Smartphone Sensor Data Mining for Gait Abnormality Detection

Shaun Gallagher

BS Computer Science Fordham University 2013

MASTER'S THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF SCIENCES

IN THE DEPARTMENT OF COMPUTER SCIENCE

AT FORDHAM UNIVERSITY

New York
February 2014

Table of Contents

Acknowledgments

Thanks to all who made completion of this thesis possible.

Advisor
Dr. Gary Weiss

Defense Committee
Dr. Damian Lyons
Dr. Yanjun Li

Albert Einstein College of Medicine
Dr. Joseph Verghese
Dr. Roee Holtzer
Emmeline Ayers

WISDM Lab
Tony Pulickal

Chapter 1: Introduction

Smartphones are a ubiquitous part of daily life. They are integral to modern communication and are rarely far from hand. These small but powerful devices are equipped with a range of sensors that enable them to gather information about the world around them. Among these sensors are accelerometers and gyroscopes, which measure acceleration and rotation, such as that generated by a person's gait, or walking pattern. A smartphone, from its usual position in your pocket, is well placed to capture this information.

A person's gait contains a large amount of information. It is possible to identify people by their gait, such as with a fingerprint [12]. Furthermore, various personal characteristics manifest themselves in a person's gait, such as his or her height, weight, and sex [34]. Gait speed has even been considered a fifth vital sign, and is a general indicator of healthiness or lack thereof [26]. Researchers at the Albert Einstein College of Medicine have explored the connection between specific abnormal gait patterns and neurological diseases such as non-Alzheimer's dementia. They theorize that the analysis of gait patterns could determine whether patients are at risk for developing dementia [27]. Due to this great wealth of information, monitoring gait could have extremely beneficial results for public health.

The application of the ubiquity, power, and capabilities of the common smartphone to gait monitoring would be a tremendous advance. Current gait analysis involves the use of expensive and immobile equipment that is often restricted to lab use [31]. If smartphone sensor data could be used to record and analyze the motion caused by walking, it would have numerous benefits to the current practice of gait monitoring and research. The sensors in the modern smartphone are far more sophisticated than those in traditional commercial gait monitoring devices such as pedometers. Smartphones could be used as supplements to or even replacements for expensive medical hardware. They would permit monitoring to occur outside of lab environments, in a more natural setting.

Such monitoring could be used to validate lab tests or increase opportunities for data collection. Finally, smartphones would allow for real-time monitoring of at-risk patients, so that doctors could be aware of any dangerous changes in their patients' gait. Such technology could possibly even save lives.

Gait Monitoring

Gait monitoring is the practice of recording or observing the periodic body motions caused by a person walking. This specific pattern of repeating movements is that person's gait. There are many reasons to record gait because many characteristics manifest themselves in this specific walking pattern. First, there is reason to suspect that individuals have a unique gait, as it is possible to accurately identify people by their gait patterns [34]. The WISDM project at Fordham University has demonstrated the viability of detecting an individual's sex, height, weight, and actions from a gait pattern [33], and researchers at the Albert Einstein College of Medicine theorize that abnormalities in gait may be predictors of future neurological illnesses [27, 30].

Dr. Verghese and Dr. Holtzer at Albert Einstein have conducted a number of studies to associate gait with various neurological conditions, including the Einstein Aging Study [7]. Patients diagnosed with dementia, for example, tend to have a history of prior gait problems [30]. Their assessment is based on a number of quantifiable variables such as velocity, stride length, cadence (steps/min), swing time, stride length variability, and swing time variability [26]. Deviations in these variables are indicative of abnormal gait. Their studies have found that patients diagnosed with dementia, neuropathy, and other neurological diseases commonly manifest such abnormalities [7, 29].

Researchers at Albert Einstein use a set of specialized equipment to quantitatively record a patient's gait. The GAITRite sensor mat is a 20-foot long walkway that detects footfalls. It can measure the distribution of pressure about the area of a patient's footprint, their speed as they walk along the mat, the distance between their footfalls, and other advanced metrics. The SwayStar sensor device is a small pack that attaches to a patient's

torso. It contains sensitive sensors that detect rotational acceleration caused by any swaying motion of the patient [7, 27, 28, 29]. The combination of these two instruments allows the researchers to measure a large number of variables that describe a patient's movement.

The use of these sophisticated devices permits for an extremely accurate assessment of a patient's gait. However, their use also has a major drawback – testing is restricted to a lab environment. The GAITRite mat, for example, is 20 feet long and contains delicate electronics. As such, it is difficult to move and set up, and could be damaged outside of the lab. Patients may act unnaturally in a lab environment, which could potentially alter results. Additionally, it is difficult to conduct testing on patients who would have difficulty traveling to the lab, and any patients must schedule and arrange an appointment in order to be assessed. It is simply impractical to conduct remote monitoring or non-lab testing.

Wireless Sensor Data Mining

The Wireless Sensor Data Mining (WISDM) project at Fordham University uses data mining and analysis techniques to deduce information about people based on smartphone sensor records of their movements [11, 12, 34]. Data mining is a technique for extracting trends and patterns from data in order to associate or classify data points. The movement recorded by smartphone sensors during walking forms a record of a person's gait, and this information contains patterns that can be used to detect personal attributes and actions. The WISDM group organizes data collection efforts to collect sensor data and attribute survey information from volunteers, which are then used for experimental purposes [34]. The project has done extensive work in detecting physical characteristics such as height, weight, and sex. WISDM has also pursued activity recognition, to detect specific actions such as walking or jogging, culminating in the Actitracker activity-detection application available for Android.

The WISDM project has had success in detecting a person's actions based on smartphone sensor data, specifically of accelerometers and gyroscopes. Accelerometers and gyroscopes are able to detect acceleration and rotation applied to the phone in three dimensions, respectively. From a person's pocket, these smartphone sensors are able to gather a recording of his or her movement pattern. Volunteers with smartphones in their pockets performed a variety of activities during recording sessions, and data mining of the sensor data recorded during these sessions yielded a set of patterns that indicate whether that person was walking, sitting, standing, jogging, using stairs, or laying down at any given point in time [11]. The Actitracker Android application, published by the WISDM project, is a deployment of this activity recognition concept.

Actitracker (https://www.actitracker.com) is an application developed by the WISDM project for the Android smartphone platform. Android is a free, open source operating system maintained by Google [6]. It is currently used by the majority of smartphones in use today, including many powerful and sophisticated devices. Actitracker collects accelerometer and gyroscope data from the phone, and securely transmits that data to servers operated by the WISDM lab. There that information is analyzed using data mining techniques and the results are stored in a user profile. The user can log into an online account to view charts and graphs that display a record of his or her activities throughout the day [16, 33]. Actitracker is an example of smartphone sensor data being used for real-time activity recognition, and demonstrates a successful implementation of sensor data mining.

The WISDM project has also used smartphone sensor data mining to detect biometric traits [34]. Biometric traits are distinct human qualities. Descriptive qualities are known as soft biometrics, while identifying characteristics such as fingerprints or identity itself are known as hard biometrics. For this study, volunteers filled out a survey requesting soft biometric traits such as height, weight, shoe size, ethnicity, and sex, and then walked around a track with a smartphone in their pockets. The smartphone again collected a record of accelerometer and gyroscope data. Analysis of these records led to

the detection of patterns that can be used to determine several of these characteristics – most notably, height, weight, and sex [34]. It was also possible to distinguish each volunteer from any other with a startling degree of accuracy, indicating that gait might be used for hard biometric purposes [12].

The WISDM Project has shown that it is possible to detect actions and qualities using smartphone sensor data alone, so it may also be possible to differentiate between normal and abnormal gait using these same sensors. Researchers at the Albert Einstein College of Medicine suspect that gait might be an indicator of neurological diseases [7, 27]. In this case, a model that can differentiate between normal and abnormal gait could be used as a tool for diagnosis of neurological diseases. An application similar to Actitracker that could detect gait abnormalities would be deployable to smartphones, granting it a number of benefits. Gait monitoring would no longer be restricted to a lab environment, so researchers would be able to perform testing anywhere, or even remotely. Any results obtained in a lab could be confirmed in a natural environment, reinforcing their validity. Additionally, smartphones are commonplace, so anyone with access to such a device would have instant access to the application, reducing the need for expensive equipment. Finally, doctors could even monitor their patients' health in real-time, allowing them to detect dangerous developments as they happen. Such an application would have countless uses to any gait monitoring effort, and could even save lives.

The Model Induction Process

The goal of this thesis is to demonstrate a proof of concept for a model that can detect gait abnormalities using commercially available smartphone sensors. To accomplish this task, a set of experimental models was constructed from test data obtained in collaboration with the Albert Einstein College of Medicine. The process of building an experimental model occured in three stages. First, experimental data was collected from patients at Albert Einstein. Secondly, that data was subjected to analysis

and data mining techniques in order to extract patterns that are indicative of abnormal gait. These patterns form the basis of a model. Finally, the generated models were assessed and analyzed. Their performance is indicative of the potential for a model that can detect abnormal gait, and the patterns found in the test data provide insight into the aspects of human movement that characterize abnormal gait.

Data collection was performed in collaboration with researchers at Albert Einstein, who agreed to perform IRB-approved sensor data collection alongside their current data collection procedure. First, it was necessary to develop a procedure that was compatible with the preexisting Einstein procedure, and to supply the necessary tools. These tools included a smartphone sensor-collecting application and an appropriate device. The application was a simplified version of the Actitracker application, for the sake of continuity with other WISDM work. As such, the application required an Android smartphone equipped with accelerometers and gyroscopes. In addition to sensor data, the Einstein researchers agreed to provide anonymous gait classification information regarding the presence or absence of any abnormality and any diagnoses of neurological disorders, which any models attempt to predict.

Before the raw sensor data could be analyzed, it was cleaned and processed. First, the data was cleaned of any extraneous data that might interfere with analysis of the patient's gait before feature generation. Feature generation is a process designed to extract any salient characteristics from the sensor data. These characteristics, or features, summarize the sensor data for use by data mining algorithms, and must be carefully crafted and chosen to appropriately describe the data. These algorithms discovered any feature patterns that are indicative of presence or absence of gait abnormality as defined by the researchers at Albert Einstein. These patterns form models that can be used to detect gait abnormalities in any new sensor data.

Finally, these models must be assessed. All feature sets and data mining algorithms were assessed using leave-one-out cross-validation. This means that the data of every patient but one comprised a training set, which was used to build the model. The

remaining patient formed the test set, of which the model attempted to detect the presence or absence of gait abnormality. If the model's classifications matched the diagnoses received from Albert Einstein, then they were considered correct. This process was performed for each patient, and the percentage of correct classifications is an informative measurement of the model's overall accuracy. The model's accuracy is an indicator of the viability of detecting gait abnormalities or neurological diseases themselves using smartphone sensor data exclusively. Furthermore, the performance of the model on different patients provides insight into the reasons for accuracy or lack thereof. Success at different prediction tasks offers insight into the interaction of gait and neurological illness.

The collection and analysis of sensor data performed in this thesis culminates in model that could influence future work in gait monitoring. The viability of smartphone sensor gait abnormality detection is reinforced by the performance of these experimental models, and the results of this thesis invite the opportunity for future work and the development of an application that can be used for gait detection in pursuit of new capabilities in gait monitoring.

Overview of the Thesis

Each chapter of this thesis discusses one step in the model induction process. First, sensor data was collected to form a foundational dataset. The tools and procedures involved in data collection are described in Chapter 2. Once the data was collected, it was processed and analyzed to build a model. The analysis techniques used to interpret the data and construct experimental models are detailed in Chapter 3. The performance of these models is discussed in Chapter 4, along with the implications of their results, including potential expansions upon this work.

Chapter 2: Data

Any predictive model requires a foundation of sample data to provide examples of each category to be predicted. In this case, in order to predict gait abnormalities and neurological diseases using smartphone sensor data, it was necessary to collect data from individuals both with and without gait abnormalities and neurological diseases. This collection was performed in collaboration with the Albert Einstein College of Medicine, where researchers in gait analysis perform studies that attempt to establish an association between gait abnormalities and neurological diseases in elderly patients [7, 27, 30]. Dr. Verghese and Dr. Holtzer agreed to offer participants in their study the opportunity to supply sensor data, gait classifications, and disease diagnoses anonymously over the course of their study. Towards these ends, it was necessary to design tools and procedures that would interfere minimally with the Einstein procedure, but still ensure that sensor data collection would be rigorous and controlled.

