimeSense Carmine (Primesense 2013), make the proposed data acquisition step of the proposed methodology unconfined to a specific hardware.

Step 2: Data Preprocessing

The Data preprocessing step of the proposed methodology aims to remove irrelevant/noisy data from the initial raw data captured from Step 1, leaving only the XYZ data needed for the data mining machine learning step (Step 3). The Data Preprocessing step also includes the calculation of velocity

($\frac{position(t_i) - position(t_{i-1})}{t_i - t_{i-1}}$) and acceleration ($\frac{velocity(t_i) - velocity(t_{i-1})}{t_i - t_{i-1}}$) of each node, based on the original

XYZ position data of each of the 20 nodes. Here, the change in time ($t_i - t_{i-1}$) from one instant in time (t_{i-1}) to a subsequent instant in time (t_i) is approximately 33ms. The Microsoft Kinect employs a randomized decision forest algorithm to automatically assign landmarks on a human body that correspond to different human joint locations for the 20 joint locations that it is capable of tracking (Lai, Konrad, and Ishwar 2012; Ye et al. 2011; Clark et al. 2012). The Kinect is capable of tracking the skeletal joints across a wide range of different body sizes, skin pigmentations and outfits (Shotton et al. 2013). During the machine learning training step of the proposed method however, the performance of the resulting machine learning models may be reduced if the heights of the test subjects significantly differs from that of the training subjects, as measured primarily by the Y position direction of the skeletal data. These model inaccuracies may be overcome by normalizing patients' height data or training the model to be more robust to wide variations in patients' height. Therefore in this work, the ratio in each dimension from each pair of nodes in the position, velocity and acceleration data are also generated in order to normalize variations in human characteristics (size, height, weight) that may cause biases in the resulting data mining predictive models. The result is a total of 1890 features generated, comprising of 60 position features (i.e., 20 position features pertaining to the X node, 20 position features pertaining to the Y node and 20 position features pertaining to the Z node), 60 velocity features (i.e., 20 velocity features pertaining to the X node, 20 velocity features pertaining to the Y node and 20 velocity features pertaining to the Z node), 60 acceleration features (i.e., 20 acceleration features pertaining to the X node, 20 acceleration features pertaining to the Y node and 20 acceleration features pertaining to the Z node). The remaining 1710 features represent ratio features generated using the XYZ position, velocity and acceleration data (i.e., enumerative combination of ratios of each node relative to the remaining nodes). The features can be considered as *input* variables while the *class* variable, can be considered as the *output* variable. The *class* variable in this case is binary in nature (i.e., whether a patient has Parkinson's disease [denoted as PD] or not [denoted as non-PD]). It is demonstrated in this work that features predicted by the data mining algorithms (explained in Step 3 of the

proposed methodology) are relevant in distinguishing PD vs non-PD patients using an *argmax* function. Furthermore, the authors of this work demonstrate that the features identified as relevant in the resulting data mining models, are consistent with those found in previous works such as (Huang et al. 2012; Lewek et al. 2010) that employ different techniques (such as wearable technologies or visual inspection) to predict early stage PD.

Step 3: Data Mining Knowledge Discovery

The data mining/knowledge discovery step of the proposed methodology aims to develop a function $f(X) = Y$ that would map the selected features $X = \{x_1, x_2, \dots, x_n\}$ to the class variable (Y), where n is the total number of features relevant to the class variable Y . The Data Mining step (Step 3) develops a model that quantifies novel feature combinations that are relevant to predicting the class variable (in this case binary). In this section, several data mining algorithms are discussed based on their applicability, accuracy and complexity in binary classification problems (A and B 2006), (Ramani, R. Geetha 2011), (Youn and McLeod 2007), (D'heygere, Goethals, and De Pauw 2003).

Naïve Bayes Algorithm

The Naïve Bayesian classification algorithm has been extensively employed in the data mining research community and has demonstrated accuracies comparable to more advanced machine learning algorithms, while being less computationally complex (Zhang and Su 2004). Mathematically, the Naïve Bayes algorithm can be succinctly represented as:

$$P(y_i | x_1, \dots, x_n) = p(y_i) \prod_{j=1}^n p(x_j | y_i) \quad (1)$$

Where,

- x_j : represents a feature/input variable (j) included in the model
- n : represents the total number of features in the data set
- $p(y_i)$: is the prior probability of the class/output variable
- $p(x_j | y_i)$: is the likelihood of the feature, given the class variable y_i

Support Vector Machines

Support Vector Machines (SVMs) are a relatively new branch of data mining classification techniques that aim to maximize the linear boundary in logistic regression modeling. SVMs attempt to construct a decision boundary based on a kernel function which could optimally separate two classes by maximizing the geometric margin between data points in two classes. In SVMs, the fitting process is novel, where problematic observations are treated differently and could finally determine the “best” boundary by constructing two parallel lines without any observations within the space between them. In practice, since data cannot be linearly separated in some cases, SVMs would also be able to transform the current data to a higher dimensional space and construct the decision boundary (Hastie et al. 2005). The general fitting function of the model is.

$$f(x_1, x_2, \dots, x_n) = \sum_{j=1}^n (w_j x_j) + b \quad (2)$$

Where,

b : is the misclassification tolerance

n : is the total number of input variables

w_j : is the weight corresponding to any input variable x_j

C4.5 Decision Tree Induction

The C4.5 decision tree classifier is a supervised machine learning algorithm that iteratively tests each feature for its ability to reduce uncertainty (randomness) in the *class/output variable* (Quinlan 1993). The methodology comprises of three steps: best feature evaluation, splitting point selection and model training and evaluation. The feature evaluation step of the methodology attempts to select the most informative node in each subset of the training data set by maximizing the *gain ratio* metric (Quinlan 1993). The splitting point selection determines the best numerical split point which has the minimum misclassification error. The last step of the algorithm trains the resulting tree model, based on a predefined stopping criteria and evaluates its predictive performance. The general modeling forms are shown here (Ramani, R. Geetha 2011), (Maroof 2012), (Ozcift 2012).

$$Info(D) = -\sum_{i=1}^m p_i \log_2(p_i) \quad (3)$$

$$Info_A(D) = \sum_{l=1}^v \frac{|D_l|}{|D|} \times Info(D_l) \quad (4)$$

$$Gain(A) = Info(D) - Info_A(D) \quad (5)$$

$$SplitInfo_A(D) = -\sum_{l=1}^v \frac{|D_l|}{D} \times \log_2 \frac{|D_l|}{D} \quad (6)$$

$$GainRatio(A) = Gain(A) / SplitInfo(A) \quad (7)$$

Where,

$Info(D)$: is the expected information (entropy) needed to classify a tuple in D

p_i : is the probability that an arbitrary tuple in D belongs to class y_i , calculated by $|y_{i,D}|/|D|$, for a total of m mutually exclusive class variables

$Info_A(D)$: is information needed (after using feature A to split D into v partitions) to classify, where l represents a mutually exclusive value of a given feature A , and v represents the number of mutually exclusive values of feature A

$Gain(A)$: is information gained by partitioning the data set on attribute A

$SplitInfo(A)$: is the normalization of the $Gain(A)$ to account for attributes with a wide range of mutually exclusive values

$GainRatio(A)$: is the metric used by the algorithm to evaluate each attribute during tree generation

Random Forest

The Random Forest (RF) model is an effective classifier that capitalizes on flexible fitting procedures, hereby potentially improving on accuracies, compared to traditional tree based models. The general procedure of RF is as follows. First, M cases are randomly sampled (M is the predetermined sample size) with replacement which is the training data set for each tree. Next, an n variable sample would also be created (n is the variable sample size) to help split in each node which is similar in the single tree

construction. Finally, each tree is grown to the largest extent and the maximum voted class in the forest is accepted as the final decision (Amasyali and Dirir 2006).

IBK

The classifier IBK is another type of the instance-based learner. In the IBK classifier, a normalized distance is applied so that each feature has the same impact on the distance measure, and each observation would be classified in the most similar class assigned to the majority of its K neighbors (K is the number of nearest neighbors in prediction model). For example, if $K = 1$, then each new observation would be classified in the class by its nearest neighbors as shown in Equation 8 (Ramani, R. Geetha 2011), (Ozcift 2012).

$$Similarity(x, y) = -\sqrt{\sum_{i=1}^n f(x_i, y_i)} = -\sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (8)$$

Where,

x_i : is the value of i^{th} feature of observation x

y_i : is the value of i^{th} feature of observation y and

n : is the total number of features.

