Australian National University

Machine Learning in Parkinson's Disease Diagnosis

Bachelor's Thesis

Max Wang

Supervised by:

Dr Deborah Apthorp
Research School of Psychology and Computer Science

Adj/Prof Hanna Suominen
Research School of Computer Science and Data61, CSIRO

Abstract

Parkinson's disease (PD) is a degenerative neurological disorder, affecting around one percent of the population by the age of 70. Currently, no objective test for PD exists, and studies suggest expert misdiagnosis rates of up to 34 per cent. Hence, there is interest in investigating if machine learning can provide a more reliable and objective diagnosis.

Current machine learning literature tests models and their ability to differentiate between already diagnosed PD and control subjects. This setup does not mirror real-life diagnosis as neurologists must exclude disorders with similar symptoms and diagnose individuals exhibiting minimal symptoms. Most studies are also based on small (N<50) datasets which suffer from a tendency to bias and overfitting due to Freedman's paradox.

Current literature on microphone and accelerometer based diagnosis was replicated on the 6,000 participant mPower dataset, which consists of crowdsourced recordings from smartphone sensors. Results reveal that the simple models used in current literature are insufficient to reliably perform diagnosis on the larger and noisier mPower data. Techniques in non-linear signal processing and deep learning based automatic feature engineering were consolidated to develop more powerful and robust models, showing major performance improvements compared to current state of the art models.

The experimental setup in current work is also problematic, as neurologist diagnosed PD subjects are compared with healthy controls. This setup is unavoidable due to the lack of publicly available datasets, but does not mirror real life diagnosis. This thesis in computer science explores innovative approaches to extrapolate the performance of machine learning with current datasets. We discover that machine learning can detect indicators of Parkinsonian speech imperceptible to humans, showing a promising future for machine learning in PD diagnosis.

Keywords: Parkinson's Disease, Signal Processing, Machine Learning, Classification, Human Activity Recognition, Dysphonia, Deep Learning, Neural Networks.

Acknowledgements

I would like to thank my supervisors, for their motivation and passion for the project; my family, for their continuous support over the years; and my friends, for the many late nights turned early mornings.

Declaration

This thesis is an account of research undertaken between February 2017 and October 2017 at the Research School of Computer Science in the Australian National University, Canberra, Australia.

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.

Max Wang
October, 2017

Contents

Introduction

This Honours thesis in Computer Science comprehensively explores the machine learning process from raw sensor data to results. It has been written for all audiences; however, a background in machine learning may be useful to understand and follow assumptions in the methodology. The work spans multiple disciplines and we have opted to concisely summarise these fields and provide references to seminal or well-written papers in the area for readers interested in an in-depth understanding. The mathematical or algorithmic formulations for most techniques are abstracted in favour of their intuition.

Throughout the thesis, highlights and footnotes are used to improve flow and reading. Highlights convey or re-iterate important information for those skim-reading and footnotes[1] provides contextual background information.

Highlight. Highlights re-iterate crucial information.

Chapter 1 consolidates all literature relevant to this thesis, starting with an outline of Parkinson's Disease and why it is difficult to diagnose. *Section* 1.1 examines the applicability of machine learning in the task, and the limitations of current datasets and techniques. *Section* 1.2 revisits the underlying biology of Parkinson's disease, and relates the biological changes to the signal processing techniques aimed at quantifying them. Finally, *Section* 1.3 provides an overview of the machine learning process including the modern advances in neural networks.

Chapter 2 begins with a literature review of relevant works in PD machine learning, covering their results and shortcomings. In *Section* 2.1, the dataset we use (mPower) is examined, and the data pre-processing described. *Section 2.2* replicates current literature on the much larger mPower dataset and reveals that current techniques fall short of their reported performance. This is likely due to the larger variety of participants in the mPower dataset compared to the small datasets used in prior work.