To implement a data collection process, it was first necessary to determine which tools were needed, and then how to incorporate those tools into a testing procedure. In this case, it was imperative that any procedure interfere minimally with the Albert Einstein testing procedure, which added to the design complexity. Once the toolset had been determined – in this case, a smartphone and sensor collection application – it was designed and implemented. Then, given the Einstein procedure, the sensor data collection procedure was carefully constructed to meet the criteria mentioned before. These tools and procedures were designed to collect sensor data, gait classifications, and disease diagnoses for each participating patient, designated by a Patient ID. This information forms the dataset that was used to construct experimental models.

Data Collection Tools

First and foremost, a tool was needed to collect smartphone sensor data. A custom application that can collect sensor data according to experimental specifications was developed to satisfy this need. The WISDM Actitracker Android application contains the necessary functionality, but it also has additional functionality that is superfluous to the purposes of this project, and could potentially complicate any collection efforts. In order to increase ease of use and accuracy, a simplified derivative application was developed to facilitate data collection. As a result, the data collection application, like Actitracker, makes use of the Android platform [6]. A smartphone with the appropriate sensors was needed to run the application. The Samsung Galaxy S3 was chosen to run the application because its processing power and large feature set help ensure accurate sensor readings.

The sensor collector application removes much of the functionality from the Actitracker application, but keeps its sensor sampling mechanism. Its user interface was rewritten for simplicity, and its sensor sampling mechanism was modified to increase accuracy at the expense of concerns such as battery life and usability [16]. The application simply collects sensor data and records it. In this case, the application samples the accelerometer and gyroscope sensors every 50 milliseconds and records the values. In other words, the application records whatever acceleration or rotational forces are applied to the phone at that particular instant. Then, the application can email the information to an email account dedicated to this project. This is the minimum amount of functionality necessary for collecting sensor data, so as to avoid unnecessarily complicating the application, and making it as easy to use as possible.

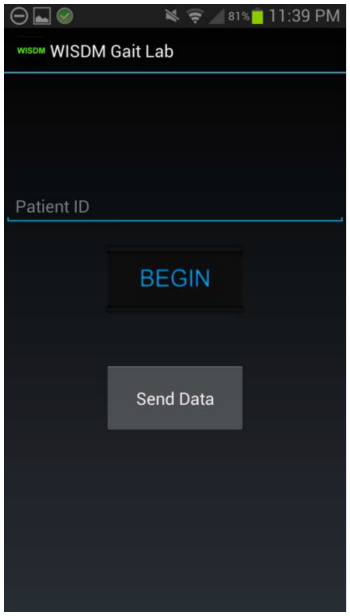

Figure 1. The data collection application has a simple and intuitive interface as shown here.

The application user interface is likewise very simple. Upon opening the application, the user is presented with the main user interface display (Figure 1). This screen displays two buttons and a text entry field. The text entry field is used to enter Patient ID for the current volunteer. The "Begin" button begins sensor recording and changes the screen to a "Now Recording" image. The "Stop" button on the "Now Recording" page stops sensor recording and reverts the screen to the original user interface. The "Send" button at the bottom of the screen emails the collected data to an email account dedicated to this project for analysis.

The smartphone chosen to run the application is the Samsung Galaxy S3. The S3 is a computationally powerful device with a large array of features, including sensors such as accelerometers and gyroscopes [24]. Its popularity qualified it as a representative Android smartphone, so applications designed for its feature set would be by extension applicable to the average Android device. Its commercial success increased the likelihood

that its features would become typical of future smartphones, and so application functionality is not likely to be impeded by hardware changes in or removal of functionality from future devices. Finally, the S3's computational power reduces the possibility of data collection errors. Maintaining a constant sampling rate on a commercial software platform is heavily dependent on CPU, and indeed, the WISDM project has experienced sampling problems on slower devices [16]. The Galaxy S3's powerful CPU greatly reduces the possibility of any degradations in the sampling rate. Two Galaxy S3 devices were supplied to researchers at Albert Einstein for the purposes of sensor data collection. These devices had the sensor application preinstalled, and had redundant Internet connections to ensure that it would always be possible to send sensor data.

Data Collection Procedure

The procedure used by the researchers at the Albert Einstein College of Medicine involved two pieces of specialized equipment. The GAITRite sensor mat is a 20-foot long track that records a patient's footfalls, including the distribution of pressure about a patient's footprint and the time and distance between successive footsteps. The SwayStar device is attached to the patient's torso, and records any swaying motion generated by his or her walking motion. The Einstein procedure was focused about recording the patient's gait using these two sensors [7, 27, 28, 20, 31].

First, the researcher administering the test asked the patient to stand at one end of the GAITRite track. If this test involved the SwayStar device, it was secured at the small of the patient's back. Then, the researcher started recording on the GAITRite device interface and the SwayStar device if applicable. At the researcher's cue, the patient walked to the opposite end of the track at his or her normal pace. Then, the researcher would ask the patient to perform any of a series of stationary tasks to test cognition and balance. Again, at the researcher's cue, the patient walked back to the beginning of the track at his or her normal pace. The researcher then stopped the GAITRite and SwayStar

recording. This marked the end the Einstein testing procedure.

The modified sensor data collecting procedure should alter the preexisting Albert Einstein procedure as little as possible in order to avoid interfering with their research. However, it is still necessary to ensure that the sensor data collection is thorough, controlled, and rigorous. The sensor data collection tools were designed to be as unintrusive as possible, so it was only necessary to add a small number of steps to the existing procedure. Five researchers at Albert Einstein were trained in the use of these tools and given a demonstration of their use. The researchers also conducted a practice collection in order to familiarize themselves with the procedure.

First, the Albert Einstein researcher administering the test asked the patient whether they wished to participate in this study. If he or she agreed, the patient was asked to wear loose-fitting pants with pockets to his or her appointment, and collection proceeded as follows. If the patient declined to participate, then the Einstein procedure proceeded without sensor data collection as described above. With the patient's agreement, the researcher asked the patient to stand at one end of the GAITRite track. Then, the researcher opened the sensor collector application on one of the supplied Galaxy S3 devices and entered the Patient's arbitrary ID number in the appropriate text field. The researcher then pressed the "Begin" button and handed the smartphone to the patient. The patient was then instructed to place the device in his or her front right pants pocket with the phone positioned upright and the screen facing away from the patient (Figure 2). Then, the Einstein procedure progressed normally. The researcher started recording on all Einstein sensor devices. Then the patient was asked to walk to the opposite end of the GAITRite track at a normal pace, stop, perform some stationary tasks, and walk back at a normal pace. Upon arriving back at the beginning of the GAITRite track, the researcher stopped recording of the Einstein devices as normal, then asked the patient to remove the Galaxy S3 from his or her pocket and hand it back to the researcher. The researcher then pressed the "Stop" button to cease collection, thus ending the collection session.

Figure 2: Placing the phone in the patient's pocket.

The researcher pressed the "Send Data" button on the application after the test was completed, as the sending process could take some time. This button sent all previously unsent data files as an email attachment to an address dedicated to this project using the phone's WiFi connection. Then, the researcher appended the Patient ID, a gait classification, a diagnosis, and other descriptive information to a spreadsheet for transmission.

The Dataset

The Albert Einstein researchers collected sensor data and diagnoses from 54 patients who elected to participate in this project over the course of 5 months. About 3 patients participated in sensor data collection each week. This rate of data collection was slower than anticipated, and the number of participating patients was very small for a data mining task. Specific demands of the Albert Einstein study, however, prohibited continued sensor data collection.

The sensor data is stored according to patient. Each patient contributed data to this study once, and was given an arbitrary unique identification number, or Patient ID. Each Patient ID has an associated set of sensor data, a gait classification, and a diagnosis. The sensor data describes the movement of the patient during the study procedure. The gait classification is an observational measure given by researchers at Einstein based on how much difficulty the patient has moving independently, and the diagnosis, also assigned by the Einstein researchers, indicates whether the patient has a neurological disease, a non-neurological disease, or both. Patterns in the sensor data as they relate to gait classification and diagnosis will form the basis of a predictive model, and the ability to interpret sensor data as spatial movement is pivotal to understanding these patterns.

The sensor data collection application takes readings from the accelerometer and gyroscope sensors every 50 milliseconds. These readings are arranged chronologically in data files. This chronological organization of sensor readings is called a time series. Each sensor data file sent from the sensor collection application contains about a full minute of sensor data. Of this, about 15 seconds of data describe the motion of the patient walking. The excess sensor data is removed in the feature generation step, however. The accelerometer and gyroscope each record in 3 dimensions (Figures 3 and 5), so a full sensor data sample consists of six time series, each of which describes some different aspect of the patient's walking pattern.

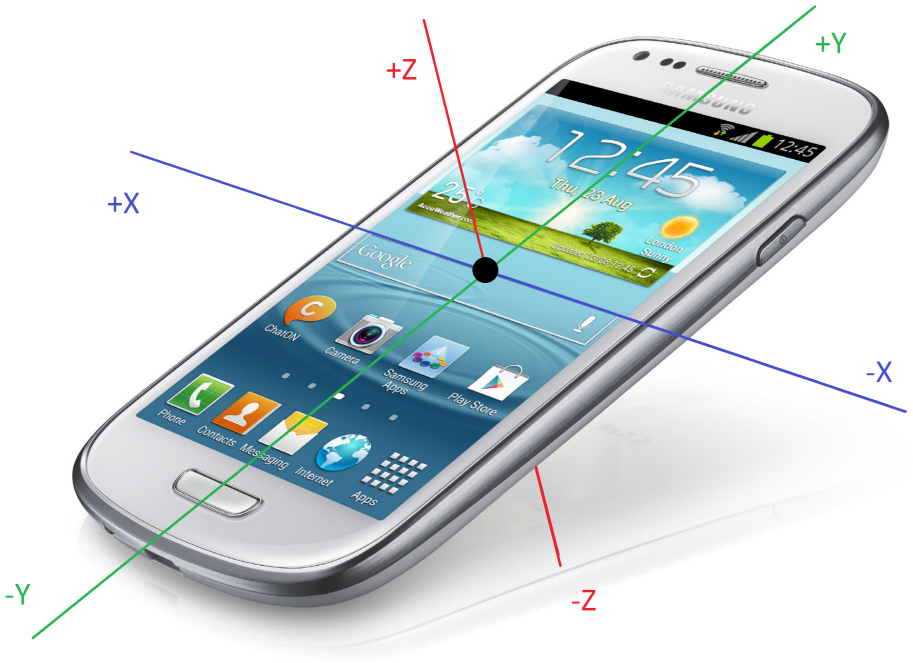

Figure 3: The Accelerometer detects acceleration along 3 axes.[1]

The accelerometer's Y-axis, which is oriented normal to the ground within the patient's pocket. This series typically oscillates about the 10 meter/second^2 value due the force of gravity on the phone. The vertical movement of the patient's leg is recorded by the Y-accelerometer. The accelerometer's X-axis records the smartphone's lateral movement. In the patient's pocket, the time series corresponds to the forward and backward swinging motion of the patient's leg. The side-to-side motion of the phone in the patient's pocket is captured by the Z-axis. This axis detects force normal to the smartphone's screen. Any acceleration towards or away from the patient's other leg is recorded by this axis. This axis is useful for detecting any swaying in the patient's gait.

1. Original image source: Samsung Galaxy S3 review - how does the Android phone that changed everything stack up now? http://cdn4.pcadvisor.co.uk/cmsdata/products/3355374/samsung-galaxy-s3-side-190.jpg

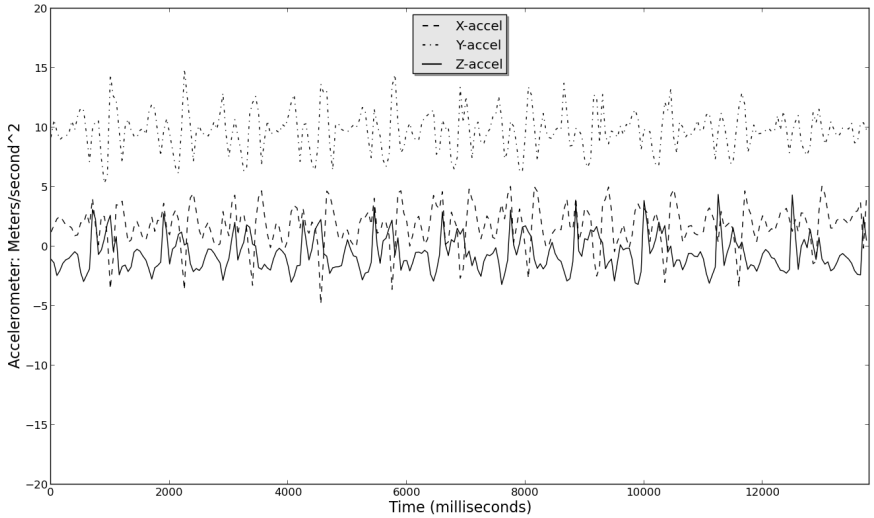

Figure 4: The combined accelerometer values reveal a clear cyclical walking pattern.