Based on the machine learning algorithms above, the early PD status would be modeled and quantified based on a combination of relevant features. Furthermore, these identified features may serve as evidence to help doctors screen potential early PD patients (explained in detail in the next session).

Step 4: Healthcare Decision Support

Step 4 outlines how the results from the data mining step, can serve as an integrated healthcare decision support system, connecting patients with physicians for early stage PD detection and progression. In order for the proposed methodology to be viable as a healthcare decision support system, the accuracy and robustness of PD models need to be clearly outlined and visualized. The methodology outlines two stages of validation: i) validation of data mining algorithms through n -fold cross validation using sample

data to both train and test the model and ii) validation of model predictive power using unseen data from both PD patients and controls.

Stage 1: Validation of Data Mining Algorithms through n -fold Cross Validation

The n -fold cross validation (CV) method is employed in this work in order to evaluate algorithm performance and consistency. Given a sample data set containing both PD patients and controls, CV involves partitioning the data set into n subsets, training the predictive model on one subset called the *training data set*, and testing/validating the model performance using the remaining $n-1$ subsets called the *test data sets*. Then, multiple rounds of CV are performed and the validation results are averaged (Kohavi 1995). In the computer science domain, 10-fold CV has been shown to generate consistent results (Tibshirani, Robert and Hastie, Trevor and Narasimhan, Balasubramanian and Chu 2002), (Ambroise, Christophe and McLachlan 2002) and is employed in this work to demonstrate the robustness of the resulting data mining models in consistently predicting the correct class values.

Stage 2: Validation of Data Mining Models' Predictive Power Using Unseen Patient Data

Once the robustness of data mining algorithms have been determined and a stable set of data mining algorithms have been selected, the goal of stage 2 is to determine whether the data mining models have predictive power. Unlike stage 1 validation that randomly partitions each data sample during the model generation stage and uses some of the randomly partitioned unseen data samples to test the model, the stage 2 model validation tests the data mining models based on an entire sample set acquired from subjects not included in the data mining model generation step. Stage 2 therefore measures the data mining models' ability to predict PD vs controls from an actual subject pool.

The performance of the machine learning algorithms will be assessed using the Correctly Classified Instances (CCI) measure. The CCI uses statistics from the confusion matrix shown in Table 1 to evaluate the average accuracy of each classifier using the following mathematical equation (Eq.9).

$$CCI = \frac{TP + TN}{TP + TN + FP + FN} * 100\% \quad (9)$$

Eq.9 contains four values predicted by a given data mining model: true positive (TP), false positive (FP), false negative (FN) and true negative (TN). These four outcomes are also used to derive the *precision*, *recall*, *F-measure* and *ROC curve* performance metrics (Fielding and Bell 1997).

Table 1: Outline of a confusion matrix

Actual/Ground Truth	Classified as	
	True	False
True	True positive (TP)	False positive (FP)
False	False negative (FN)	True negative (TN)

Precision and *recall* also provide additional insights into the robustness of the resulting data mining model. *Precision* indicates the type I error in statistics while the *recall* indicates the type II error. The *precision* represents the number of true positives divided by the total number of elements labeled as belonging to the positive class (shown in Eq.10), while *recall* (shown in Eq.11) represents the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items that are not labeled as belonging to the positive class but should have been).

$$precision = \frac{TP}{TP + FP} \quad (10)$$

$$recall = \frac{TP}{TP + FN} \quad (11)$$

The F-measure combines *precision* and *recall* and aims to explain the weighted average value of the *precision* and *recall*, ranging from 0 (worst) to 1 (best). The calculation is shown in Eq.12. The last performance evaluation metric proposed in this methodology is the Receiver Operating Characteristic (ROC) curve. The ROC describes different model performances based on varying values of threshold so that the optimal operating point is selected. The best operating point might be chosen so that the classifier gives the best trade-off between the costs of failing to detect positives against the costs of raising false

alarms. These costs need not be equal; however this is a common assumption. Note that the best place to operate the classifier is usually the point on its ROC which lies on a 45 degree line closest to the upper-left corner (0, 1) of the ROC plot.

$$F = 2 * \frac{precision * recall}{precision + recall} \quad (12)$$

The PD evaluation would be quantified based on the proportion of “PD” frames (data points captured every 33ms that are classified by the data mining model as PD) to the total captured frames during, for example, a 5 second walk towards the multimodal sensor hardware. In addition, the relevant PD gait features are also identified from the selected model in order to explore the potential effective way to capture and evaluate PD gait patterns in the future. As a healthcare decision support system, the proposed methodology would quantify the predictive accuracy of the data mining models and identify novel feature combinations relevant to early stage PD diagnosis, all from a non-clinical location. The resulting model could then be used to “flag” patients with abnormal gait patterns for follow up PD consultation with physicians and healthcare decision makers, rather than wait for the disease to progress to significantly noticeable stages before visiting a physician, which at that point greater than 50% of the dopaminergic neurons have been lost (Gil and Manuel 2009).

4. PD DETECTION STUDY

A study including both PD patients and controls is presented in this section to demonstrate the feasibility of the proposed data mining driven methodology. Fourteen subjects (seven PD patients and seven controls) were recruited from a tertiary movement disorders clinic for the data mining model generation and validation stage (stage 1 validation explained in step 4, section 3). In addition to the fourteen subjects, three additional subjects (one control and two patients diagnosed with showing signs of PD) were recruited for the stage 2 validation in order to quantify the data mining algorithms’ predictive power for unseen data. PD diagnosis was established by a movement disorders specialist (prior to the start of this study) using published criteria (Hughes, Ben-Shlomo, et al. 1992; Hughes, Daniel, et al. 1992). Prior to the start of the study, subjects were evaluated and confirmed free of major and acute neurological disorders,

hypothyroidism, vitamin B12 and folate deficiencies, and kidney and liver disease. Written informed consent was obtained for each subject, in accordance with the Declaration of Helsinki. The research protocol was reviewed and approved by the Penn State Hershey Medical Center Institutional Review Board (IRB protocol number 28989). The details relating to the experimental set up and sensor data acquisition are presented in Step 1. Data preprocessing is presented in Step 2, followed by data mining knowledge discovery in Step 3. Finally, the practicality of the proposed methodology as a decision support system is discussed in Step 4.

Step 1: Sensor Data Acquisition

In the experiment, the human gait movement was captured through the Microsoft Kinect multimodal sensor which results in capturing 3D skeleton images in a non-invasive manner. The MS Kinect is an off the shelf, (relatively) low cost sensor that is capable of capturing multimodal sensor data with the appropriate Application Programming Interfaces (APIs) and drivers installed. Although initially released by the Microsoft Corporation as a wireless controller for their Xbox gaming console, the Microsoft Kinect sensor has also been applied for research with motion tracking and virtual reality applications (Gonçalves et al. 2012).

During the experiment, the Kinect was configured at an elevation of 3 feet and 10 inches above the floor as seen in Figure 3. Then, the whole-body representation of each subject was verified and the sensor angle calibrated by having the subject stand relaxed while facing the Kinect at a distance of 10 feet. Each subject was then instructed to walk forward (front), which took 4-6 seconds, depending on the subject. Therefore, collecting one new data point every 33 ms, each subject generated an average of 206 data points (individual rows in the data table in Figure 2) during the entire 4-6 second data collection run. In the experiments, subjects were asked to first take 2-3 steps backward (4 feet) from the point of sensor calibration, still remaining within the distance limit of the device. Then, subjects were instructed to walk

comfortably towards the Kinect and were not given any specific instructions regarding side of initiation. The overall Kinect setup diagram is shown in Figure 3.