[1] *Footnotes* provide contextual background information

Section 2.3 investigates the applicability of the novel features we have introduced in *Section* 1.2.4. An improvement is observed; however, the task is clearly more difficult than suggested by prior literature. *Section* 2.4 explores the difficulties of the task by visualising the current and novel features employed in the task. *Section* 2.5 explores techniques to improve machine learning performance, including data augmentation, feature selection and ensemble models. *Section* 2.6 applies deep learning to the task of engineering features, inspired by the latest developments in speech recognition and computer vision. Finally, in *Section* 2.7 we consolidate our work and discuss what machine learning can offer to PD diagnosis. *Section* 2.8 describes the platforms used to implement the project. The code is published online at https://github.com/maxwg/parkinsons-mpower, and the accompanying presentation at https://maxwg.github.io/honours/presentation

1 Background

Parkinson's disease (PD) will affect around one percent of the population by age 70 [1]. It is a degenerative neurological disorder characterised by a regression of movement, speech, and memory. There is currently no objective test for PD and diagnosis is especially difficult in its early stages as symptoms (*Table* 1.1) have not fully manifested [2]. Studies suggest that motor symptoms only manifest once 20-40 per cent of dopamine[1] producing neurons have deteriorated [3]. The exact underlying causes of Parkinson's disease are still unknown [1].

Table 1.1: Symptoms of Parkinson's disease [1]. Although this disorder is commonly associated with tremor, only around 70 per cent of patients experience resting tremor [4].

Movement	Voice	Non-motor
Resting Tremor	Reduced Volume	Hallucinations
Rigidity	Monotonous Speech	Reduced Cognitive Ability
Bradykinesia (Slow Movement)	Imprecise Articulation	Sleep Disorders
Dyskinesia (Involuntary Movement)	Slurred Speech	Mood Disorders
Akinesia (Freezing of Gait)	Hesitant Speech	Vision Problems
		Physical Changes

Current treatments provide temporary relief from symptoms and have been shown to slow disease progression [5, 6, 7] . Thus, a correct early diagnosis is crucial to ensuring a higher quality of life later in life. PD is currently diagnosed with a standardised, yet subjective test by a neurologist [8]. This test involves qualifying visible symptoms such as tremor and assessing the patient's response to drugs such as Levodopa[2]. As visible symptoms do not manifest until later stages, an early stage diagnosis is rare.

[1] *Dopamine* is a neurotransmitter that aids communications between neurons — the basic working unit of the brain. As PD causes deterioration in dopamine producing neurons, this leads to a decline in functionality of the Basal Ganglia which is associated with fine motor and cognitive control.

[2] *Levodopa* is the most common medication for Parkinson's disease. It is converted to dopamine in the brain — replenishing the patient's deficit. It often results in side-effects such as depression and fatigue.

The primary difficulty in diagnosis is differentiating from other Parkinsonian[3] disorders such as Multiple System Atrophy, Supranuclear Palsy, and Essential Tremor [10, 11]. Confirmation of diagnosis is generally only possible with an autopsy. As there is no definitive test and symptoms resemble other neurological disorders, diagnosis is not simple. Studies suggest the misdiagnosis rate is high, ranging from 9–34 per cent depending on methodology [8, 2, 12].

Highlight* 1.1 *(Diagnosis). PD is diagnosed with a standardised, yet subjective test administered by a neurologist. The misdiagnosis rate is high — up to 34 per cent with the primary difficulty being differentiating from other Parkinsonian disorders.

As there is no consensus for PD diagnosis, the search for a more objective measure for diagnosis is a hot topic in the research community. This ranges from more standardised diagnosis criteria such as the UK Parkinson's Disease Society Brain Bank criteria [8, 13, 14], to discovering more quantifiable biomarkers such as gene expression [15, 16] and proteins in bodily fluids [17]. Although the discovery of objective biomarkers shows promise, it is likely that cost would be prohibitive for most early stage patients. *Machine learning* is another viable option, offering an objective and low-cost tool to assist the neurologist in diagnosis.

***Highlight* 1.2.** Machine learning may prove to be an objective and low-cost tool in the diagnosis of Parkinson's disease.