The composite accelerometer graph of a patient's walking motion reveals a clear cyclical pattern that matches the swinging of the patient's leg (See Figure 4). All three axes peak as the patient's leg accelerates forward and slightly inward for the next step. The slight second peak before the valley of each wave represents the movement of the patient's left leg as it swings forward. Finally, the valley of the wave represents the backswing of the patient's right leg as it pushes off the ground towards its next step.

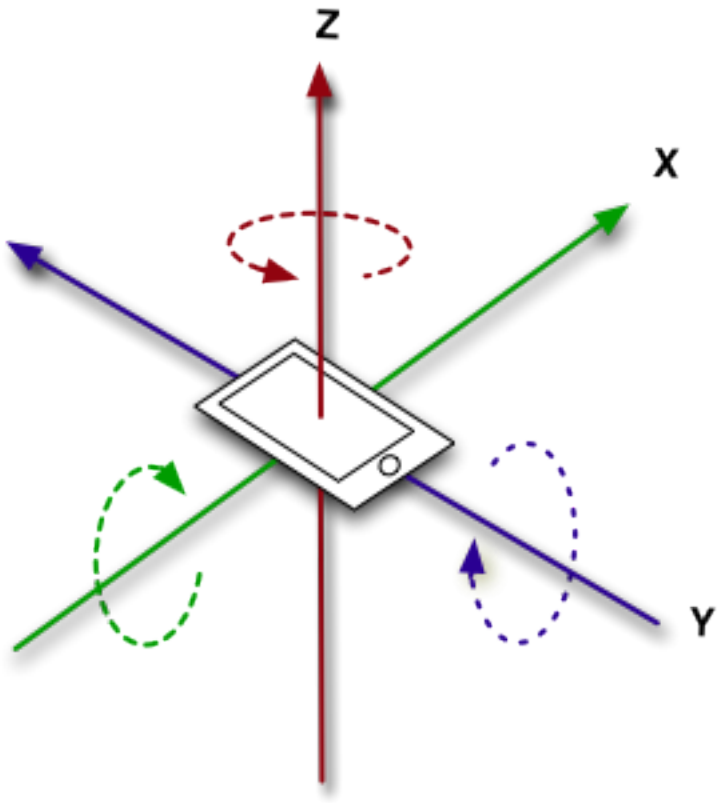

Figure 5: The Gyroscope detects rotation about 3 axes.[2]

The X-gyroscope detects rotation about the phone's horizontal midline. Any rotation exerted on the phone by the movement of the patient's leg towards or away from his or her other leg will be represented by this axis. The Y-gyroscope sensor data represents the rotation about the phone's vertical midline. The rotational force caused by the patient's turning foot will be captured by this axis's data series, and any rotation of the patient's body about the spinal axis will be captured by this axis. The Z-gyroscope axis detects rotational force about the axis of the phone that lies normal to the device's screen. This rotation is caused primarily by the swinging of the patient's leg about his or her pelvis. The Z-gyroscope captures this forward-and-backwards movement as a rotational force.

---

2. Image source: "CommandFusion iViewerScripting Documentation." http://www.commandfusion.com/docs/scripting/sensors.html

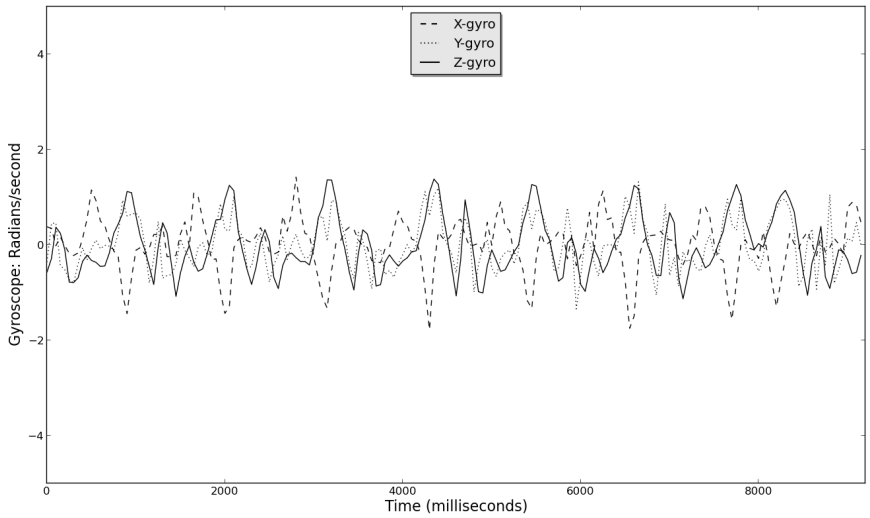

Figure 6: The combined gyroscope values reveal a clear cyclical walking pattern.

The composite Gyroscope image also reveals a clear cyclical pattern (Figure 6). The Z rotation peaks as the X and Y rotations fall as the phone rotates back during the backswing of the patient's leg. This is due to the counterclockwise rotation along the Z-axis caused by the swinging of the patient's leg. During the swing, the leg rotates towards the heel, causing the X rotation to peak. Additionally, the patient's leg moves out slightly as part of the same motion, causing a peak along the Y rotation axis.

The collected sensor data depicts a patient's movements while walking, including any gait abnormality. There are some data mining algorithms that can process time series, but these algorithms usually only measure the distance between two series. Distance is a relatively uninformative metric by itself. More detailed analysis requires decomposition of the time series into a number of descriptive values. Therefore, the data must be preprocessed and converted to a set of descriptive metrics called features before data mining model creation can begin. This phase of model creation is described in Chapter 3.

Patient Data

The Albert Einstein researchers provided attributes for the 54 patients who contributed sensor data. The researchers recorded a number of different characteristics (Table 1) for their own testing procedure. These data points are useful for identifying imbalances in the results that could eventually affect model results. Recorded items include the patient's age, sex, height, and weight. The researchers also included a pair of measurements – single-task and dual-task velocity – collected using their own equipment used for their own research. The purpose of this project is to demonstrate that prediction can be performed without the use of such equipment, so these values will not be used for model creation. Finally, the researchers labeled each patient according to how much difficulty that patient experienced in walking. Each patient was also given a diagnosis label that indicates whether the patient has a gait-affecting disease.

Age and sex are two characteristics that may indirectly affect gait [34], so it is beneficial to assess the distribution of these factors across the set of patients. The average age of the patients is about 76, and the oldest patient is 90, while the youngest patient is 66. The age of the patients is fairly evenly distributed, and there is not an imbalance in any age group. Females comprise 61.11% of the dataset, however, which is a not insubstantial imbalance. This may bias the classifier and make it more difficult to detect the attributes of males, if such differences manifest in walking motion.

Height and weight are two factors of the human physique that obviously affect an individual's walking pattern. A taller person will have longer strides, and a shorter person may take shorter steps. A heavier person might balance differently or walk more slowly than a lighter person. The average patient height is 164.08 cm, and the range of heights is 42.5 cm. It is difficult to determine the average weight because there may be unit conversion errors in the provided dataset. However, since these factors probably affect gait, in an ideal situation these variables would be controlled for. The study might be restricted to patients that fall within a specific height range in order to minimize any

affect the variation of height might have on gait. However, due to the small size of the patient pool, such selectivity is impractical.

The Einstein researchers record a patient's Single-Task Velocity and Dual-Task Velocity during their experimental procedure using the GAITRite track. The track records the movement of the patient's feet and can use this data to calculate velocity. Single-Task velocity refers to the patient's velocity while that patient is simply walking. A patient's Dual-Task velocity is recorded while the patient performs another action while walking, such as speaking [30, 31]. This change in coordination causes an evident change in the patient's velocity. These numbers are used by the Einstein researchers for their own research purposes. This project cannot use this information for prediction because it would defeat the purpose of demonstrating that sensor data alone might be able to detect gait abnormalities. Dual-task data collection was conducted separately from sensor data collection, so this change in behavior, which evidently affects gait, will not impact any analytical models.

Each patient was given a pair of labels by the Einstein researchers (Table 2). The first label classifies a patient's ability to walk independently. A higher number indicates that a patient has more difficulty walking. The second label indicates whether the patient has a non-neurological disease (1), a neurological illness (2), both (3), or neither (0). These two labels are descriptive of whether the patient has an abnormal gait. Therefore, the predictive model will attempt to classify the patients according to these last two values. In the following chapters, the sensor data belonging to each patient will be analyzed in order to build a model. This model will be used to predict the class values associated with that patient. A model with the ability to correctly classify these values could prove to be an invaluable diagnostic asset.

Patient ID	Age	Gender	Height (cm)	Weight (kg)	Single-task Velocity (cm/s)	Dual-task Velocity (cm/s)	Walking Difficulty	Disease Diagnosis
159	69	M	180	197	111.35	82.6	2	2
164	67	M	164	64.5	108.4	90.8		
171	75	F	157	72	109.35	36.7	1	0
182	77	M	173	73.1	104.6	69.3	2	1
183	75	M	175	78	120.6	115.5	1	0
185	72	F	150	50	124.15	105.4	1	0
186	79	M	177	204	83.05	81.6	2	
187	66	M	105	185	115	45.2	1	0
190	78	F	157	60	86.1	68.5	2	
191	70	F	162.5	79.7	83.45	56.8	1	0
192	68	F	157	55.9	137.8	104.8	2	2
196	69	F	166	130	117.05	103.3	1	0
216	75	F	160	155	117	83	2	2
218	74	F	168	67	129.7	104	1	0
220	67	F	148	114		108.2	1	0
221	89	F	150	67	100.2	73	1	0
226	81	F	169	62	64.6	35.3	1	0
228	81	F	164	54	97.1	59.1	1	0
229	73	M	180	80	131.2	103.7	1	0
241	71	M	182	85	119.5	47.3	1	0
243	90	F	148.5	48.6	97.3	57.4	2	1
245	84	M	174	93	78.6	64.3	1	0
246	82	M	176	79	122.3	84.3	1	0
248	81	F	166	77	93.1	79.3	1	3
262	78	F	156	63	102.7	64.3	1	0
269	72	M	180	96		84.8	1	0
275	76	F	157	82.5	92.4	52	2	3
282	88	M	160	67	103.35	56.4	2	2
286	81	F	154	90	101.05	57.8	2	2
348	73	F	158	49.5	69.7	81.7	1	2
349	73	M	172.7	94.6	84.05	60.2	2	2
352	86	F	158.5	58.6	72.2	66	1	0
354	76	M	172.8	83.9	74.15	57.7	2	3
355	78	F		54	69.45	21.9	2	
356	71	M	171	64.8	114	98.5	1	0
357	76	M	173	69.7	140.95	73.3	1	0
361	70	M	180.8	92.1	130.9	64.9	2	
362	71	F	152.4	73.1	113.4	89.8	1	0
364	71	F	166	106.6	89.8	66.2	1	0
366	88	F	155	65	98.05	75.2	1	0
367	72	M	177.8	89.8	127.55	73.1	1	0
371	80	M	170	67.8	119.55	79.4	1	0
373	77	F	172	87.8	128.7	48.9	1	0
374	75	F	169	59	111.1	93.6	1	0
375	77	F	162	60	116.5	78.2	1	0
376	74	F	150	84	78.6	47	2	
378	79	F	164	55		104.4	1	0
391	70	F	176	73		62.8	2	2
392	69	M	164	69	118.2	79.7	1	0
395	73	F	163	140	116.85	80.5	2	2
398	83	F	152	100	59.55	54.3	2	3
399	83	M	172	97	90.75	43.8	3	3
401	75	F	168	68	133.7	90.8	1	0
408	81	F	160	74		75.7	2	3

Table 1: Patient characteristics.

The collected data, referred to collectively as the dataset, forms the foundation of any predictive model. A significant effort was conducted to acquire data for the construction of this dataset. The tools and procedures developed for collection were designed to make data gathering as simple and fast as possible. The researchers at Albert Einstein conducted this study for 5 months in accessory to their own assessment procedure and collected a total of 7,180 sensor readings. The collected sensor data comprises a description of the movement patterns of 54 individuals who agreed to participate. The patterns in this data will enable analysis techniques to distinguish between patients with different labels as assigned by the Albert Einstein researchers. To encourage further research, this dataset will be made publicly available on the WISDM website (http://www.cis.fordham.edu/wisdm/dataset.php).