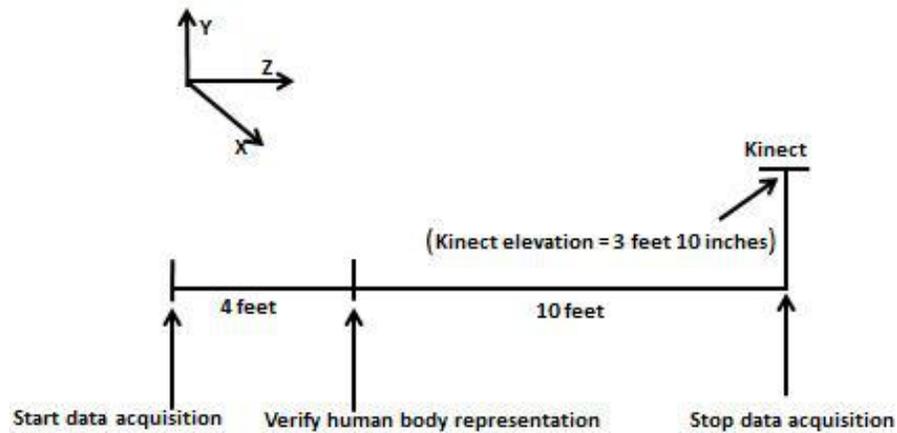


Figure 3: PD Data Acquisition Experiment Setup

Step 2: Data Preprocessing

Once the data collection is complete, the data is cleaned to reduce noise in the original data set and extract the correlated feature set. For example, using the Kinect for human gait data acquisition, data such as *tracking status* (e.g., tracked or not tracked), *user ID* and *frame number* are initially captured in the original data set and hence can be removed. The cleaning includes several components as follows.

1. The velocity and acceleration of each node are generated in X, Y and Z coordinate space, with the position of the sensor taken as reference point (i.e., (0, 0, 0)).

2. Position, velocity and acceleration ratios are generated between each of the nodes in X, Y and Z coordinates. Here X is the horizontal direction where the right direction of the zero point is considered as positive. For the Y direction, the upward direction is considered the positive direction. For the Z direction, subjects walking towards the Kinect, represent the positive direction.

3. For the binary class variables, PD is considered a subject that has Parkinson's disease (value of 1) and CONTROL is considered a non-Parkinson's disease subject (value of 0);

4. Two data sets are generated called PD-OFF data set and CONTROL data set, with 1890 features included in each data set. The PD-OFF data set represents data generated by the seven PD patients. If a PD patient was on medication, they were taken off medication for at least 1 hour prior to the experiment so that their natural gait patterns could be observed, without being masked by any PD medications. The CONTROL data set represents the gait data captured from the control group comprising of seven subjects.

Step 3: Data Mining Knowledge Discovery

Once the data acquisition and preprocessing steps are complete, the Waikato Environment for Knowledge Analysis (WEKA) software (Weka 2012) is used to discover novel, previously unknown features pertaining to patient gait by employing the aforementioned data mining/machine learning algorithms. The WEKA default parameters for each of the data mining predictive models are used to generate the predictive models. The proposed methodology is able to model and predict PD gait patterns based on the selected features presented in the previous subsection. Furthermore, the researchers are also able to evaluate performances of different algorithms in order to select the best classifier/model for healthcare decision support and obtain more insights into early stage PD diagnosis. The authors perform a 10-fold cross validation in order to measure the robustness of the data mining models to consistently predict the correct class variable, while varying the test data used to validate the models. Empirical studies have shown that 10 fold cross validation is a good measure of model predictive consistency (Kohavi 1995).

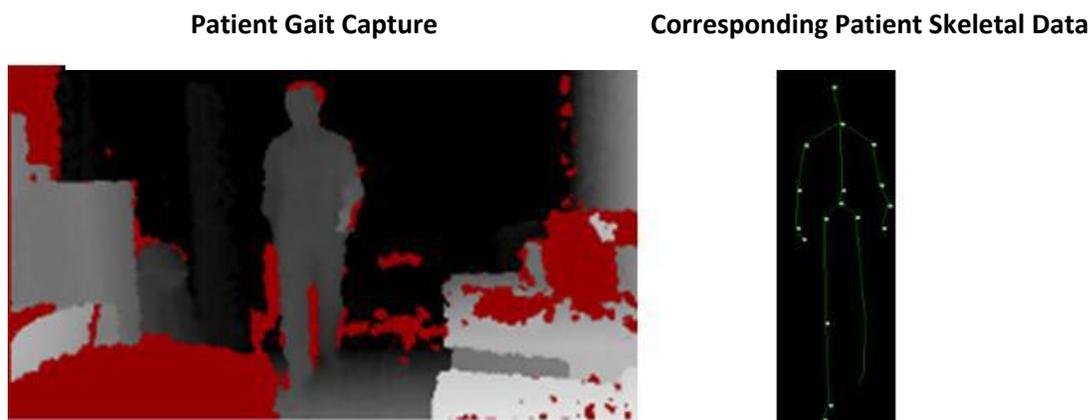


Figure 4: Visualization of Non-invasive Patients' Gait Data Collection

Figure 4 presents a visualization of the experiment in progress. As the patient walks towards the sensor hardware, non-identifiable depth sensor data is collected, with the corresponding skeleton data (right side of Figure 4), used for the actual data mining pattern discovery. After a 4-6 second run, the algorithm is able to classify a subject as either exhibiting symptoms of PD (determined by the relevant features in the data mining model) or a control group. Step 4 of the case study focuses on the veracity of the model in order to demonstrate the viability of the proposed methodology in being integrated into the healthcare system as an early warning/disease progression/management tool.

Step 4: Healthcare Decision Support

In order to investigate the robustness of the proposed data mining algorithms in predicting PD patients from controls, the authors repeated the study, varying the walking patterns of both the PD patients and controls (Figure 5). The original study (walking forward towards the sensor) was used as the baseline for comparison. Three additional experiments were conducted: i) walking away from the sensor, ii) walking to the right of the sensor and iii) walking to the left of the sensor. For each of these studies, the skeleton data captured by the non-invasive sensor may differ. For example, the sensor data captured by a subject walking towards the sensor would include symmetrical data points from both the left and right sides of the body. However, the sensor data captured by a subject walking to the right of the sensor would primarily capture the right side body data, leaving the left side body data primarily concealed from the view of the sensor.

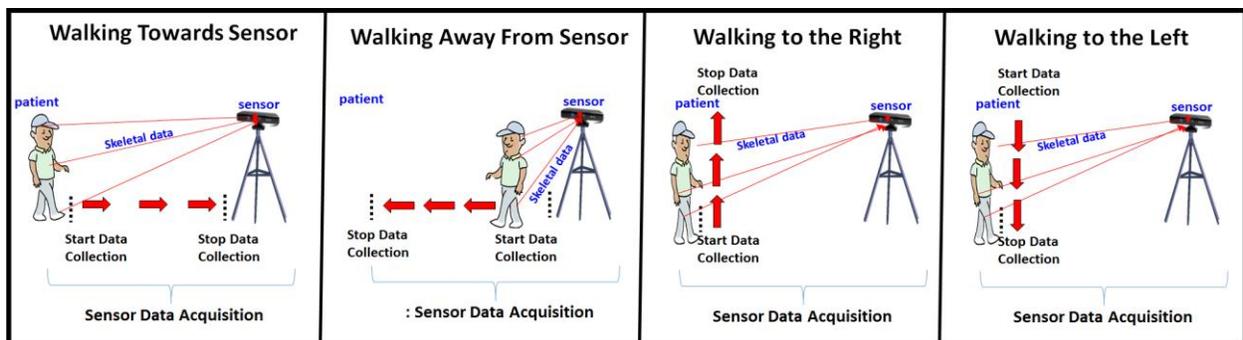


Figure 5: Extension of Study to Include Variations in Walking Directions

The PD detection study presents results involving the two validation stages explained in Section 3 of this work: i) validation of data mining algorithms through n -fold cross validation using sample data to both train and test the model and ii) validation of model predictive power using unseen data from both PD patients and controls.

Stage 1: Validation of Data Mining Algorithms through 10-fold Cross Validation

By testing the resulting data mining models against the unseen test sample data using 10-fold cross validation, the authors are able to identify the most accurate and reliable data mining models that facilitate PD gait assessment. In addition, the relevant features are also extracted, which may serve as better evidence to measure early stage PD gait patterns. To demonstrate the potential for gait data to be captured at multiple walking positions, beyond simply walking forward and walking backwards, the authors conducted experiments where participants walked to the left and right of the sensor for 4-6 seconds. Data mining models were then generated, based on both symmetric (i.e., walking forward and backward) and asymmetric (walking left and right) data sets.