1.1 Machine Learning in Parkinson's Disease

Machine Learning can be broadly defined as a suite of computational techniques that address the challenge of making sense of the ever-increasing volume and complexity of data generated in, for example, modern information-dense healthcare systems. The technical foundation of these techniques will be examined in *Section* 1.3.

There has been a large body of work in the field — a majority with very positive results [18, 19]. However, the applicability of this research in healthcare is limited, primarily due to the small datasets associated with these publications. Most datasets used in literature consist of fewer than 40 subjects. Reported results are therefore prone to biases in the dataset, Freedman's paradox[4] [20] and overfitting on cross validation [21]. It is difficult to empirically compare results of different papers, and often, later work consolidating the methodologies of prior work on a new dataset achieves worse results [22].

[3] *Parkinsonian* syndromes share many symptoms with PD [9]. They are differentiated by the underlying cause, and treatment options vary.

[4] *Freedman's paradox* describes a common issue in model fitting where variables with no predictive power appear important. It is especially prevalent when the number of features exceeds the number of data points.

There is also a fundamental issue with the setup of experiments in current PD machine learning literature. Machine learning is used to differentiate subjects already diagnosed with Parkinson's (therefore, likely having noticeable symptoms) with healthy subjects. This artificial setup simplifies the complexities involved in a neurologist's diagnosis for PD, where they must exclude a number of other causes for the symptoms. There has been some preliminary investigation in the applicability of machine learning in differentiating PD and other Parkinsonian disorders [23, 24]; however, these suffer from similar issues with limited data.

***Highlight* 1.3.** Current research uses machine learning to differentiate PD and healthy individuals. This is a simpler problem than that faced by neurologists, who must rule out a number of other possibilities.

To precisely compare the effectiveness of machine learning to neurologist diagnosis, a large *longitudinal dataset* would be required. Collecting this would involve monitoring subjects prior to any Parkinsonian symptoms until their passing, where the existence of PD can be confirmed through autopsy. This data would allow the comparison of machine learning to the diagnosis by a neurologist given the same information. Such a dataset would be very costly and logistically difficult to collect. To advocate the collection of this dataset, evidence of machine learning's applicability to PD diagnosis is needed. This thesis will investigate methods of assessing machine learning's applicability to Parkinson's disease without such a dataset.

One such application of machine learning is Parkinson's disease telemonitoring [25, 26]. The progression of PD is monitored with a scale, the most common being the MDS-UPDRS [27] which quantifies the extent of 44 motor and non-motor symptoms on an integer scale between 0–4, with 0 representing no evidence of symptoms and 4 indicative of severe symptoms. It is recommended that individuals with PD visit a clinic every 3–6 months to track progression and adjust treatment options — this is costly and inconvenient. Machine learning offers the opportunity for patients to track disease progress at home with their smartphone or other wearables [28]. Monitoring is a practical avenue for machine learning given current datasets; however, it will not be explicitly explored in this thesis, as the primary focus is diagnosis. The machine learning techniques applied in monitoring and diagnosis are easily interchangeable.

Highlight* 1.4 *(UPDRS). The MDS-UPDRS [27] scale quantifies the extent of 44 motor and non-motor symptoms on a scale of 0–4. It can be administered by individuals qualified to use the scale.

This thesis will cover the full process of creating a machine learning model to diagnose

Parkinson's disease. *Section* 1.2 examines traditional feature engineering — the process of converting raw sensor data into numbers interpretable by a machine learning model. *Section 1.3* investigates machine learning techniques used to classify Parkinson's disease from the features as well as some recent 'deep learning' based approaches of extracting features from the raw sensor data. In *Chapter* 2, we develop increasingly powerful machine learning models, and pit machine learning against humans in *Section* 2.7.