Walking Difficulty		
	1 Normal	33
	2 Mild Difficulty, walks without assistance or assistive device	19
	3 Moderate difficulty, uses cane or assistive device	1
	4 Severe Difficulty, requires assistance with taking steps	0
Disease Diagnosis		
	0 No Abnormalities	31
	1 Non-neurological disease	2
	2 Neurological illness	9
	3 Both neurological illness and non-neurological disease	6

Table 2: Walking Data and Disease Diagnosis Classifications.

Chapter 3: Processing and Analysis

A set of analysis and data mining techniques were applied to the dataset in order to construct predictive models. This analysis occurred in a 3-step process. During the experimental procedure used for data collection, the application recorded sensor data even when the patient was not walking. Therefore, the sensor data that represents the walking movement was first separated from the rest of the sensor data and prepared for analysis in a preprocessing step. Second, the cleaned and separated data was converted into set of features. Features are mathematical and statistical representations that summarize the data by extracting important characteristics [2, 18]. Third, these characteristics were analyzed by data mining algorithms that find patterns in the features that indicate the presence or absence of gait abnormalities. This set of patterns forms a model that can be used to predict whether a given patient has a gait abnormality.

Due to the nature of the data collection procedure used by the researchers at Albert Einstein, the patient was not walking for the majority of the time that the smartphone application was collecting sensor data. Gait assessment only targets motions generated by the patient's walking motion, so the relevant sensor information was extracted from the collected time series. This process was performed manually, so the extracted cycles of the patient's walking motion tended to be fragmented. A trimming procedure removed cycle fragments from the beginning and end of the selected series, ensuring that the gait pattern was captured in a clean sample. Then, the data was normalized about its mean in order to reduce the effect of gravity on the accelerometer and increase the comparability between different sensor axes. Finally, the data was divided into segments to emphasize discrete time series elements such as periodic cyclic motions and diminish the effect of outlier data. These procedures are all designed to remove unimportant data and isolate gait patterns in the sensor data so that these gait patterns can be described more accurately by features.

Feature generation is a process designed to extract salient characteristics from the raw data. Features are often mathematical descriptions of the data that quantify specific properties, such as statistical functions that provide information about a distribution. Data mining algorithms use features to differentiate between different categories of data, and some features are more useful towards this end than others. Many features provide a distinctive interpretation of how the raw data relates to the patient's movement. Others are less meaningful to humans, but may be useful to a data mining algorithm, especially in combination with other features. The careful development of features is essential to the construction of an accurate classifier [18].

Data mining is a technique for detecting patterns, trends, associations, and differences in data. Patterns present in features are used to construct a model that can differentiate between data samples belonging to different categories, such as patients with gait abnormalities, or patients without gait abnormalities. The data mining algorithms used for this project are implemented in the SciKit-Learn machine learning library for the Python programming language [19]. The specific application of these algorithms in conjunction with a well-designed feature set is key to constructing an effective model.

Preprocessing

The preprocessing stage of data analysis cleans and prepares the data for feature generation. Any non-walking sensor data was manually removed from the collected time series to isolate the patient's walking sensor data. The remaining data was then algorithmically processed for feature generation. First, the isolated walking sections were trimmed to remove any cycle fragments from the beginning or end. The data was then normalized to increase comparability. Finally, the data was segmented into the smallest meaningful time series to isolate any outliers. These techniques were designed to reduce the effects of unimportant factors in the time series while emphasizing the gait patterns themselves for comparison.

The raw transmitted data contains about a minute of accelerometer data (1200 records collected at 20Hz), of which, on average, 300 records, or 15 seconds represent a patient's walking movement. It was therefore necessary to partition the time series into intervals that appropriately represent the patient's gait. The smartphone's sensors are fairly sensitive, and the patient's pocket often permitted the phone to move independently. As a result, the data is very noisy, and identifying the exact edges of the walking pattern in the raw data series was difficult. A low-pass filter removed the noise and revealed the underlying walking pattern, indicated by the evenly-spaced double peaks in the time series. It was then relatively simple for a human to select the sections of the data that correspond to the patient's walking motion (Figures 7, 8, 9).

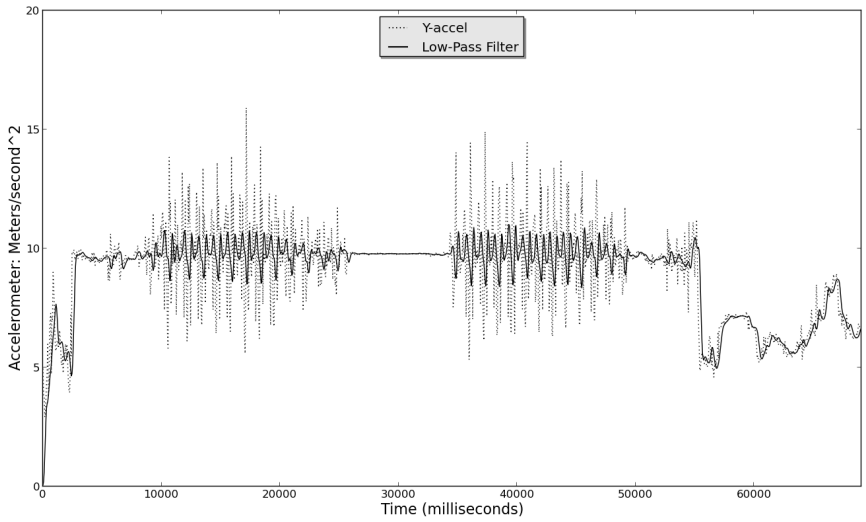

Figure 7: A low-pass filter reveals the walking patterns.

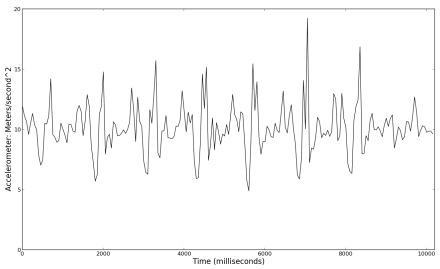

Figure 8: The first section of walking data.

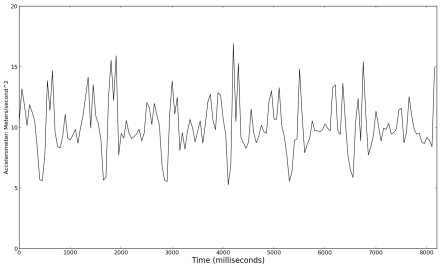

Figure 9: The second section of walking data.

The selected segments represent the patient's walking data, but the human element involved in selection left the cycles that comprise the patient's walking motion misaligned, which reduced comparability [37]. A relatively crude but effective method for aligning two waves was simply to remove data before and after the first and last full cycles in the series respectively. This process ensured that there were no fragmented cycles in the series. To accomplish this, a low-pass filter first eliminated any local maxima or minima that could have been caused by sensor noise. These data points would have interfered with the detection of the underlying shape. All samples before the first local maximum of the filtered wave or after the last local minimum were discarded. All series therefore begin at the peak of the first cycle and end at the valley of the last cycle. This approach to data trimming minimally impacted the length of the series, and reduced the amount of fragmentation present in discrete elements that appear in the series (such as cycles) [25]. In this way, the underlying shapes of the samples were brought into closer alignment (Figures 10 and 11).

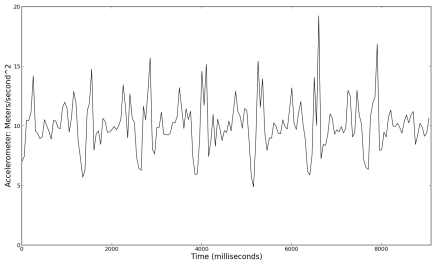

Figure 10: Fragmented data is removed from the first walking section.

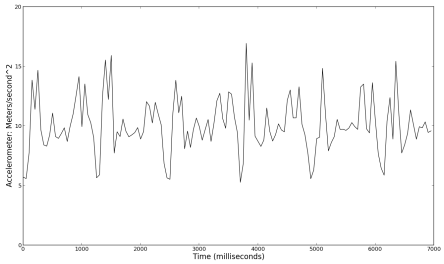

Figure 11: Fragmented data is removed from the second walking section.

Each axis was next normalized in order to increase comparability between the different axes. Normalization is a common practice used to bring all values in the appropriate time series into the same range for comparison [23]. In this context, normalization was useful for bringing Y-accelerometer data into the same range as the other axes. The Y-accelerometer values oscillate around a 10 m/s^2 value due to the force of gravity. Normalization forced the Y-accelerometer values into the same range as the

other axes while preserving the relative relationship between them. Normalized values were calculated using the following formula:

Let m and r be the mean and range for the time series S = {s_1, s_2, …, s_n}. The series S' is substituted for the original time series S, where each value of S s_i is replaced by the corresponding value of S' s_i' (Equation 1). The resulting series S' ranges between -1 and 1. Normalization reduces the impact of outlier data and minimizes the effect of gravity on the patient's movement data.

$$s_i' = \frac{s_i - m}{r} \qquad (1)$$

The data was then segmented, or divided into smaller time series. Time series segmentation has several benefits that increased the resolution of the calculated features. First, any outliers in the series only contaminated their particular segment. The entire series was not be distorted due to an outlier value. Second, reducing the length of a segment to the smallest meaningful length allowed for a more atomic analysis of particular behavior [10, 17]. Finally, the data mining classifier to which the feature vectors were finally submitted were able to produce a more meaningful model with a larger number of samples. Reducing the size of each segment to the smallest meaningful length simply gave the classifier more information regarding those particular atomic elements.

There are several common methods for segmentation. The sliding window method is among the most popular. This algorithm simply selects a number of consecutive samples of a given windows size, and then slides by an incremental parameter across the series [10, 23]. The incremental parameter is generally smaller than the window size. Therefore, some data points are analyzed more than once. However, this approach ensures that each point is analyzed from every possible context that might be provided by surrounding data points. The sliding window approach reduces the possibility of fragmentation of particular elements, such as cycles, of the time series. The drawback to

the sliding window approach is the dilution of any meaningful observations caused by analyzing every section of data repeatedly, even those that were uninformative. Each set of data points is analyzed in every possible context, but also in contexts that fragment the data and destroy valuable context. These bad segments dilute any collection of good segments and provide misleading data [9].

To retain the benefits of segmentation but preserve the context of each data point in the series, a segmentation process based on distance was applied. The GAITRite track at the Albert Einstein College of Medicine, Bronx NY, is 20 feet long. Therefore, dividing the data evenly into a set number of segments will fix the amount of distance covered in each segment. Each segment expresses the movement generated by the patient to travel a determined fraction of the 20-foot track. The context of each data point is affixed to an arbitrary section of the GAITRite track, ensuring that the sensor data of each segment logically corresponds to a specific physical motion. For this project, each walking section was divided into 8 segments so there were 16 segments per patient.

Several other segmentation strategies were tested, but each had its individual drawbacks. Intuitively, the data would be divided into individual step cycles. Algorithmically dividing the data into step cycles was inconsistent due to the noisiness of the data, however. Perhaps future research could refine techniques for isolating individual cycles. Larger numbers of evenly-sized segments excessively fragmented the data and smaller numbers of segments did not reduce the data to the smallest informative size. A segment size of 2.5 feet approximated the length of one step fairly consistently across all of the patient data.

The selected walking sections were trimmed, normalized, and segmented in order to reduce the variability among the time series, making their distinct characteristics more prominent for comparison. The features in the next section were better able to represent the differences between gait patterns once variability in the time series recordings was reduced.

Features

Features were chosen to describe the preprocessed time series data in a way that emphasized aspects that tend to characterize either the presence or absence of gait abnormalities. Features that are indicative of a gait abnormality, for example, will be recognized by a data mining algorithm and will contribute to the accuracy of the resulting model. Most features describe obvious physical characteristics of the data. However, some do not. The data mining algorithm may still find patterns in these features that contribute to a model's accuracy. The ordered set of feature values belonging to a given time series segment is called that segment's feature vector.

The Nearest Centroid data mining algorithm categorizes data points based on similarity to other data points. In this case, the algorithm calculates a centroid feature vector from a training set of vectors using a defined distance function. Each vector's assigned category is that of its nearest centroid, as determined using that same distance metric [19]. These calculated centroids were used to determine prototypical example segments of any class. The distances between the time series data of these prototypes and those of other segments are an informative metric that was used as a feature [8].