Table 2: Performance of the classification algorithms in the Forward Walking experiment

Algorithm	Actual/Ground Truth	Classified as			Accuracy
		PD	Control	Sum	
IBK		PD	Control	Sum	95.4 %
	PD	1195	62	1257	
	Control	64	1422	1486	
	Sum	1259	1484	2743	
Naïve Bayes		PD	Control	Sum	50.6%
	PD	724	533	1257	
	Control	822	664	1486	
	Sum	1546	1197	2743	
SVM		PD	Control	Sum	74.0%
	PD	851	406	1257	
	Control	306	1180	1486	
	Sum	1157	1586	2743	
J48 Decision Tree		PD	Control	Sum	88.0%
	PD	1110	147	1257	
	Control	181	1305	1486	

	Sum	1291	1452	2743		
Random Forest		PD	Control	Sum	66.3%	
		PD	623	634		1257
		Control	290	1196		1486
		Sum	913	1830		2743

The results from Step 1 validation indicate that the IBK classifier has the highest accuracy (95.4%), based on 10-fold cross validation. The accuracies of the C4.5 (represented as J48 in WEKA) and SVM models were also high with 88% and 74% respectively. That is, these three models are able to correctly recognize between 74%-95% of the PD frames among the 14 subjects. The lowest accuracy was from the Naïve Bayes and Random Forest models (50.6% and 66.3% respectively), indicating that there may exist correlations between the samples or features not captured by these algorithms. For example, the Naïve Bayes algorithm assumes independence of all features. The results from the forward walking experiment reveal that IBK and J48 are the most suitable classifiers in quantifying correlations between the input space (i.e., features) and the class variable (i.e., PD classification) as seen in Tables 2 and 3.

Table 3: Performance indicators of various classification algorithms (Forward Walking)

Algorithm	TP rate	FP rate	Precision	Recall	F-Measure	ROC
IBK	95.4%	4.6%	95.4%	95.4%	95.4%	95.0%
Naïve Bayes	50.6%	48.3%	51.5%	50.6%	50.5%	51.4%
SVM	74.0%	26.9%	74.0%	74.0%	73.9%	73.6%
J48 Decision Tree	88.0%	11.9%	88.1%	88.0%	88.1%	88.3%
Random Forest	66.3%	36.3%	66.7%	66.3%	65.4%	73.6%

Similar to the results conducted with participants walking towards the sensor (Tables 2 and 3), the results from participants walking away from the sensor (Table 4), to the left of the sensor (Table 5) and to the right of the sensor (Table 6), reveal high predictive accuracy of the J48 algorithm (using 10 fold cross validation). While the IBK algorithm had high predictive accuracy for the forward and back walking experiments, it performed poorly on the left and right walking experiments with 59.2% and 57.9% respectively.

Table 4: Performance indicators of various classification algorithms (Back Walking)

Algorithm	TP rate	FP rate	Precision	Recall	F-Measure	ROC
IBK	91.5%	8.6%	91.6%	91.5%	91.5%	91.3%
Naïve Bayes	56.7%	45.5%	58.8%	56.7%	52.2%	56.1%
SVM	72.5%	28.0%	72.8%	72.5%	72.3%	72.2%
J48 Decision Tree	94.1%	6.0%	94.1%	94.1%	94.1%	94.5%
Random Forest	68.6%	32.1%	68.9%	68.6%	68.3%	75.9%

Results from Tables 3-6 reveal that despite the symmetric (forward and back walking) and asymmetric (left and right walking) characteristics of the data points, the J48 algorithm was able to distinguish gait variations between PD patients and controls with accuracies ranging from 88%-94%. The Naïve Bayes and SVM algorithms have NaN values in Table 5 and 6 because they failed to generate results in the allocated time. The consistency of the J48 algorithm when i) test data subsets are varied through 10-fold cross validation and ii) walking positions are varied, serves as a guide for healthcare decision makers seeking to investigate which machine learning algorithms are suitable for PD detection of unseen samples (i.e., stage 2 validation).

Table 5: Performance indicators of various classification algorithms (Left Walking)

Algorithm	TP rate	FP rate	Precision	Recall	F-Measure	ROC
IBK	59.3%	54.3%	76.2%	59.3%	46.2%	53.0%
Naïve Bayes	N/A	N/A	N/A	N/A	N/A	N/A
SVM	N/A	N/A	N/A	N/A	N/A	N/A
J48 Decision Tree	91.3%	9.3%	91.3%	91.3%	91.3%	91.4%
Random Forest	70.0%	35.2%	70.4%	70.0%	68.7%	77.8%

Table 6: Performance indicators of various classification algorithms (Right Walking)

Algorithm	TP rate	FP rate	Precision	Recall	F-Measure	ROC
IBK	57.9%	48.3%	59.4%	57.9%	51.7%	55.8%
Naïve Bayes	N/A	N/A	N/A	N/A	N/A	N/A
SVM	N/A	N/A	N/A	N/A	N/A	N/A

J48 Decision Tree	93.6%	6.6%	93.6%	93.6%	93.6%	93.6%
Random Forest	71.7%	31.5%	73.3%	71.7%	70.6%	79.0%

Future work will investigate the theoretical aspects inherent in these algorithms that make them suitable for PD classification. The rate of data capture by the sensor (i.e., one new data instances every 33ms) means that classification algorithms can be iteratively retrained on new incoming data, hereby making the resulting data mining models, a function of the constantly increasing data streams of PD patients and controls.

Stage 2: Validation of Data Mining Models’ Predictive Power Using Unseen Patient Data

Based on the results from Stage 1 validation, we are now able to understand how different data mining algorithms perform under varied experimental setups. Based on the results from Stage 1, the J48 decision tree algorithm has the most consistent predictive performance, relative to the sample data set containing 7 PD patients and 7 controls. Furthermore, the extracted gait features included in these models help to identify the important features in order to assist physicians in better understanding emerging gait features of PD patients. For example, the top 10 features of the J48 decision tree classifier model are: *WristRightZPosition*, *WristLeftXPosition*, *WristRightXPosition*, *KneeLeftXVelocity*, *KneeLeftXPosition*, *HeadXPosition*, *HandRightXPosition*, *HipRightYPosition*, *HipLeftYPosition*, and *ElbowLeftXVelocity*. It is important to note that the position features found relevant by the Decision Tree model in Figure 6, primarily relate to position in the X-direction. I.e., there exists differences in the lateral position of skeletal joints between PD and controls. These findings reveal that factors such as height variations (which would be more evident in the Y-direction) are not differentiating features between PD and controls. A subset of our results are consistent with previous research studies that reveal that reduced variations in arm swing and leg movement (i.e., relevant position attributes) are important features to recognize PD gait patterns (Marconi et al. 1994), (FitzGerald and Jankovic 1989). In addition, this research study has also successfully identified additional PD features that may be relevant to early PD detection. Other novel features include left elbow velocity in the X direction and several position features relating to the knee and hip. These features could explain

intensity of gait pattern instead of just position variations. A subset of the J48 Decision Tree model is shown in Figure 6, with the accompanying numeric values for each feature.