1.2 Feature Engineering and Signal Processing

Feature engineering is the process of converting raw input data (*signals*) into meaningful numerical values[5] that can be used for machine learning. For example, with sensors such as microphones, features such as pitch and volume may be used. Features should be relevant to the machine learning task, as most models perform poorly when unrelated features are introduced. A common strategy is to engineer as many features as possible and apply feature selection to narrow down the set of features – especially when the data is not well understood. Feature selection is examined in *Section* 1.3.3.

Movement-related problems are the primary manifestation of symptoms considered by a neurologist when diagnosing PD. Human vision is very advanced, capturing and processing a great deal of information about the world around us. Through years of experience, we have learned the general behaviour of human movement, hence minor tremor and slight deviations from normal gait are very noticeable [29]. However, our ability to differentiate between forms of irregular gait is more limited [11]. Although sensors such as accelerometers only a capture a fraction of information compared to eyes, they may be more precise and better at distinguishing forms of irregular gait [30].

***Highlight* 1.5.** Our senses are good at detecting deviations from normal gait/speech, but are less proficient at detecting differences between types of abnormal gait/speech.

Although speech is only a single component of the 44 component UPDRS [27] scale, it has received a great deal of attention in machine learning. There is evidence that speech is one of the earliest indicators of PD [31] and there exists a large body of work in vocal feature engineering [32]. Furthermore, there is much less information loss when recording audio with a microphone compared to the sensors used to record movement.

Table 1.2 summarises prior work in feature engineering related to PD. As most datasets consist of data from a single sensor, machine learning focuses on quantifying a single symptom of Parkinson's disease based on that sensor. Literature can be classified as by the symptoms they are attempt to quantify [33, 34].

[5] Some models such as neural networks can work from raw sensor data (*Section* 1.3.2 and 2.6). Feature engineering is essential for most machine learning models.

Table 1.2: Our classification of prior work in the field of PD diagnosis. The signal processing of sensor data is often more important that the machine learning model.

Movement	Voice	Non-motor
Resting Tremor IMUs[6] [35, 36, 37] Smartphones [19, 38, 39]	Words and sentences [53, 54, 55]	Demographics
Postural Sway Force Plates [40, 41] IMUs [42, 37]	Sustained vowel phonation [25, 56, 57]	UPDRS Patient Questionnaire [58, 59]
Gait Force Walkways [43, 44, 45] Video [44] Multiple IMUs [46, 47, 48]		Physical Changes Gene Expression [15, 60] MRI [61, 62] EEG [63, 64] Olfactory [58] REM sleep [58, 59] Cerebrospinal Fluids [59] Gastrointestinal [65]
Handwriting [49, 50]		
Motion Capture [51]		
Tapping [52, 22]		

There is evidence that PD is heterogeneous and symptoms are present in distinct subsets [66]; however, the underlying reasons not well understood. Studies have reported speech dysfunction present in 74–94 per cent of individuals with PD [67, 68, 69, 70], tremor in 70 per cent [4] and bradykinesia in 15 per cent [71]. As neurologist diagnosis relies on judgement from observation, there is the possibility that some of these symptoms exhibit in a form imperceptible to a neurologist but detectable by a high-resolution sensor.

***Highlight* 1.6.** It is possible that some subtypes of PD exhibit symptoms imperceptible to a neurologist but detectable by a high-resolution sensor.

Unless there is evidence that '*micro-symptoms*' are present in all people with PD, feature engineering in each of these areas is equally important. *Section* 1.2.2 explores some of the biological causes of these symptoms, and we will investigate the existence of micro-symptoms in *Section* 2.7

Section 1.2.5 summarises the features used in this paper. Feature engineering is not a simple task, and information about the signal is almost always lost in the process. More recently, biologically inspired neural networks have been proposed to bypass the feature engineering step and extract information from raw representations of data. Neural networks will be detailed in *Section* 1.3.2 and their applicability investigated in *Section* 2.6.

[6] Inertial Measurement Units (*IMUs*) are electronic devices which measure both acceleration (x,y,z) and direction (pitch, roll, yaw) over time. This is generally done with an accelerometer and gyroscope.