The mean or average value of a series is a standard statistical measurement that describes the entire series [18] (Equation 2). Each value of the series represents the amount of force applied to the phone at the time of the sample. The mean of these values serves as an integral of the series divided by the length of the sample. In other words, the mean represents the average net force applied to the phone per sample. A more positive mean is indicative of a tendency towards the positive axis of the sensor, while a more negative mean represents a tendency towards the negative axis of the sample. Before normalization, Y-accelerometer samples tend to have a mean close to 10 meters per second squared due to the effect of gravity (9.8 m/s^2). The mean value of all other sensor axes is close to zero as a result of the cyclical motion of walking patterns.

The mean is calculated by Equation 2, where n is the number of samples in a series S = $\{s_1, s_2, \ldots, s_n\}$:

$$Mean = \frac{\sum_{i=1}^{n} s_i}{n} \quad (2)$$

A local maximum is defined as a data point preceded and succeeded by lower values. The frequency and distance between local maxima in the series is an informative descriptor of the series as a whole [21]. The quantity of local maxima in a sample is meaningless in a noisy sample of sensor data. However, a low-pass filter makes the overall movement of a patient's leg discernible, and the frequency of local maxima in the sample is indicative of the rhythm of a patient's gait – particularly, any peaks caused by gait abnormalities can be detected (Figure 12).

Where p is the number of local maxima in a series and n is the number of samples in a series, the average number of local maxima is defined in Equation 3:

$$Average\ Maxima = \frac{p}{n} \quad (3)$$

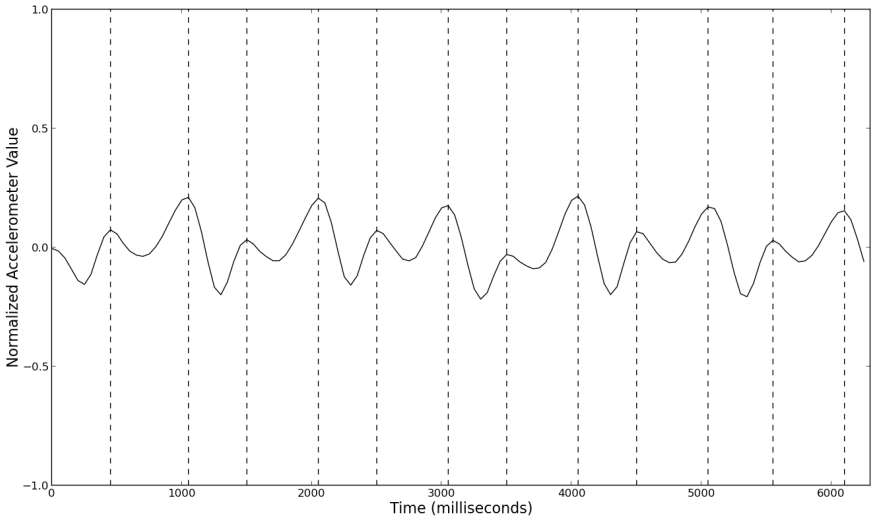

Figure 12: The local maxima occur regularly in a cyclical walking pattern.

The standard deviation represents the extent to which sensor sample values vary from the mean. It is a standard statistical measurement that characterizes the behavior of the series [18]. A movement pattern characterized by extreme application of force will have a larger standard deviation than one characterized by more gradual motions.

Where m is the mean value of the series $S = \{s_1, s_2, \ldots, s_n\}$ of size n (as calculated in Equation 2), the standard deviation is defined in Equation 4:

$$Standard\ Deviation = \sqrt{\frac{\sum_{i=1}^{n} \left(m - s_i\right)^2}{n}} \qquad (4)$$

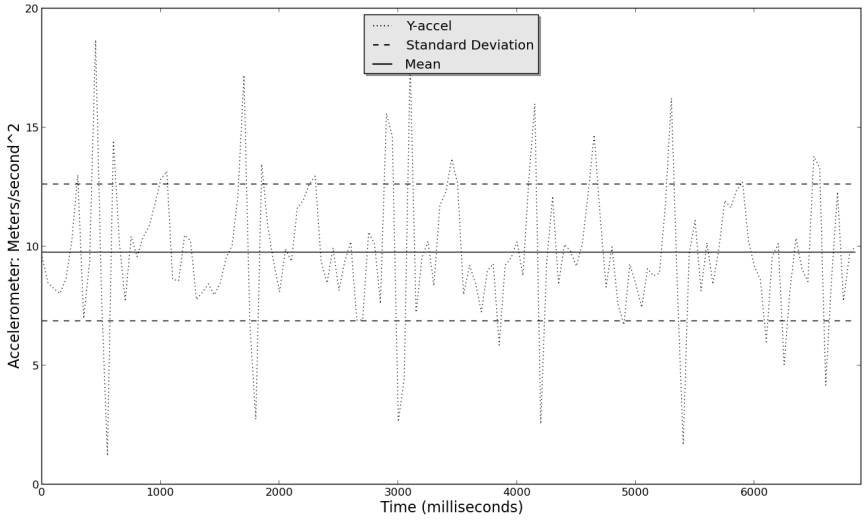

Figure 13: Accelerometer Values with mean and one standard deviation.

Skewness measures the extent to which sample values occur either below or above the mean. A skewed distribution depicts a movement pattern that tends to exert more force in one particular direction on the relevant axis. Skewness is a common statistical measurement used to characterize data distributions [18].

Where m is the mean (Equation 2) value of the series S = $\{s_1, s_2, \ldots, s_n\}$ of size n, skewness is defined in Equation 5:

$$Skewness = \frac{\sum_{i=1}^{n} \left(m - s_i\right)^3}{\left(\sum_{i=1}^{n} \left(m - s_i\right)^2\right)^{2/3}} \qquad (5)$$

The variance of a data series represents the extent to which samples tend to differ from each other (and the mean, by extension). It is the square of the standard deviation. A movement pattern characterized by extreme application of force will also have a large variance, unlike movement patterns with less abrupt movements.

Variance is defined in Equation 6, where d is the standard deviation (Equation 4) of the series S, variance is:

$$Variance = d^2 \quad (6)$$

The signal-to-noise ratio of a series quantifies the extent to which sensor samples tend towards the mean. A movement pattern with a low signal to noise ratio will tend to consist of more forceful motions. A pattern with a high signal to noise ratio will tend to consist of smaller, less forceful movements. Series with a large standard deviation will have a low signal to noise ratio.

Where m is the mean (Equation 2) and d is the standard deviation (Equation 4) of the series S as calculated above, the signal to noise ratio is defined in Equation 7:

$$Signal / Noise\ Ratio = \frac{m}{d} \quad (7)$$

The histogram of a time series measures the frequency at which values fall into specific ranges, or bins. Each bin represents a portion of the range of possible values. A 10-bin histogram simply counts frequency of occurrence of sensor samples in a series that fall within each 10^{th} of the range of possible sensor values. Each count is then divided by the length of the series to account for varying series lengths, yielding the fraction of the series that falls within each bin. The histogram characterizes the distribution of movements throughout the series, but fails to account for the arrangement of those values in time.

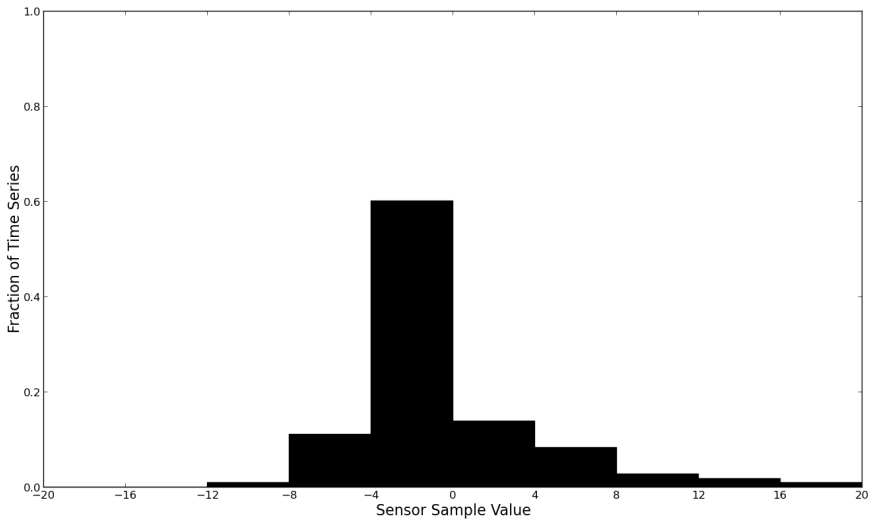

Figure 14: The sensor sample values fall into different histogram ranges.

Performing a Fourier Transform on the sensor sample series yields a frequency-domain representation of the smartphone's movement, represented by the set of Fourier coefficients. The largest coefficients alone comprise a powerful representation of the entire series [21, 23]. Therefore, determining the number of large coefficients provides a succinct and informative description of the series [21]. A histogram of the absolute values of the Fourier coefficients quantifies the distribution of frequencies throughout the series. Particularly, high-frequency movement is counted by the histogram, which is indicative of shakiness or unsteadiness in a patient's gait.

The Fast Fourier Transform algorithm used to transform the series $S = \{s_1, s_2, \ldots, s_n\}$ of length n to the frequency domain $Y = \{y_1, y_2, \ldots, y_n\}$ is defined in Equation 8:

$$y_j = \sum_{k=0}^{n-1} \left(x_k \times e^{\frac{-\sqrt{(-1)} \times j \times k \times 2\pi}{n}} \right) \qquad (8)$$

The preceding features form a preliminary feature vector, and the Nearest Centroid classifier was used to generate additional features that, when added, formed a final feature vector. The Nearest Centroid classifiers use a distance metric to calculate

feature set centroids for each class in the training set [19]. Classification of each test case is based on distance from that centroid. The Nearest Centroid classifier generates metadata that can prove valuable for classification. The calculated centroids can be used with the distance metric to provide additional descriptive metrics [7].

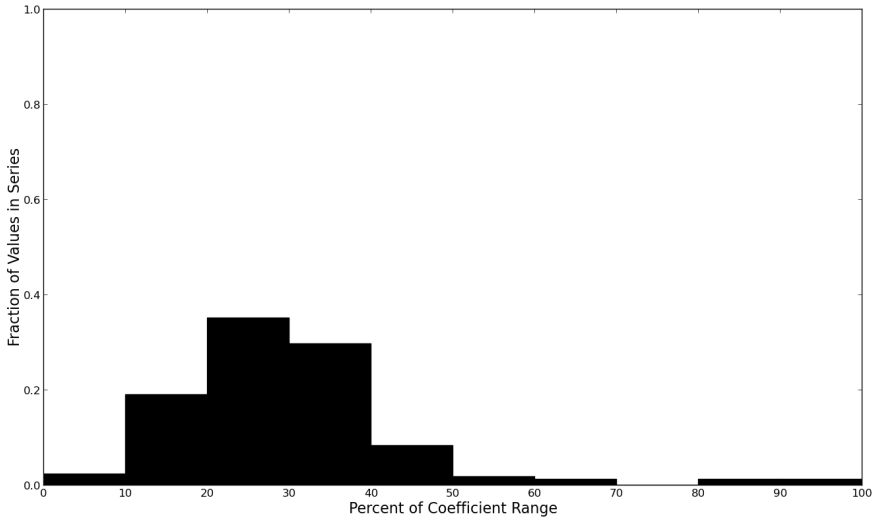

Figure 15: Fast Fourier Transform Histogram captures the distribution of frequencies in the time series.

The distance between feature vectors was calculated using a normalized variant of the Euclidean Distance. Many features operate on different scales – the standard deviation might be orders of magnitude larger than the value of a histogram value in a fast Fourier transform histogram. Therefore, dividing by the range of each feature is necessary in order to make the distance between a pair of large feature vectors meaningful. The implementation of the Nearest Centroid algorithm used can be found in the SciKit-Learn Python library, but this variant of the Euclidean Distance was coded for this project [19]:

Given two feature vectors of length n, $A = \{a_1, a_2, \ldots, a_n\}$ and $B = \{b_1, b_2, \ldots, b_n\}$ and the range of values for each feature across a given set of feature vectors as $R = \{r_1, r_2, \ldots, r_n\}$,

the normalized Euclidean Distance is defined in Equation 9:

$$Distance = \sqrt{\sum_{i=0}^{n} \left(\frac{a_i - b_i}{r_i} \right)^2} \qquad (9)$$

The Nearest Centroid classifier uses the Normalized Euclidean Distance to determine a feature vector that rests at the centroid of each set of feature vectors corresponding to a class value in the provided training set. For a two-class classification task, the classifier calculates a pair of centroids, one for each class. The pair of centroids ware compared against every other feature set to measure similarity between feature vectors of the same class. The resulting distance was appended to the feature vector.