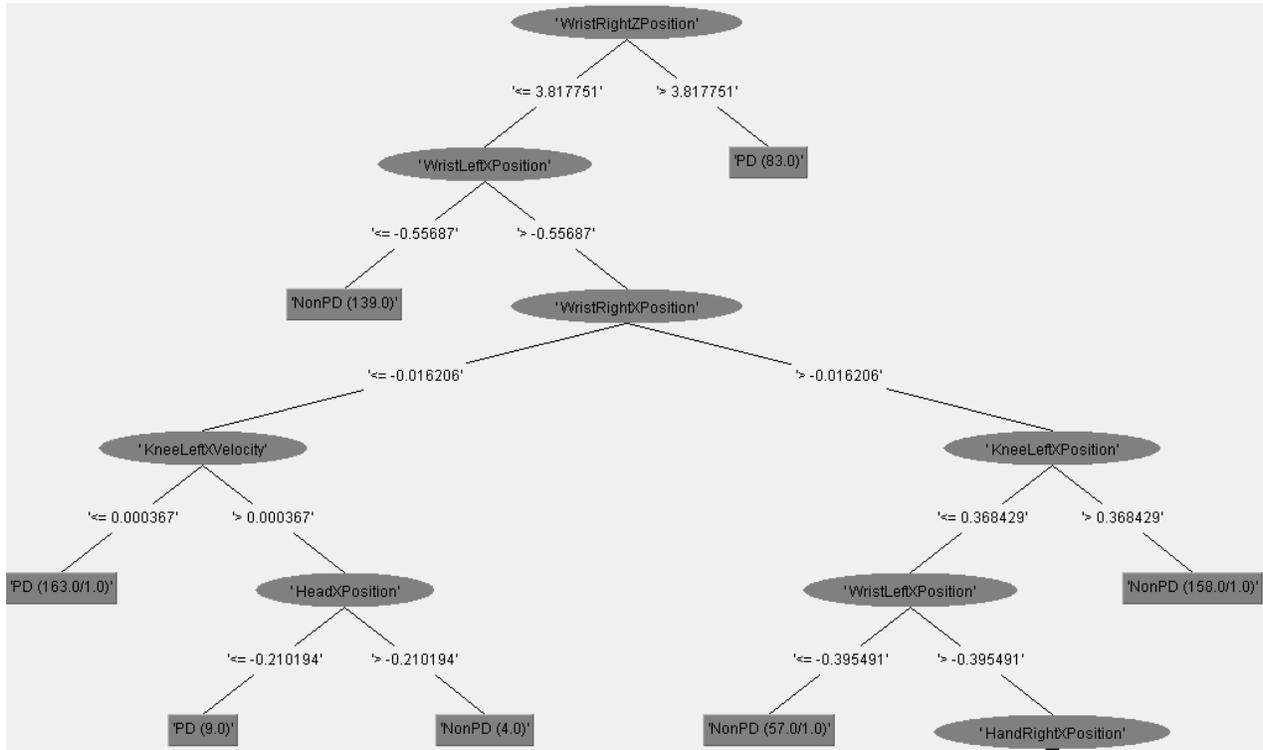


Figure 6: Data Mining J48 (C4.5) Decision Tree Model

For example, in the model shown in Figure 6, the most important feature (based on its hierarchy in the decision tree model) is Wrist Right Z Position. For example, for a given frame (captured at a 33 ms interval), if the Wrist Right X Position value is less than 3.817m, and the Wrist Left X Position is greater than -0.5568m, and the Wrist Right X Position is less than -0.0162 and the Knee Left X Velocity is less than 0.000367, then this individual is classified as a PD patient for that captured frame. For a given 4-6 second experimental run, an average of 206 data instances are captured per subject. A final PD prediction is based on the total number of frames being classified as PD vs. control) That is:

$$PD \text{ Classification} = \arg \max_{i \in \{1, \dots, Y\}} (p(y_i)) \quad (13)$$

Where,

Y: number of mutually exclusive class values. In this case, the classification is binary (PD, non-PD), although this can be extended to greater partitions (e.g., moderate PD, severe PD) in subsequent works

$p(y_i)$: represents the probability of the class value i

Table 7: Data Mining Model Predictive Accuracy Based on Unseen Data from Control Subject

Front			Back			LHS			RHS		
	Predicted			Predicted			Predicted			Predicted	
	PD	CONTROL		PD	CONTROL		PD	CONTROL		PD	CONTROL
Actual (CONTROL)	78 (50%)	78 (50%)	Actual (CONTROL)	27 (71.1%)	11 (28.9%)	Actual (CONTROL)	9 (7.6%)	110 (92.4%)	Actual (CONTROL)	50 (45.0%)	61 (55.0%)

Unseen Test Data for Subject 1 (CONTROL): The control subject used for the stage 2 validation represents an individual with normal body functions that has been cleared by a neurologist as being free of PD related gait symptoms. As can be observed in Table 7, the proposed methodology accurately predicts the control subject based on the left (LHS) and right (RHS) walking experiments with 92.4% and 55% respectively. However, the front and back walking experiments does not predict the control subject, resulting in 50% and 28.9% predictive accuracies respectively, as can be seen in Table 7. The numerical values to the left of each percentage values in Tables 7 and 8, represent the raw instance counts (i.e., tuples in the data set).

Unseen Data for Subjects 2 and 3 (PD at risk patients): Subjects 2 and 3 in Table 8 represent actual patients that have been flagged by an expert neurologist as exhibiting signs of PD. Their classification of PD at risk was characterized prior to this study. Typically, a follow up examination is required, which may include a DaTscan to measure dopamine levels in the brain. The stage 2 validation of the proposed model aims to determine whether the J48 decision tree model generated for a particular walking direction, has predictive power in classifying an individual as having PD or not.

Table 8: Data Mining Model Predictive Accuracy Based on Unseen Data from PD at Risk Patients

Front			Back			LHS			RHS		
	Predicted			Predicted			Predicted			Predicted	
	PD	CONTROL		PD	CONTROL		PD	CONTROL		PD	CONTROL
Actual (PD Patient 1)	104 (61.9%)	64 (38.1%)	Actual (PD Patient 1)	92 (89.3%)	11 (10.7%)	Actual (PD Patient 1)	52 (44.8%)	64 (55.2%)	Actual (PD Patient 1)	17 (12.4%)	120 (87.6%)
Actual (PD Patient 2)	94 (66.7%)	47 (33.3%)	Actual (PD Patient 2)	49 (67.1%)	24 (32.9%)	Actual (PD Patient 2)	29 (22.3%)	101 (77.7%)	Actual (PD Patient 2)	52 (35.9%)	93 (64.1%)

Table 8 presents a summary of the J48 Data Mining models when tested on the PD at risk patient 1 and PD at risk patient 2. As can be seen from the results in Table 8, both the front and back experiments provide consistent classification of PD patients as actually being PD. For example, for the forward walking position for PD at risk patient 1, 104 unseen tuples (61.9%) pertaining to this patient’s gait, were classified as exhibiting characteristics of a PD patient, compared to 64 unseen tuples (38.1%) classified as exhibiting characteristics of a non-PD (i.e., control) subject. For the back walking position for PD at risk patient 1, 92 unseen tuples (89.3%) pertaining to this patient’s gait, were classified as exhibiting characteristics of a PD patient, compared to 11 unseen tuples (10.7%) classified as exhibiting characteristics of a non-PD (i.e., control) subject. Similar results can be observed for the forward and backward walking positions of PD at risk patient 2 in Table 8. However, the J48 algorithm does not accurately predict PD patients when the LHS and RHS experiments are conducted, as seen in Table 8. This may be attributed to the fact that the LHS and RHS walking experiments primarily capture the skeleton data of a subject’s profile view, instead of a symmetric data collection such as those of the front and back walking experiments. This hypothesis is

consistent with the literature that has found the absence of symmetrical gait to be an indicator of the emergence of PD (Huang et al. 2012).

Table 7 and 8 can be used to determine the walking direction with the most consistent predictive accuracy across participants (both PD and controls). I.e., the aggregated front walk experiments (summing down using Tables 7 and 8) correctly classifies 276 of the total 465 tuples captured from both the CONTROL and PD patients, yielding an overall classification accuracy of 59.3%. The aggregated back walking experiments correctly classifies 152 of the total 214 instances captured from both the CONTROL and PD patients, yielding an overall classification accuracy of 71.0%. The aggregated LHS experiments correctly classifies 191 of the total 365 instances captured from both the CONTROL and PD patients, yielding an overall classification accuracy of 52.3%. The aggregated RHS experiments performed the worst, only correctly classifying 130 of the total 393 instances captured from both the CONTROL and PD patients, yielding an overall accuracy of 33.07%. If the authors are to select only one walking direction for all subjects (e.g., front), then the results would indicate that two out of the three subjects were accurately classified (i.e., the PD patients). For the front walking direction for the CONTROL subject, the methodology has a predictive accuracy of only 50%. The authors postulate that the challenge in accurately predicting any one subject accurately, regardless of walking direction is due to the class imbalance problem which is a well-known issue in the machine learning research community that affects the robustness of machine learning classifiers (Japkowicz and Stephen 2002). I.e., there are fourteen subjects selected for this study, with 50% of them being PD patients. The prevalence of PD in the United States population is much lower at 0.3% (“National Parkinson Foundation - Parkinson’s Disease” 2015). Therefore, the skewness of the training data set may be reducing the robustness of the resulting data mining models. The authors included seven CONTROLS (e.g., instead of only one) in order to have a large enough sample to train the machine learning model. However, this issue can be resolved by increasing the sample size to accommodate more CONTROLS in the training data set, without altering the steps within the methodology.