1.2.1 General Signal Processing

This thesis will focus on signal processing for time-series sensor data. A signal can be represented as an array with time on one axis and the sensor measurements on the other. The *frequency* of a signal refers to the rate at which measurements are made (in measurements/second). For example, an average microphone would record the value of a sound wave at around 44.1kHz, whereas an IMU would record six values for acceleration and rotation in the x, y and z direction at a frequency ranging from 50Hz to 4000Hz. *Noise* refers to deviations between the measured and true values, typically introduced by low quality recording equipment. This section outlines simple signal processing techniques which can be applied in most domains.

Moments are basic statistical descriptors of a signal, with the first three moments representing mean, variance, skewness. Typically up to five moments are used in the signal processing of biological signals. For waveform signals such as voice, mean is generally uninformative and variance corresponds to volume, whereas with accelerometer data, the mean represents the average velocity of acceleration. The zero or mean *crossing rate* is a measure of how rapidly the signal oscillates around a certain value. It is a very simple measure of the lowest frequency of the symptom.

Entropy describes the amount of information in a piece of data, if it were modelled by a Bernoulli scheme. In the context of signal processing, it is a simple measure of the complexity of a signal. When there are two dimensions of data (e.g, x and y of an accelerometer) *mutual information* and *cross correlation* can be applied. Mutual information is a measure of the amount of information obtained of one signal when observing the other and cross-correlation is a measure of the similarity of the two signals. For continuous time signals these measures are approximate by binning the values, with a recommended $\sqrt{\frac{len}{5}}$ bins [72].

The *Fourier* transform is one of the most fundamental tools in signal processing, decomposing a time-series signal into the magnitudes of frequencies that compose it. This is referred to as mapping from the *time* domain to the *frequency*, or *spectral* domain. Given an accelerometer signal, the Fourier transform can determine the amount of tremor in certain frequency bands — for example, PD tremor is often stronger in the $3.5 - 7$Hz band [35]. The *short-time*[7] *Fourier transform* (STFT) is often used when modelling evolving signals such as those generated during speech and walking [73, 74].

[7] *Short-time* signal processing involves analysing short 'windows' of the data to understand how it evolves over time. This provides more information but increases the complexity of analysis. Features extracted on the short time Fourier domain are often referred to as spectral features.

1.2.2 Voice

PD diagnosis with vocal features is a promising option for machine learning. Minor vocal symptoms have been shown to be detectable before other symptoms [31], and microphones are readily available. A high quality microphone is not required to perform diagnosis, with research showing that phone-quality audio is sufficient to perform diagnosis [25]. This gives rise to the possibility of at-home diagnosis or monitoring with a smartphone.

Biological Background

Speech production consists of two components: the vocal folds and vocal tract.

The vocal folds are housed in the larynx and consists of a flap called the *glottis*, which can be opened and closed. During speech production (phonation), air expelled from the lungs builds pressure below the glottis. The imbalance of pressure above and below the glottis causes it to oscillate, producing sound. Muscles in the vocal folds allow adjustments to the frequencies of sound produced within a certain range. The lowest of these frequencies — the *fundamental frequency*, f_0 — correlates to duration of one oscillation and is denoted as the *glottal cycle* or *pitch period*. The higher frequencies are referred to as the *harmonics* or *overtones*. Physical characteristics such as age and especially gender affect the size of the vocal folds and range of frequencies producible.

The vocal tract comprises the components between the larynx and lips such as the mouth and nose. These components act as a resonator, 'shaping' the sound by amplifying and attenuating certain frequencies produced by the vocal folds. The vocal folds and tract can be viewed as a *source-filter model*, where the vocal folds (source) generates the sound (signal) which is shaped by the vocal tract (filter).

Traditionally, the source-filter relationship of the vocal tract was assumed to be *linear*[8] and *time invariant*[9]. This assumption grants the use of a rich set of tools in the well-understood field of linear, time invariant systems theory. However, recent works analysing speech show strong evidence that these linear assumptions do not hold for most speech signals [75, 76, 77]. Non-linear signal processing is a less precise field, with most algorithms providing estimations of the true properties of underlying phenomena.