The distances between the feature vectors of each sample and the centroid feature vectors were used by the Nearest Centroid classifier to generate predictions. Therefore, the elements closest to each centroid might be considered prototypical examples of each class. This approach has been used with similar clustering algorithms to great effect [7]. Consider a pair of samples from the training set. Prototype_A is the sample with the feature vector that is the least distance from Centroid_A, and Prototype B is the sample closest to Centroid_B. The sensor data time series of the Prototypes are prototypical examples of each class. Based on this assumption, metrics were applied that measure similarities between the time series themselves – namely, the Euclidean Distance and the Dynamic Time Warping distance. The time series that correspond to the preliminary feature vectors were compared to each other time series in both the training and the test sets using a pair of distance metrics. These two distances were appended to the corresponding preliminary feature vector to be used for classification. The time series were simply referenced for the purposes of calculating distance as another feature.

This approach has a potential caveat. If a given class value is actually represented by more than one characteristic centroid feature vector, then a prototype will only be partially applicable [7]. Regardless, any information gained from metrics applied to the prototype and other feature vectors may be informative to other classifiers.

The Euclidean Distance is the n-dimensional distance between two points. It is a fundamental measurement that has countless applications in time series comparison [13, 16, 19, 28]. With respect to a pair of time series with equal length, the Euclidean distance is the square root of the sum of the squares of the distances between each corresponding point in the pair of time series. The resulting figure is a measurement of the pairwise distance between the two waves.

Given two feature vectors of length n, A = {a_1, a_2, …, a_n} and B = {b_1, b_2, …, b_n}, the Euclidean Distance is defined in Equation 10:

$$Distance = \sqrt{\sum_{i=0}^{n} \left(a_i - b_i\right)^2} \qquad (10)$$

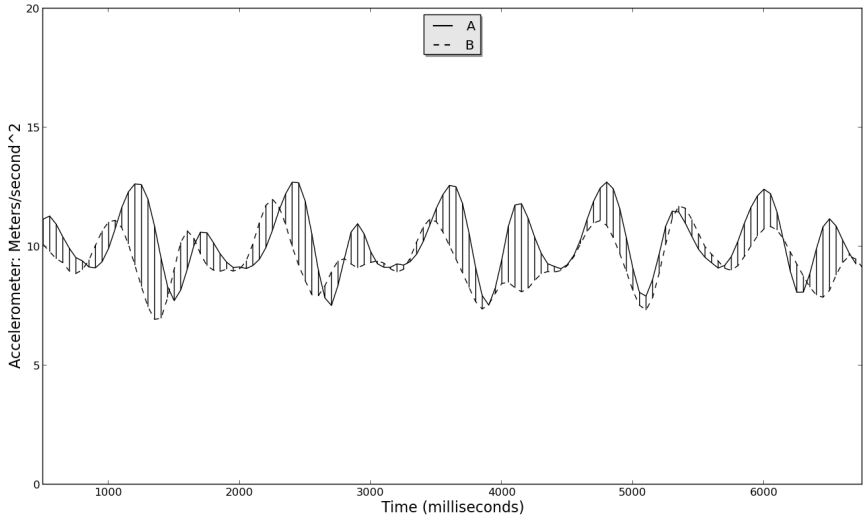

Figure 16: Euclidean Distance measures the pointwise distance between two time series.

Dynamic Time Warping (DTW) addresses the problem of time series elements that are out of phase. Each point in a wave is matched to the corresponding point in the wave to which it is compared. The difference between the points in each pairing produces

a measure of the difference between the two waves. Waves of different shape will have a higher DTW distance. DTW is a very powerful tool used in time series analysis. It is considered one of the best metrics for directly comparing two time series, and has applications in a large number of fields, including signal processing and genetics [1, 23].

A particular parameter that governs the DTW algorithm, is the window size [22]. The DTW algorithm will only match a given data point in the first series to a corresponding data point in the second series that is less than the specified window size from its location in the time series. Often, this window size is the length of the second series, effectively removing the window parameter. Otherwise, it is often an arbitrary 10% of the length of the second series [22]. Indeed, for the analysis of noisy time series with nearly aligned elements, it is often more informative to use a small window size [22, 36]. Therefore, the DTW algorithm was implemented with a window that is 10% of the length of the second series in order to ensure that the distance metric is as descriptive as possible.

The final feature vectors each contain 182 features. Each feature was designed to emphasize some aspect of a gait pattern as it manifests in the collected smartphone sensor data. Most of these features can be easily translated to a specific motion, while others are less transparent. Some of the features are significantly more informative than others. However, each in some way enabled data mining classifiers to differentiate between patients that have gait abnormalities and patients that lack gait abnormalities based on the collected sensor data in the dataset. A list of all features can be found in Table 3.

Feature Name	Values per Segment	Description
Mean	6	Average sensor values of axis.
Average Local Maxima	6	Number of maxima per sensor value.
Standard Deviation	6	Standard Deviation of axis sensor values.
Skewness	6	Skewness of axis sensor values.
Variance	6	Variance of axis sensor values.
Signal-To-Noise Ratio	6	Signal to noise ration of axis sensor values.
Histogram	60	Histogram of sensor values in each axis.
FFT Histogram	60	Histogram of the absolute values of the Fourier Coefficients each axis.
Distance From Centroid	2	Normalized Euclidean Feature vector distance from each centroid.
Euclidean Distance	12	Euclidean distance of each axis from each prototype.
Dynamic Time Warping	12	Dynamic Time Warping distance from each axis of each prototype.

Table 3: Features summarize the raw sensor data.

Data Mining Classifiers

Data mining algorithms find patterns in data in order to glean some information that is often not detectable by humans. Data mining techniques have countless applications, and more are introduced every day because the human race is generating more data today than ever before. There are a number of popular algorithms, but not all are suited to the purposes of this thesis. These algorithms are implemented in various incarnations across different software platforms and programming libraries. All algorithms used in this project are implemented in the SciKit-Learn library for the Python programming language [19]. While these algorithms are unaltered for this project, their performance is entirely dependent on the features and parameters with which they are supplied.

Data mining techniques have a number of analysis applications, but one of the most common applications for data mining is classification. Classification involves dividing data into predefined categories, or classes, based on feature patterns. For example, a patient might be classified as having or not having a gait abnormality based on patterns in the features of his or her sensor data. The patterns learned from training data form a model called a classifier that can assign a class to test data of a class unknown to the classifier. This class assignment is based on the test data's adherence to those patterns. Classifier algorithms find features that tend to differ across class lines to be more informative when analyzing training data, leading to a more accurate model. Therefore, features that emphasize distinctive characteristics of specific classes lead to the creation of better classifiers. The data mining algorithms that are traditionally very successful at performing such differentiation are decision tree classifiers [4].

Decision tree classifiers use a predefined function to determine how well a feature is able to differentiate between a set of classes. There are several popular decision tree implementations, but SciKit-Learn provides an enhanced version of the CART algorithm [19]. The CART algorithm uses a metric called the Gini impurity to determine the value

at which each feature is best able to differentiate between each class [3]. A differentiation based on this value is called a rule, and these rules are organized in a hierarchical tree based on how well they are able to distinguish between classes. The resulting set of rules identifies a set of patterns in the data, which are then used to assign a class to test data.

Decision trees have the advantage of being more interpretable than most other data mining classifiers [5]. The clear organization of decision trees allows humans to easily identify patterns from any graphical representations. Their hierarchical nature illustrates the interaction between different features, which can be translated to specific physical movements. Decision trees offer the opportunity to visualize the physical characteristics of gait abnormalities. The patterns found by decision trees pinpoint specific aspects of human gait and emphasize the relationship between those patterns and the predicted classes.

Problem Formulation

There are three prediction tasks that would prove useful in gait monitoring. First, the ability to detect whether a patient has difficulty walking is useful because such difficulty is closely linked with a number of diseases. Second, a model capable of detecting the presence of diseases themselves could prove to be a valuable diagnostic tool. Third, the two of these conditions are so closely linked, so it may be possible to predict the presence of either one or the other if detecting them individually proves difficult. The ability to successfully predict values in any of these three tasks would successfully demonstrate the utility of smartphone sensor data for gait monitoring. (The class value descriptions are in Table 2).

The dataset has a relatively small number of patients for such a complex classification task, so the two variables to be predicted were divided into binary classification problems [13]. A binary classification problem has two categories, the positive and negative. The negative class shows the lack of an abnormality, or when a patient is considered to have a normal gait. The positive class contains those patients who

have a gait abnormality. For labeling purposes, the negative class is denoted as ‘0’, and the positive class is ‘1’.

The negative class for Walking Difficulty represents those patients who have no difficulty walking. This class consists of those patients who were assigned a value of ‘1’ by the Einstein researchers. This class is 32 patients in size. The positive class consists of those patients who were assigned a value of {2, 3, 4}. This class consists of 18 patients.

The negative class for Diagnosis represents those patients who have no diseases or illnesses. These patients were assigned a value of ‘0’ by the Einstein researchers. This class consists of 30 patients. The positive class consists of those patients who were assigned a values of {1, 2, 3}. These patients have either a non-neurological disease, a neurological illness, or both. The size of this class is 16 patients.

Many of the patients belong to the positive class for Walking Difficulty and the positive class for Diagnosis. This is to be expected as the researchers at Einstein have established these two conditions to be closely linked [6, 27, 29]. Therefore, the patients who belong to only one of the positive classes may share characteristics with those who only belong to the other for underlying reasons. The positive class for the third classification task consists of those patients who have either a Walking Difficulty or a Disease or both. The size of this class is 20 patients. The remaining patients have both no Walking Difficulty and no Disease, and these patients compose the negative class. There are 30 patients in this class.

For each classification problem, each patient was assigned either a ‘0’ or a ‘1’ according to their labels. The data mining algorithm analyzed those patients’ feature vectors for differences between the two classes and generate a model based on those differences. The success or failure of the model at predicting the class value of features without class labels determines the accuracy of that model. These prediction tasks test the ability of the dataset and features to capture elements of gait that can distinguish between each of the binary classes. Any success is suggestive of the existence of patterns in the

sensor data. Success supports the possibility that sensor data might one day be used for large scale data monitoring. The next Chapter will detail the methods used for measuring model performance and discuss the implications of their results.

Chapter 4: Results and Discussion

The performance of data mining classifiers for a prediction task indicates how well that classifier is able to detect differences between the classes in the dataset and features. To demonstrate the feasibility of using sensor data for gait monitoring, this project assessed the performance of three classifiers. The first classifier attempted to detect whether a patient has any difficulty walking. The second classifier predicted whether a patient has a gait-affecting illness. The third classifier differentiated patients who have either difficulty walking or an illness from those who have no difficulty and no illnesses. These classifiers assigned class predictions to feature vectors. The predictions were scored using a majority voting strategy and compared to a straw man strategy in order to measure the classifier's accuracy.

The success or failure of each classifier has implications for the relationship between gait and disease. The patterns discovered by successful classifiers can even express the characteristic elements of motion for these conditions. Unfortunately, there were a number of factors that limited the ability of this project to create a fully useable model for gait monitoring. Even limited success, however, demonstrates that there is promise for the use of sensor data in gait monitoring.

Interpreting the Classifiers

The model performance for each classification task was tested using a strategy called leave-one-out cross-validation. The classifier was constructed from a set of feature vectors called the training set. It was then presented with a test set, and assigned a class label to each feature vector in the test set. The data mining algorithm knew the class values of the training set in order to build a classifier, but the class labels of the test set were withheld. In a leave-one-out cross-validation strategy, each patient is used as the test set once, and the rest of the patients are used as the training set. Therefore, a classifier was generated for each patient.

Each classifier's predictions were then compared to the actual class label of the patient in the test set. If the values match, the prediction is correct. If they are different, the prediction is incorrect. The percentage of correct predictions is the vector score for that patient. A classifier is generated to test each patient, so the average vector score across every patient is a preliminary measure of the algorithm's accuracy.

Once every patient had been assigned a vector score, a majority voting scheme was used to determine a discrete classification. Each patient is assigned a class label prediction equal to the class label that the classifier gave to the majority of that patient's feature vectors. Each patient had two walking data series per collection, and 8 segments per walking data series. Therefore, each patient had 16 feature vectors. If a patient had 9 or more correctly classified feature vectors, then the patient is considered to have been correctly labeled. If the patient has 8 or fewer correctly classified feature vectors, then the patient is considered to have been incorrectly labeled. The percentage of correctly labeled patients is the algorithm's voting accuracy.

The results of the majority voting scheme were then compared to a straw man strategy. A straw man is the simplest prediction strategy possible – guessing. If a classifier cannot do better than guessing, then its results are meaningless. For a classification problem, a straw man would simply guess the majority class, and it would be correct most of the time. The percentage of patients belonging to the majority class is the straw man accuracy. The algorithm's predictions are only meaningful if its voting accuracy is higher than the straw man accuracy.