Instead of determining which single walking position (front, back, LHS, RHS) generalizes to include both CONTROL and PD patients, a subject (either CONTROL or PD) could walk in all four directions, with a final disease classification made by selecting the class that has the highest distribution of assigned tuples. Based on this criteria, the CONTROL subject in Table 7 would be classified as non-PD with 260 tuples correctly classified as CONTROL and 164 tuples misclassified as PD (i.e., 61.3% correctly classified instances). For PD at risk patient 1, aggregating their front, back, LHS and RHS walking directions results in 265 tuples correctly classified as PD, with the remaining 259 tuples incorrectly classified as CONTROL (i.e., 50.6% correctly classified instances). For the PD at risk patient 2, aggregating their front, back, LHS and RHS results in 224 tuples correctly classified as PD, while the remaining 265 tuples are misclassified as CONTROL (i.e., 45.8% correctly classified instances). Therefore employing the *argmax* function for each subject, given their four walking directions, would result in two (CONTROL and PD at risk patient 1) out of the three subjects being correctly classified. However, the decision to make disease predictions based on a specific walking direction or across all walking directions will be further explored in future work with a larger population of subjects.

To put this research study into perspective, given fourteen subjects to train the data mining model and three additional subjects to test its validity and robustness, the proposed methodology was able to correctly classify two out of the three test subjects using the *argmax* function for classifying across all walking directions. Furthermore, each subject only had to walk in front of the sensor system for approximately 4-6 seconds per experiment, resulting in a total of approximately 25 seconds worth of data collected for each patient. For a patient sample of 14 subjects, this amounts to less than 10 minutes of data collection that is demonstrating evidence of predictive power. This paper aims to demonstrate the feasibility of the proposed methodology in capturing gait data in a non-invasive manner, employing data mining models that are robust across different walking experiments and demonstrate the model's predictive power using unseen data captured by additional subjects.

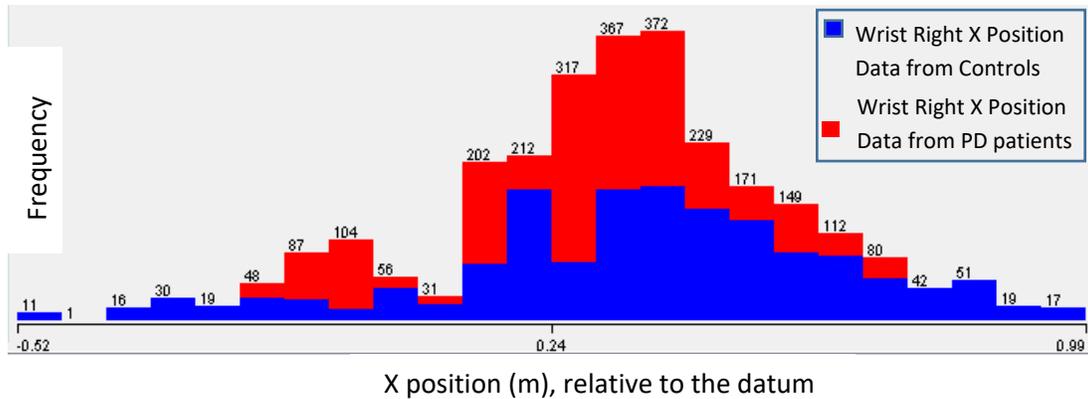


Figure 7: Histogram of Wrist Right X Position between PD and Controls

In addition to the predictive accuracies, our preliminary findings are consistent with previous related research and help provide evidence of significant gait differences between PD patients and controls (Huang et al. 2012). For example, the right hand wrist position in the X coordinate is plotted for the 7 PD patients and 7 controls in Figure 7 and reveals reduced arm swing in PD patients (red), compared to controls (blue). The visualization of the results is consistent with the research findings in (Huang et al. 2012), indicating a reduced arm swing in PD patients since the range of red (represent arm swing range in PD patients) is less than the range of blue (represent arm swing range in controls).

Based on the proposed methodology, human gait can be captured and modeled using non-invasive techniques in an effort recognize PD patients from controls. By generating data mining predictive models based on the gait data captured by the non-invasive sensors, researchers are able to determine the predicted status (i.e., PD or not PD) of each observation (i.e., frame), and an early stage PD. The final prediction would be the proportion of PD frames to the total number of frames collected. In terms of potential clinical application, this research aims to demonstrate the feasibility of utilizing non-invasive sensors and data mining algorithms to predict PD from non-hospital settings. For example, potential PD patients may be able to have PD gait screening, prior to a hospital visit. The proposed methodology serves as a quantitative reference for flagging patients that may be developing PD. While the current study was limited by sample size, our results are in agreement with previous characterizations of gait impairment in early PD. Specifically, several studies have suggested that gait impairment in early PD may involve reduced swing

of the upper limbs during ambulation (Huang et al. 2012; Lewek et al. 2010; Isaias et al. 2012; Crenna et al. 2008; Carpinella et al. 2007; Roggendorf et al. 2012) . Moreover, the agreement of our results with previous studies supports the notion that whole-body data mining techniques may be used to reveal novel markers of gait dysfunction. Future remote sensing studies using larger sample sizes may offer new insights regarding high-level characteristics of gait symptoms and PD progression.

5. CONCLUSION

In this paper, a data mining based methodology is proposed to analyze and predict PD patients from controls through the low-cost, non-invasive sensor hardware. The proposed methodology enables researchers and physicians to utilize low cost, non-invasive data acquisition techniques to model and predict the emergence of Parkinson's disease. The proposed methodology maintains patient privacy by only collecting the three-dimensional (3D) geospatial data pertaining to a patient's node location. The proposed methodology is also successful at extracting relevant gait features in order to model and predict early stage PD. The researchers in this work aim to provide healthcare decision makers with the ability to quantify PD movement abnormalities, independent of clinic location (e.g., from a patient's home), hereby serving as an early warning system that can be communicated to physicians and healthcare decision makers, towards telehealth solutions.

In future work, the proposed methodology may be expanded to other neurodegenerative disorders and possible medication effectiveness evaluation. In addition, this decision support may serve as a non-location based screening system hereby detecting serious health conditions at their infancy and recommending individuals for follow up consultations with their healthcare providers.

6. ACKNOWLEDGEMENTS

This research is funded through the *NSF I/UCRC Center for Healthcare Organization Transformation (CHOT)*, NSF I/UCRC grant # 1067885.