PD vocal symptoms can be broadly classified as dysphonia [78] — impairment in the production of sounds and dysarthria [79] — difficulties in the articulation of speech. Dysphonia arises from problems in the vocal folds, and dysarthria the vocal tract.

[8] Mathematically, a *linear function* f satisfies $f(a+b) = f(a) + f(b)$ and $f(ab) = af(b)$.
[9] *Time invariant* filters produce the same result for the same data independent of time or position.

Dysphonia is often described as a 'breathy' or 'hoarse' voice. As fine motor control is diminished in people with PD, they exhibit incomplete vocal fold closure. Turbulent airflow causes each glottal cycle to vary more than a healthy speaker. However, similar phenomenon occurs when the vocal cords are damaged or irritated by causes such as colds. It is unknown whether differentiation between neurological and physical causes of dysphonia with microphones is possible.

Dysarthria arises from the loss of both motor and cognitive control. People with dysarthria experience hesitant speech due to slower cognition, and slurred or imprecise articulation from the loss of fine motor control in the vocal tract. It is more difficult to quantify as signal processing must be done in the short time domain.

Speech Signal Processing

Parkinson's disease diagnosis with speech diverges to two distinct sub-fields: quantifying dysarthria in spoken sentences and quantifying dysphonia with sustained vowels (e.g, /aa/). To obtain a reliable diagnosis, both dysphonia and dysarthria related features will have to be considered.

Although changes in speaking patterns (dysarthria) are noticeable to human ears, they are difficult to quantify with current technologies. There are also a number of complexities involved in modelling *spoken language*, with a wide variation of accents and styles. Hazan et al. [53] investigates PD diagnosis on English and German sentences, observing that machine learning models trained on the English speakers do not generalise well to the German speakers and vice versa.

The Interspeech 2015 [54] competition featured a sub-challenge where the extent of PD dysarthria (as rated by the UPDRS) was estimated based on sentence and word pronunciations. The challenge dataset consists of pronunciations of isolated words and sentences from 50 patients in a controlled environment with a professional grade microphone. The best performing papers in this sub-challenge only managed Pearson correlations of 0.4 to 0.64 against neurologist diagnosis [80, 81, 82]. However, recent works point to evidence that speech can be a powerful predictor with better signal processing approaches. Vasqeuz et al. [83] enhanced noisy PD speech data using a technique proposed in Wang et al. [84], which decomposes speech into signal and noise subspaces. Orozco et al. [55] showed that the transitions between voiced and unvoiced speech is a strong indicator of PD.

***Highlight* 1.7.** As dysarthria (speech) is difficult to quantify, dysphonia (vowel phonation) based signal processing methods currently show more promise.

Sustained vowel phonations are the preferred method of quantifying dysphonia. A visualisation of heavy dysphonia in vowel phonation is presented in *Figure* 1.1. Early dys-

phonia analysis is based on variations of jitter, shimmer, and the harmonics-to-noise ratio. *Jitter* measures the variation in the length of each glottal cycle, and *shimmer* [85, 86] the variation in amplitude (volume). The harmonics-to-noise ratio (*HNR*) [87] measures the amount of noise in a signal, which correlates with the 'hoarseness' or 'breathiness'of speech, which arises from the incomplete closure of the glottis. The Glottal to Noise Excitation (*GNE*) ratio was introduced by Michaelis et.al [88] and is a more robust measure of dysphonia than HNR [89].