Precision and recall are a pair of analytical metrics that assess how well a classifier produces meaningful test results. Precision is the fraction of the patients classified as positive that actually belong to the positive class. This metric measures how often a positive result is accurate. Recall is the fraction of the positive class that was correctly classified. It is a measure of how often the classifier correctly identifies positive classes. A separate straw man measure is used to assess precision and recall. This straw man will always predict the positive case, so its recall will be 100%, but it will have a

low precision. The F1 score combines precision and recall into a composite assessment of the classifier's ability to identify positive values [20]. The F1 score is calculated in Equation 11:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{11}$$

The performance of a classifier given a prediction task is represented by a confusion matrix. A confusion matrix shows the correct and incorrect guesses for each class – positive (1) and negative (0). For a binary classification problem, there are 4 possible categories for a given prediction. When the classifier is presented with a patient from the positive class, if the classifier predicts correctly, that prediction is a True Positive (TP). If the classifier predicts incorrectly, then it must have predicted the negative class. Since the patient belongs to the positive class, this prediction is a False Negative (FN). On the other hand, if the patient in question is from the negative class, a correct prediction is called a True Negative, and an incorrect prediction is a False Negative (FN). These values are displayed in a square matrix (Table 4). The values on the left are the actual classes, and the values along the top are the predicted classes. The sum of the values from top left to bottom right is the number of correct predictions, and the sum of all other values is the number of incorrect predictions.

		Predicted Class	
		1	0
Actual Class	1	TP	FN
	0	FP	TN

Table 4: A sample Confusion Matrix.

Results

The classifiers trained to detect whether patients had any difficulty walking (Table 5) achieved a voting score of 64%. The straw man score is also 64%, which means that this classifier is only doing as well as guessing. The precision is 50%, and the recall is 44%. Therefore, the F1 score is 47%. The positive class straw man used to assess precision and recall has a precision of 36% and an F1 score of 53%. A classifier with an F1 score lower than its straw man cannot be trusted to provide meaningful positive predictions.

		Predicted Class	
		1	0
Actual Class	1	8	10
	0	8	24

Table 5: Walking Difficulty Confusion Matrix.

		Predicted Class	
		1	0
Actual Class	1	4	12
	0	5	25

Table 6: Disease Diagnosis Confusion Matrix.

		Predicted Class	
		1	0
Actual Class	1	13	7
	0	8	22

Table 7: Either Walking Difficulty or Disease Confusion Matrix.

The decision trees generated to predict diseases (Table 6) were less successful. These classifiers have a 63.04% voting accuracy, while the straw man is 65.22%. Precision is 44%, recall is 25%, and the F1 score is 32%. The positive straw man has a precision of 43% and a F1 score of 60%. This classifier was extremely poor at predicting the positive class.

The classifiers were better able to distinguish between those patients who had either difficulty walking or a disease and those patients who had neither (Table 7). This classifier had a voting accuracy of 70%, and the straw man for this classification task is 60.00%. Precision is 62% and recall is 65%. The F1 score is 63%. The positive straw

man has a precision of 40% and an F1 of 57%. This classifier's predictions, although not very accurate, are better than guessing, indicating that the classifier was able to find a distinction between these two categories of patients. A summary of the classifier scores and their respective straw men can be found in Tables 8 and 9.

Classification Task	Majority Voting	Average Vector	Straw Man
Difficulty Walking	64.00%	59.37%	64.00%
Disease	63.04%	57.06%	65.22%
Difficulty Walking OR Disease	70.00%	60.75%	60.00%

Table 8: Summary of Classifier Accuracies.

Classification Task	Precision	Recall	F1	Straw Man F1
Difficulty Walking	50.00%	44.44%	47.05%	52.94%
Disease	44.44%	25.00%	32.00%	60.38%
Difficulty Walking OR Disease	61.90%	65.00%	63.41%	57.14%

Table 9: Summary of Precision, Recall, and F1 Scores.

While the decision tree algorithm did not achieve good performance in detecting the presence of a patient's difficulty walking or a gait-affecting disease, it was slightly more successful at detecting the presence of either condition. The failure of the first two tasks and the humble success of the third provide insight into the connection between motion and diseases.

Discussion

The results of all three classification tasks have implications for the relationship between walking motion and gait-affecting diseases. The poor performance of the first two classifiers serves to neither assert nor refute the potential for smartphone sensor use in gait monitoring due to a large number of limiting factors. For the third classification problem, data mining algorithms did not prove particularly successful at detecting either

walking difficulty or disease. These results are promising, however, because they indicate the presence of feature patterns that are able to distinguish between the two classes most of the time. They suggest that smartphone sensors are able to detect some movement characteristics that those patients who have difficulty walking share with those afflicted by diseases.

Very few patients who have either difficulty walking or a disease are not affected by both conditions. Albert Einstein studies have firmly established the link between these two afflictions [7, 24, 25, 26, 27]. The increase in accuracy when attempting to detect the presence of either condition indicates that those patients who have difficulty walking but do not have a disease share some common gait characteristic with those patients who have a disease but do not have difficulty walking. Furthermore, those patients that the models failed to accurately classify were not those who only belonged to one of the aforementioned categories. There seems to be some set of characteristics that the gaits of these patients share.

Analysis of a decision tree created by the data mining algorithm yielded a description of such characteristics. The decision trees differ from patient to patient since a different model was generated for each. Analysis of each tree yields insight into the composition of each training set. Very generally, feature vectors from the positive class tended to have less extreme data. These vectors tended to have lower values for features such as the higher Fast Fourier Transform Histogram bins and the Standard Deviation. This would indicate that individuals with gait abnormalities tend to have slower, more reserved motions, while patients with less difficulty walking have more brisk, forceful movements. This generalization is very broad, as the decision trees are varied and very large, with hundreds of nodes. However, the use of decision tree algorithms allows such interpretation for specific cases.

The decision tree algorithm performed poorly when attempting to differentiate patients who have difficulty walking from those who do not, and patients who were affected by diseases from those who were not. The classifier performed marginally better

at distinguishing patients who were affected by neither of these conditions from patients who had at least one of them. Even such weak performance is better than guessing, however, which suggests that there are distinctive patterns in sensor data that could be used to differentiate between these classes. The existence of such patterns supports claims made by Einstein researchers regarding the close association between difficulty walking and disease.

The underwhelming performance of the classification algorithms by no means detracts from the feasibility of using sensor data for gait monitoring. The dataset used for this project was extremely small and limited, which greatly reduced the expected performance of any data mining classifier. Not only were there very few patients, but each patient contributed only a small amount of walking sensor data. Additionally, walking difficulty classifications, while well defined, were still made visually. This could have contributed error to the first and third classifier sets. The expansion of the dataset would drastically increase the expected accuracy of all three classifier sets.

The dataset is extremely small at only 54 patients, some of which were missing individual class labels. Detecting trends as subtle and variable as gait disorder from a source as noisy and imprecise as smartphone data is far from impossible, but an extremely complex task. There are simply a large number of factors at play, and data mining classifiers need a large amount of data to sift through the noise and find the actual patterns that distinguish classes. A classification task of this complexity usually requires hundreds of examples of each class in order to generate meaningful predictions. A significant expansion of the volunteer pool would be expected to significantly improve results.

The positive classes were extremely under-represented in this study. Only 36% of patients who were given walking difficulty classifications had trouble walking, and 34.88% of the patients with diagnoses had diseases. These imbalances biased the classifier towards the majority negative class, skewing any predictions. These imbalances would explain the low recall for the first two classification tasks, but fail to account for

the low precision. A common method used to compensate for such imbalance is artificial balancing [35]. This involves discarding random patients from the negative class until the two classes are of equal size, and then training the classifier. However, this strategy reduced the dataset to a size too small to achieve meaningful results.

The patients that did participate in this project were predominantly female, and of well-distributed but greatly varying height. 61.11% of the patients in the dataset were female, which could account for some bias. The patients were of greatly varying heights as well. The tallest patient was 42.5cm taller than the shortest, and the heights of the other patients were distributed along this range. In an ideal study, patient height would be a controlled variable in order to reduce its effects, should gait abnormalities manifest differently for individuals of different height. However, the initial shortage of patients made such selection impossible. For such a great variability in height, a much greater number of patients would have been appropriate.

The patients who participated in this study were elderly individuals, many who had difficulty walking on their own. Therefore, it was only possible to collect a limited amount of sensor data in one session. On average, a patient contributed 15 seconds worth of walking sensor data. This is a very small amount of data considering the complexity of this prediction task. If patients had been able to even contribute a minute of sensor data, results could be expected to improve, as the data mining algorithms would have more samples from which to derive more descriptive patterns.

The assignment of Walking Difficulty classifications was performed visually during the Albert Einstein study. The goal of this project is to demonstrate that smartphone sensors are suitable supplements to existing gait monitoring equipment. Ideally, therefore, this study would attempt to predict values that equipment had quantitatively determined. The ability to come to the same conclusions as dedicated equipment would strongly support the use of smartphone sensors for the same quantitative tasks. Classes assigned manually by humans are naturally less precise, and this imprecision could have contributed to the underwhelming results of the classifiers.

The large number of limiting factors that impact the dataset used in this project shape the meaning of the classifier performance. The first two classification tasks, for which the data mining algorithms demonstrated poor performance, had extremely imbalanced classes. The third class had a more balanced class, and achieved at least marginal success. Only 10% of the patients who were classified with a disease did not have difficulty walking. These assessments could have been mistakenly recorded. The performance of the first two classifier sets certainly does not assert the viability of using sensor data for gait monitoring, but neither does it detract from this possibility due to the large number of mitigating factors that afflict these two sets of patients. The success of the third classifier set, on the other hand does assert the feasibility of sensor data use for gait monitoring because it was able to achieve even humble results despite a large number of limitations. It demonstrates that there are very likely detectable distinctions in the data. A larger, deeper dataset would allow for more refined detection of these distinctions, and a more successful classifier.

Future Work

The promise provided by the third classifier invites future studies to use the information in sensor data for gait monitoring purposes. There are significant challenges to this pursuit, especially in data collection. However, the resulting model could be very accurate. It could even see use as a diagnostic or remote monitoring tool. A smartphone application could deploy such a model by distributing this technology to any smartphone on the planet, making gait monitoring an accessible and easily implemented technique.

First however, the data requirements for a larger model would need to be met. A relatively strong model would probably require around 300 patients. Optimally, around half of these patients would be from each class to be predicted. It is somewhat difficult to find volunteers with neurological conditions, so this project would likely require collaboration with one or more research institutions that specialize in gait, which operate at their own pace. For this project, it took 5 months to gather data for 54 volunteers,

which is an excruciatingly slow rate. It would take years to build a reasonable dataset at this rate.

If possible, the study should control for height, weight, sex, and any other variables that could affect gait. In ideal conditions, the patients would have almost identical gait-affecting physical characteristics. This means that the patients would be approximately the same height and weight. This would yield a classifier better at identifying gait abnormalities in individuals that fit that same description. A classifier trained to identify gait abnormalities in individuals of any height or weight would require sensor data from a larger number of individuals of varying and evenly distributed height and weight. The initial scarcity of volunteers makes such selectivity difficult, but an expansive study might have access to enough patients to control for such characteristics.

Additionally, a future study should attempt to acquire a larger amount of walking data from patients. For this project, the Albert Einstein procedure only permitted for about 15 seconds worth of walking data per recording session. Ideally, the patient would contribute at least a minute of walking data in order to provide a more reliable recording of their gait pattern. The majority of individuals who have abnormal gait due to a disease or illness are elderly, and due to the focus of the experiment, have difficulty walking. These patients might have trouble walking for a full minute, limiting the amount of data they could contribute. Nevertheless, a future study should attempt to collect as much data from each patient as possible considering safety and medical concerns.

The inclusion of additional mobile sensor platforms could allow for more accurate gait abnormality detection. A person's gait manifests itself throughout his or her entire body, and a smartphone placed in a pocket may have difficulty detecting at least part of the motion involved in the larger gait pattern. The use of supplemental devices with sensors could increase results. For example, a set of increasingly popular devices called smartwatches contain accelerometers that could be used for gait abnormality detection (https://www.getpebble.com). These devices are worn on the wrist, and connect wirelessly to a smartphone. They can transmit any sampled accelerometer data back to

the phone for transmission or analysis. Such data could be used to supplement the smartphone's own sensors for increased accuracy. Future studies should consider making use of such devices.