REFERENCES

- A, David G I L, and Magnus Johnson B. 2006. "Diagnosing Parkinson by Using Artificial Neural Networks and Support Vector Machines," 63–71.
- Amasyal, M Fatih, and Banu Diri. 2006. "Automatic Turkish Text Categorization in Terms of Author, Genre and Gender." In *Natural Language Processing and Information Systems*, 221–26. Springer.
- Ambroise, Christophe and McLachlan, Geoffrey J. 2002. "Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data." *Proceedings of the National Academy of Sciences* 99 (10): 6562–66.
- Bartels, Anna L., Yacov Balash, Tanya Gurevich, Joanna D. Schaafsma, Jeffrey M. Hausdorff, and Nir Giladi. 2003. "Relationship between Freezing of Gait (FOG) and Other Features of Parkinson's: FOG Is Not Correlated with Bradykinesia." *Journal of Clinical Neuroscience* 10 (5): 584–88.
- Barth, Jens, Michael Sunkel, Katharina Bergner, Gerald Schickhuber, Jurgen Winkler, Jochen Klucken, and Bjorn Eskofier. 2012. "Combined Analysis of Sensor Data from Hand and Gait Motor Function Improves Automatic Recognition of Parkinson's Disease." In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, 5122–25.
- Behoorra, Ishan, and Conrad S. Tucker. 2014. "Quantifying Emotional States Based on Body Language Data Using Non Invasive Sensors." In *ASME 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, V01AT02A079–V01AT02A079. American Society of Mechanical Engineers.
- Behoorra, I., and C. Tucker. 2015. "Machine Learning Classification of Design Team Members' Body Language Patterns For Real Time Emotional State Detection." *Design Studies* 39 (1): 100–127. doi:10.1016/j.destud.2015.04.003.
- Bonato, Paolo, Delsey M Sherrill, David G Standaert, Sara S Salles, and Metin Akay. 2004. "Data Mining Techniques to Detect Motor Fluctuations in Parkinson's Disease." *Conference Proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference 7* (January): 4766–69. doi:10.1109/IEMBS.2004.1404319.
- Carpinella, Ilaria, Paolo Crenna, Alberto Marzegan, Marco Rabuffetti, Mario Rizzone, Leonardo Lopiano, and Maurizio Ferrarin. 2007. "Effect of L-Dopa and Subthalamic Nucleus Stimulation on Arm and Leg Swing during Gait in Parkinson's Disease." In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, 6664–67. IEEE.
- Clark, Ross A., Yong-Hao Pua, Karine Fortin, Callan Ritchie, Kate E. Webster, Linda Denehy, and Adam L. Bryant. 2012. "Validity of the Microsoft Kinect for Assessment of Postural Control." *Gait & Posture* 36 (3): 372–77.
- Crenna, PAOLO, I. Carpinella, L. Lopiano, A. Marzegan, M. Rabuffetti, M. Rizzone, M. Lanotte, and M. Ferrarin. 2008. "Influence of Basal Ganglia on Upper Limb Locomotor Synergies. Evidence from Deep Brain Stimulation and L-DOPA Treatment in Parkinson's Disease." *Brain* 131 (12): 3410–20.
- Critchley, E M. 1981. "Speech Disorders of Parkinsonism: A Review." *Journal of Neurology, Neurosurgery & Psychiatry* 44 (9). BMJ Publishing Group Ltd: 751–58.
- Darley, Frederic L, Arnold Elvin Aronson, and Joe Robert Brown. 1975. *Motor Speech Disorders*. Vol. 304. Saunders Philadelphia.
- Dastgheib, Zeinab A, Brian Lithgow, and Zahra Moussavi. 2011. "Application of Fractal Dimension on Vestibular Response Signals for Diagnosis of Parkinson's Disease." In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, 7892–95.
- "Developers Purchase Online Get Your Sensor - Primesense." 2013. *PrimeSense*. Accessed September 10. <http://www.primesense.com/developers/get-your-sensor/>.

- D'heygere, Tom, Peter L M Goethals, and Niels De Pauw. 2003. "Use of Genetic Algorithms to Select Input Variables in Decision Tree Models for the Prediction of Benthic Macroinvertebrates." *Ecological Modelling* 160 (3). Elsevier: 291–300.
- Domingos, Pedro, and Geoff Hulten. 2000. "Mining High-Speed Data Streams." In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 71–80. ACM.
- Fahn, Stanley. 2003. "Description of Parkinson's Disease as a Clinical Syndrome." *Annals of the New York Academy of Sciences* 991 (1): 1–14.
- Fielding, Alan H, and John F Bell. 1997. "A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/absence Models." *Environmental Conservation* 24 (1). Cambridge Univ Press: 38–49.
- FitzGerald, P M, and J Jankovic. 1989. "Lower Body Parkinsonism: Evidence for Vascular Etiology." *Movement Disorders : Official Journal of the Movement Disorder Society* 4 (3): 249–60. doi:10.1002/mds.870040306.
- Freed, Curt R., et al. 1992. "Survival of Implanted Fetal Dopamine Cells and Neurologic Improvement 12 to 46 Months after Transplantation for Parkinson's Disease."
- Gil, David, and Devadoss Johnson Manuel. 2009. "Diagnosing Parkinson by Using Artificial Neural Networks and Support Vector Machines." *Global Journal of Computer Science and Technology* 9 (4).
- Gonçalves, Nuno, José L Rodrigues, Sandra Costa, and Filomena Soares. 2012. "Automatic Detection of Stereotypical Motor Movements." *Procedia Engineering* 47. Elsevier: 590–93.
- Hastie, Trevor, Robert Tibshirani, Jerome Friedman, and James Franklin. 2005. "The Elements of Statistical Learning: Data Mining, Inference and Prediction." *The Mathematical Intelligencer* 27 (2). Springer: 83–85.
- Ho, Aileen K, Robert Ianssek, Caterina Marigliani, John L Bradshaw, and Sandra Gates. 1999. "Speech Impairment in a Large Sample of Patients with Parkinson's Disease." *Behavioural Neurology* 11 (3). IOS Press: 131–37.
- Hoff, J I, E A H Wagemans, J J van Hilten, and others. 2001. "Accelerometric Assessment of Levodopa-Induced Dyskinesias in Parkinson's Disease." *Movement Disorders* 16 (1). Wiley Online Library: 58–61.
- Huang, Xuemei, Peter Chen, Daniel I Kaufer, Alexander I Troster, and Charles Poole. 2006. "Apolipoprotein E and Dementia in Parkinson Disease: A Meta-Analysis." *Archives of Neurology* 63 (2). Am Med Assoc: 189.
- Huang, Xuemei, Joseph M Mahoney, Mechelle M Lewis, Guangwei Du, Stephen J Piazza, and Joseph P Cusumano. 2012. "Both Coordination and Symmetry of Arm Swing Are Reduced in Parkinson's Disease." *Gait & Posture* 35 (3). Elsevier B.V.: 373–77. doi:10.1016/j.gaitpost.2011.10.180.
- Hughes, Andrew J, Yoav Ben-Shlomo, Susan E Daniel, and Andrew J Lees. 1992. "What Features Improve the Accuracy of Clinical Diagnosis in Parkinson's Disease A Clinicopathologic Study." *Neurology* 42 (6). AAN Enterprises: 1142.
- Hughes, Andrew J., Susan E. Daniel, Linda Kilford, and Andrew J. Lees. 1992. "Accuracy of Clinical Diagnosis of Idiopathic Parkinson's Disease: A Clinico-Pathological Study of 100 Cases." *Journal of Neurology, Neurosurgery & Psychiatry* 55 (3): 181–84.
- Hulten, Geoff, Laurie Spencer, and Pedro Domingos. 2001. "Mining Time-Changing Data Streams." In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 97–106. ACM.
- Isaias, Ioannis U., Jens Volkman, Alberto Marzegan, Giorgio Marotta, Paolo Cavallari, and Gianni Pezzoli. 2012. "The Influence of Dopaminergic Striatal Innervation on Upper Limb Locomotor Synergies." *PloS One* 7 (12): e51464.
- Japkowicz, Nathalie, and Shaju Stephen. 2002. "The Class Imbalance Problem: A Systematic Study." *Intelligent Data Analysis* 6 (5): 429–49.