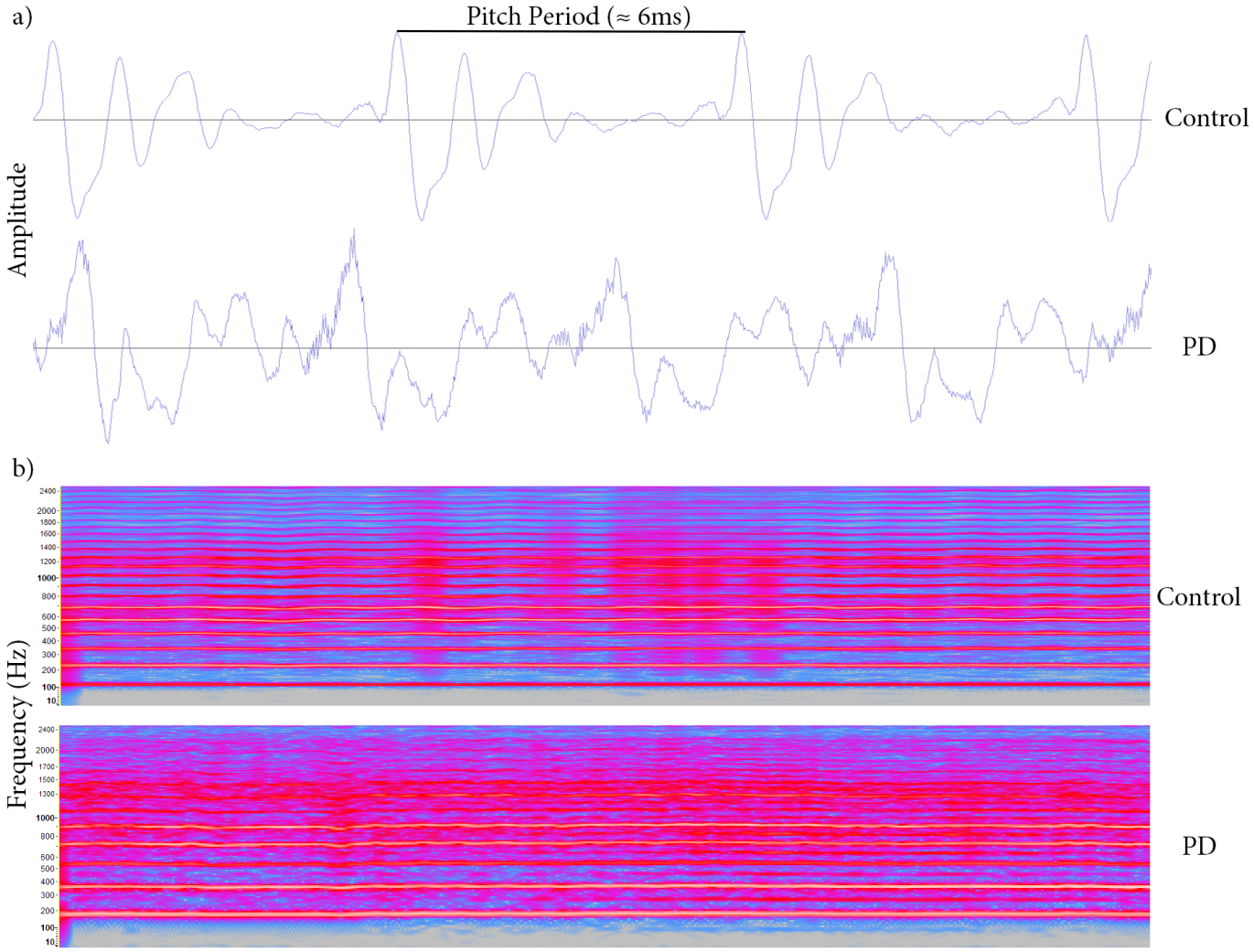

Figure 1.1: A visualisation of prominent dypshonia in /aa/ phonation on the time (a) and short time spectral domain (b, Mel-scale [90]). Cases are generally not as extreme and the natural variation in voice makes differentiation a difficult task.

Tsanas [91] extends GNE to develop the *Vocal Fold Excitation Ratio* (*VFER*). VFER is another quantification of the 'breathiness' effect in dysphonic speech from turbulent airflow. Tsanas also introduces the Glottal Quotient (*GQ*), which measures the standard deviation of the duration when the glottis is opened versus closed. Both VFER and GQ are built upon concepts of the DYPSA [92] fundamental frequency estimation algorithm.