An application developed around a gait abnormality detection model could offer a large amount of functionality. The application could sample the smartphone sensors and securely and anonymously send the resulting data to a server for analysis, much like the Actitracker application. A doctor approved by the patient could then see the patient's sensor data in real time and be alerted to any dangerous changes. Alternatively, the application could collect sensor data and analyze it on the phone itself to provide instantaneous gait classification. The major smartphone operating systems maintain 'App Stores' that allow any user to easily download and install the application [6]. The established infrastructure available to smartphones would make large scale deployment of any application extremely easy. The functionality and portability of the smartphone platform would enable gait monitoring to occur anywhere, and even remotely.

The creation of a model capable of better detecting gait abnormalities poses a substantial challenge. The data collection requirements for this endeavor are extremely daunting. The scarcity of volunteers coupled with a limited ability to collect data places severe restrictions on any data gathering effort. A thorough study would need to develop over the course of several years in order to acquire the data necessary for accurate predictions. However, the resulting model could see a number of uses in gait abnormality detection and remote gait monitoring due to the smartphone platform's versatility. The cost of a thorough future study would be high, but the benefits of success could be well worth the cost.

Chapter 5: Conclusion

This project represents a significant effort to demonstrate a proof of concept for smartphone sensor gait monitoring. This project initiated the development of sensor collecting software for the Android device, along with an experimental procedure for collecting sensor data from elderly patients. This thesis explored processing and feature generation strategies that will likely see future applications in wireless sensor data mining. The results of data mining efforts, while not immediately promising, do offer a substantial argument for the feasibility of future developments in sensor data gait monitoring.

Data collection was performed in collaboration with researchers at the Albert Einstein College of Medicine who were conducting a study to determine whether gait could be an indicator of neurological illness. Specially designed tools and procedures were developed to collect sensor data in accessory to the Einstein researchers' ongoing experiments. Sensor data was collected from 54 patients by two of the smartphone sensors, each of which senses motion in three dimensions. This motion depicts the patient's gait, and therefore characteristics that indicate the existence of any gait abnormalities. Patient attributes collected as part of the Einstein study provide context to the data, and the patient's diagnoses were used to generate a predictive model.

First, however, the data was preprocessed and summarized before analysis. The preprocessing step involved isolating data that described the patient's walking pattern and preparing it for future analysis. The data was normalized and segmented to emphasize only information that distinguishes each patient's gait. The data was then interpreted in 20 different ways as features, which together form a description of the important aspects of each section of walking data. Several features used a data mining classifier to associate other features before comparing the time series they represent directly. Then, the processed and summarized data was analyzed by a data mining algorithm to form classifiers. These classifiers attempted to predict whether each patient had a gait

abnormality.

The accuracy of these predictions indicates how well smartphone sensor data is able to capture important information useful for gait monitoring. Decision tree classifiers were formed for three different tasks, and the classifiers for each task were scored separately to determine which characteristics the sensor data could record. Unfortunately, the dataset is limited by practical considerations that likewise limited the accuracy of the classifiers. However, one of them scored relatively well and indicates that the sensor data contains at least some information useful for detecting gait abnormalities.

The Einstein collaboration yielded a modest amount of sensor data from elderly patients with varying ability to walk independently and sometimes gait-affecting diseases. The amount of data in the set, however, is simply too limited to allow for truly meaningful predictions. Future work attempting to identify gait abnormalities based on smartphone sensor data should invest a larger amount of time than was available to this project in the collection of sensor data. There are challenges involved in collecting sensor data from individuals with neurological illnesses who have difficulty walking, but the larger, deeper dataset would allow for refined classifications that could be used to create a meaningful classifier that might one day even be used as a diagnostic tool.

Although the results of the classifiers were poor, the fact remains that gait monitoring is a worthy pursuit for its numerous applications, especially in medicine. Smartphones are a powerful vehicle for gait monitoring, not only for their sensors, but also for their ubiquity and power. The somewhat successful results of one of the classifiers created for this project demonstrate that there is at least some useful information to be found in sensor information. The fact that such results were achieved even in the face of the many limitations reinforces this finding, and insinuates that future work that uses sensor data for gait monitoring could be successful. This thesis has established a methodology for using smartphone sensor data to perform gait monitoring that could be used to great effect on a richer dataset. The potential benefits from using smartphones for gait monitoring could be well worth future efforts.

Bibliography

[1] Aach, John, and George M. Church. "Aligning gene expression time series with time warping algorithms." Bioinformatics 17.6 (2001): 495-508.

[2] Anstey, Jonathan S., Dennis K. Peters, and Chris Dawson. "An improved feature extraction technique for high volume time series data." *Proceedings of the Fourth IASTED International Conference on Signal Processing, Pattern Recognition, and Applications*. ACTA Press, 2007.

[3] Breiman, Leo, ed. *Classification and regression trees*. CRC press, 1993.

[4] Chiu, Bill, Eamonn Keogh, and Stefano Lonardi. "Probabilistic discovery of time series motifs." Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.

[5] Geurts, Pierre. "Pattern extraction for time series classification." Principles of Data Mining and Knowledge Discovery. Springer Berlin Heidelberg, 2001. 115-127.

[6] Google. "Android Developers." http://developer.android.com/reference/. 2013.

[7] Holtzer, Roee, et al. "Cognitive processes related to gait velocity: results from the Einstein Aging Study." Neuropsychology 20.2 (2006): 215.

[8] Keogh, Eamonn J., and Michael J. Pazzani. "An Enhanced Representation of Time Series Which Allows Fast and Accurate Classification, Clustering and Relevance Feedback." KDD. Vol. 98. 1998.

[9] Keogh, Eamonn, and Jessica Lin. "Clustering of time-series subsequences is meaningless: implications for previous and future research." Knowledge and information systems 8.2 (2005): 154-177.

[10] Keogh, Eamonn, et al. "An online algorithm for segmenting time series." Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on. IEEE, 2001.

[11] Kwapisz, Jennifer R., Gary M. Weiss, and Samuel A. Moore. "Activity recognition using cell phone accelerometers." ACM SIGKDD Explorations Newsletter 12.2 (2011): 74-82.

[12] Kwapisz, Jennifer R., Gary M. Weiss, and Samuel A. Moore. "Cell phone-based biometric identification." Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on. IEEE, 2010.

[13] Levi, Kobi, Michael Fink, and Yair Weiss. "Learning from a small number of training examples by exploiting object categories." Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on. IEEE, 2004.

[14] Lin, Jessica, et al. "A symbolic representation of time series, with implications for streaming algorithms." Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery. ACM, 2003.

[15] Lipowsky, Constanze, et al. "Alignment of Noisy and Uniformly Scaled Time Series." Database and Expert Systems Applications. Springer Berlin Heidelberg, 2009.

[16] Lockhart, Jeffrey W., et al. "Design considerations for the WISDM smart phone-based sensor mining architecture." Proceedings of the Fifth International Workshop on Knowledge Discovery from Sensor Data. ACM, 2011.

[17] Lonardi, Jessica Lin Eamonn Keogh Stefano, and Pranav Patel. "Finding motifs in time series." Proc. of the 2nd Workshop on Temporal Data Mining. 2002.

[18] Nanopoulos, Alex, Rob Alcock, and Yannis Manolopoulos. "Feature-based classification of time-series data." Information processing and technology (2001): 49-61.

[19] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." *The Journal of Machine Learning Research* 12 (2011): 2825-2830.

[20] Powers, D. M. W. "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation." *Journal of Machine Learning Technologies* 2.1 (2011): 37-63.

[21] Rafiei, Davood, and Alberto Mendelzon. "Efficient retrieval of similar time sequences using DFT." arXiv preprint cs/9809033 (1998).

[22] Ratanamahatana, Chotirat Ann, and Eamonn Keogh. "Making time-series classification more accurate using learned constraints." Proceedings of SIAM international conference on data mining. Lake Buena Vista, Florida, 2004.

[23] Ratanamahatana, Chotirat Ann, et al. "Mining time series data." Data Mining and Knowledge Discovery Handbook. Springer US, 2010. 1049-1077.

[24] Samsung Electronics Co., LTD. "SPECIFICATIONS." http://www.samsung.com/global/galaxys3/specifications.html

[25] Shatkay, Hagit, and Stanley B. Zdonik. "Approximate queries and representations for large data sequences." Data Engineering, 1996. Proceedings of the Twelfth International Conference on. IEEE, 1996.

[26] Subra, J., et al. "Gait speed: a new vital sign for older persons in primary care." *J Frailty Aging* 1.2 (2012): 50-58.

[27] Verghese, Joe, et al. "Abnormality of gait as a predictor of non-Alzheimer's dementia." New England Journal of Medicine 347.22 (2002): 1761-1768.

[28] Verghese, Joe, et al. "Epidemiology of gait disorders in community-residing older adults." Journal of the American Geriatrics Society 54.2 (2006): 255-261.

[29] Verghese, Joe, et al. "Gait dysfunction in mild cognitive impairment syndromes." Journal of the American Geriatrics Society 56.7 (2008): 1244-1251.

[30] Verghese, Joe, et al. "Quantitative gait dysfunction and risk of cognitive decline and dementia." Journal of Neurology, Neurosurgery & Psychiatry 78.9 (2007): 929-935.

[31] Verghese, Joe, et al. "Quantitative gait markers and incident fall risk in older adults." The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 64.8 (2009): 896-901.

[32] Wang, Changzhou, and X. Sean Wang. "Supporting content-based searches on time series via approximation." Scientific and Statistical Database Management, 2000. Proceedings. 12th International Conference on. IEEE, 2000.

[33] Weiss, Gary M., and Jeffrey W. Lockhart. "A comparison of alternative client/server architectures for ubiquitous mobile sensor-based applications." Proceedings of the 2012 ACM Conference on Ubiquitous Computing. ACM, 2012.

[34] Weiss, Gary M., and Jeffrey W. Lockhart. "Identifying user traits by mining smart phone accelerometer data." Proceedings of the Fifth International Workshop on Knowledge Discovery from Sensor Data. ACM, 2011.

[35] Xi, Xiaopeng, et al. "Fast time series classification using numerosity reduction." Proceedings of the 23rd international conference on Machine learning. ACM, 2006.

[36] Yi, Byoung-Kee, H. V. Jagadish, and Christos Faloutsos. "Efficient retrieval of similar time sequences under time warping." Data Engineering, 1998. Proceedings., 14th International Conference on. IEEE, 1998.

Index of Figures

Index of Tables

Abstract

Shaun Gallagher

BS Computer Science, Fordham University

MS Computer Science, Fordham University

Smartphone Sensor Data Mining for Gait Abnormality Detection

Thesis directed by Gary Weiss, Ph.D.

Today, smartphones are a ubiquitous part of daily life. We carry them everywhere with us, and they are involved in almost every aspect of our lives. These omnipresent devices are equipped with sensors that allow them to gather information about the world around them. Among these sensors are accelerometers and gyroscopes, which measure acceleration and rotation, such as that generated by walking. A smartphone, from its usual position in your pocket, is perfectly placed to capture this information. The WISDM Lab has shown that this information can determine the qualities, identity, or actions of an individual, but such technology might be used to diagnose injuries and neurological disorders as well. In collaboration with the Albert Einstein College of Medicine, we built a model from smartphone sensor data that can detect gait abnormalities, which are often symptomatic of neurological illnesses such as non-Alzheimer's dementia. As part of this project, medical students from Albert Einstein collected data using smartphones as part of their normal clinical gait assessment. The smartphones run a custom-built application that collects accelerator and gyroscope data and transmits it back to the WISDM server. The data was cleaned, converted to representative features, and analyzed using data mining algorithms to build a model. The performance of this model indicates the viability of smartphone sensor data as a tool for detecting gait abnormalities. Further research upon and deployment of the techniques developed in this thesis could result in an application that could be used to detect gait abnormalities or neurological illnesses themselves. Such technology would prove to be a valuable tool for gait monitoring in medical and commercial settings.

Vita

VITA

Shaun Gallagher was born on April 16th, 1991 in New Jersey. After graduating from St. Joseph's Preparatory School in Philadelphia, Pennsylvania, he entered Fordham university as the recipient of a National Merit Finalist Scholarship. In 2013, he received the Bachelor of Science degree in Computer Science.

From May 2010 to May 2011 he worked for Fordham University User Support as a technician. He also joined the Wireless Sensor Data Mining Lab in May 2010, where he performed data mining and informatics research, as well as programming and server administration. He continued to pursue this research while working for his Masters degree in Computer Science under the mentorship of Dr. Gary Weiss, and jointly published a workshop paper to the KDD data mining conference. In July 2013, he began working as an information security consultant for Protiviti.