- Kamel Boulos, Maged N. 2012. "Xbox 360 Kinect Exergames for Health." *Games for Health: Research, Development, and Clinical Applications* 1 (5): 326–30.
- Kohavi, Ron. 1995. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." *IJCAI* 14 (2): 1137–45.
- Lai, Kam, J. Konrad, and P. Ishwar. 2012. "A Gesture-Driven Computer Interface Using Kinect." In *2012 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, 185–88. doi:10.1109/SSIAI.2012.6202484.
- Lang, Anthony E, and Andres M Lozano. 1998. "Parkinson's Disease." *New England Journal of Medicine* 339 (15). Mass Medical Soc: 1044–53.
- Lawrence, Andrew D, Andrew H Evans, and Andrew J Lees. 2003. "Compulsive Use of Dopamine Replacement Therapy in Parkinson's Disease: Reward Systems Gone Awry?" *44* (0): 595–604.
- Lewek, Michael D., Roxanne Poole, Julia Johnson, Omar Halawa, and Xuemei Huang. 2010. "Arm Swing Magnitude and Asymmetry during Gait in the Early Stages of Parkinson's Disease." *Gait & Posture* 31 (2): 256–60.
- Lewis, M M, C G Slagle, a B Smith, Y Truong, P Bai, M J McKeown, R B Mailman, a Belger, and X Huang. 2007. "Task Specific Influences of Parkinson's Disease on the Striato-Thalamo-Cortical and Cerebello-Thalamo-Cortical Motor Circuitries." *Neuroscience* 147 (1): 224–35. doi:10.1016/j.neuroscience.2007.04.006.
- Lewis, S J G, T Foltynie, a D Blackwell, T W Robbins, a M Owen, and R a Barker. 2005. "Heterogeneity of Parkinson's Disease in the Early Clinical Stages Using a Data Driven Approach." *Journal of Neurology, Neurosurgery, and Psychiatry* 76 (3). BMJ Publishing Group Ltd: 343–48. doi:10.1136/jnnp.2003.033530.
- Little, Max A, Patrick E McSharry, Eric J Hunter, Jennifer Spielman, and Lorraine O Ramig. 2009. "Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease." *Biomedical Engineering, IEEE Transactions on* 56 (4). IEEE: 1015–22.
- Manohar, Gautam, and Conrad Tucker. 2013. "A Privacy Preserving Data Mining Methodology for Dynamically Predicting Emerging Human Threats." In *ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, V02AT02A069–V02AT02A069. American Society of Mechanical Engineers.
- Marconi, R, D Lefebvre-Caparros, a M Bonnet, M Vidailhet, B Dubois, and Y Agid. 1994. "Levodopa-Induced Dyskinesias in Parkinson's Disease Phenomenology and Pathophysiology." In *Movement Disorders: Official Journal of the Movement Disorder Society*, 9:2–12. doi:10.1002/mds.870090103.
- Maroof, David Aaron. 2012. "Binary Logistic Regression." In *Statistical Methods in Neuropsychology*, 67–75. Springer.
- Martinez-Martin, P, A Gil-Nagel, L Morlán Gracia, J Balseiro Gomez, J Martínez-Sarriés, and F Bermejo. 1994. "Unified Parkinson's Disease Rating Scale Characteristics and Structure." *Movement Disorders* 9 (1). Wiley Online Library: 76–83.
- "National Parkinson Foundation - Parkinson's Disease." 2015. Accessed June 27. <http://www.parkinson.org/parkinson-s-disease.aspx>.
- Olanow, C Warren. 2004. "The Scientific Basis for the Current Treatment of Parkinson's Disease." *Annu. Rev. Med.* 55. Annual Reviews: 41–60.
- Ozcift, Akin. 2012. "SVM Feature Selection Based Rotation Forest Ensemble Classifiers to Improve Computer-Aided Diagnosis of Parkinson Disease." *Journal of Medical Systems* 36 (4). Springer: 2141–47.
- Patel, Shyamal, Konrad Lorincz, Richard Hughes, Nancy Huggins, John Growdon, David Standaert, Metin Akay, Jennifer Dy, Matt Welsh, and Paolo Bonato. 2009. "Monitoring Motor Fluctuations in Patients with Parkinson's Disease Using Wearable Sensors." *Information Technology in Biomedicine, IEEE Transactions on* 13 (6). IEEE: 864–73.
- Quinlan, J. 1993. *C4.5 Programming for Machine Learning*. Edited by P. Langley. Morgan Kauffman.

- Rajput, A H, B Rozdilsky, A Rajput, and others. 1991. "Accuracy of Clinical Diagnosis in Parkinsonism-- a Prospective Study." *The Canadian Journal of Neurological Sciences. Le Journal Canadien Des Sciences Neurologiques* 18 (3): 275.
- Ramani, R. Geetha, and G. Sivagami. 2011. "Parkinson Disease Classification Using Data Mining Algorithms." *International Journal of Computer Applications* 9 (32).
- Richards, Joseph W., Dan L. Starr, Nathaniel R. Butler, Joshua S. Bloom, John M. Brewer, Arien Crellin-Quick, Justin Higgins, Rachel Kennedy, and Maxime Rischard. 2011. "On Machine-Learned Classification of Variable Stars with Sparse and Noisy Time-Series Data." *The Astrophysical Journal* 733 (1): 10. doi:10.1088/0004-637X/733/1/10.
- Roggendorf, J., S. Chen, S. Baudrexel, S. van de Loo, C. Seifried, and R. Hilker. 2012. "Arm Swing Asymmetry in Parkinson's Disease Measured with Ultrasound Based Motion Analysis during Treadmill Gait." *Gait & Posture* 35 (1): 116–20.
- Schrag, A, Y Ben-Shlomo, and N Quinn. 2002. "How Valid Is the Clinical Diagnosis of Parkinson's Disease in the Community?" *Journal of Neurology, Neurosurgery & Psychiatry* 73 (5). BMJ Publishing Group Ltd: 529–34.
- Shotton, Jamie, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. 2013. "Real-Time Human Pose Recognition in Parts from Single Depth Images." *Commun. ACM* 56 (1): 116–24. doi:10.1145/2398356.2398381.
- Stewart, Celia, Linda Winfield, Ann Hunt, Susan B Bressman, Stanley Fahn, Andrew Blitzer, and Mitchell F Brin. 1995. "Speech Dysfunction in Early Parkinson's Disease." *Movement Disorders* 10 (5). Wiley Online Library: 562–65.
- Tibshirani, Robert and Hastie, Trevor and Narasimhan, Balasubramanian and Chu, Gilbert. 2002. "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression." *Proceedings of the National Academy of Sciences* 99 (10): 6567–72.
- Tsanas, Athanasios. 2010. "New Nonlinear Markers and Insights into Speech Signal Degradation for Effective Tracking of Parkinson's Disease Symptom Severity." *Age (years)* 64 (8.1): 63–66.
- Tsanas, Athanasios, Max A Little, Patrick E McSharry, and Lorraine O Ramig. 2010. "Enhanced Classical Dysphonia Measures and Sparse Regression for Telemonitoring of Parkinson's Disease Progression." In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 594–97.
- Tsanas, Athanasios, Max A Little, Patrick E McSharry, Jennifer Spielman, and Lorraine O Ramig. 2012. "Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease." *Biomedical Engineering, IEEE Transactions on* 59 (5). IEEE: 1264–71.
- Tucker, Conrad, Ishan Behoora, Harriet Black Nembhard, Mechelle Lewis, Nicholas W Sterling, and Xuemei Huang. 2015. "Machine Learning Classification of Medication Adherence in Patients with Movement Disorders Using Non-Wearable Sensors." *Computers in Biology and Medicine*. Accessed September 10. doi:10.1016/j.combiomed.2015.08.012.
- Tucker, Conrad S., and Harrison M. Kim. 2011. "Trend Mining for Predictive Product Design." *Journal of Mechanical Design* 133 (11): 111008.
- Wang, Haixun, Wei Fan, Philip S. Yu, and Jiawei Han. 2003. "Mining Concept-Drifting Data Streams Using Ensemble Classifiers." In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 226–35. KDD '03. New York, NY, USA: ACM. doi:10.1145/956750.956778.
- Weka. 2012. "Weka 3 - Data Mining with Open Source Machine Learning Software in Java." <http://www.cs.waikato.ac.nz/ml/weka/>.
- "Xtion PRO LIVE - Multimedia - ASUS." 2013. Accessed September 10. http://www.asus.com/Multimedia/Xtion_PRO_LIVE/.
- Ye, Mao, Xianwang Wang, Ruigang Yang, Liu Ren, and M. Pollefeys. 2011. "Accurate 3D Pose Estimation from a Single Depth Image." In *2011 IEEE International Conference on Computer Vision (ICCV)*, 731–38. doi:10.1109/ICCV.2011.6126310.

- Youn, Seongwook, and Dennis McLeod. 2007. "A Comparative Study for Email Classification." In *Advances and Innovations in Systems, Computing Sciences and Software Engineering*, 387–91. Springer Netherlands.
- Zhang, Harry, and Jiang Su. 2004. "Naive Bayesian Classifiers for Ranking." In *Machine Learning: ECML 2004*, edited by Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, 501–12. Lecture Notes in Computer Science 3201. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-540-30115-8_46.