Mel-Frequency Cepstral Coefficients (*MFCC*) have long been used for speech recognition [93], and have also shown promise in detecting dysphonia [94]. They are the most common and often the only feature used in speech recognition systems, but they lack interpretability and are very sensitive to noise [95]. There are also a variety of feature sets used in general speech classification, such as the 6,368 feature ComParE set [96]. Although these features may not be designed specifically for dysphonia, they are effective in fields such as speaker trait classification and may be useful in providing contextual information for complex machine learning models. The incidence of PD varies based on age, gender and race [97, 98], and it is likely that dysphonia presents itself differently depending on speaker traits. We refer to Eyben [32] for a comprehensive description of these features as well as a summary of feature sets used in speech classification.

Methods used in non-linear dynamical systems[10] have also been effective in dysphonia quantification [99]. *Detrended Fluctuation Analysis* (*DFA*) was originally introduced as a measure of the autocorrelation[11] of a signal [100]. Little et al. [99] shows this correlates with the amount of turbulent airflow in speakers with dysphonia. Little et al. also proposes Recurrence Period Density Entropy (*RPDE*), which characterises the periodicity of a signal. These measures are expected to be lower for speakers with dysphonia due to the noise introduced by turbulent airflow. Little et al. [25] builds upon RPDE to develop Pitch Period Entropy (*PPE*) which is a better measure of the impaired control of pitch experienced by PD patients.

1.2.3 Movement

Despite a similar amount of literature existing in movement and voice feature engineering, signal processing in the voice domain is more developed. Feature engineering for movement data diverges into a number of subfields, each developing different measurements for different sensors to quantify the extent of a movement disorder.

People with PD exhibit increased tremor, particularly in the 3.5-7Hz range [35], as well as distinct patterns of postural sway. Mediolateral (left-right) sway is generally a better indicator of PD than Anteroposterior (forwards-backwards) sway [101]. A Fourier transform can quantify the amount of tremor in IMU sensor data. Postural sway is best measured when the subject attempts to stand as still as possible. Both IMUs and force plates can quantify this — IMUs have the advantage of being cheaper and more accessible; however, have lower resolution and may not be spatially accurate [37]. Medication such as Levodopa is known to significantly increase the amount of postural sway [101].

[10] Dynamical systems theory is used to describe the behaviour of deterministic systems which appear to exhibit unpredictable behaviour based on a number of initial conditions. This will be explored further in *Section* 1.2.4

[11] *Autocorrelation* describes the similarity of a signal to itself when offset by a given interval.

Abnormal gait is another major indicator of Parkinson's disease. Gait related symptoms are mostly presented in the forms of Bradykinesia and Akinesia. *Bradykinesia* describes the slowness of movement, with individuals with PD generally showing a decreased cadence (steps per minute) and stride length [23]. *Akinesia* describes the involuntary loss of movement, and often presents itself in the form of 'freezing of gait,' where a patient experiences a sensation like their feet being glued to the ground [102]. Individuals with PD also exhibit interesting heel-to-toe characteristics, where their foot lands flat on the ground (as opposed to the heel striking the ground first, as in regular gait) [103, 44].

Gait can be quantified using IMUs, force walkways and motion capture. IMUs are undoubtedly the cheapest and most available option; however, gait characteristics can only be estimated. IMUs attached to the foot shank allow relatively precise estimation of step length and cadence [46, 48], whereas handheld or in-pocket gait estimation is a more challenging task [47, 104]. IMUs and other sensors embedded in the shoe can quantify heel-to-toe characteristics. Force Walkways and motion capture are more precise, but costly alternatives for measuring gait.

Smartphones

Smartphones are one of the fastest adopted technologies in history, becoming increasingly prominent in developing countries. As they possess a number of sensors such as IMUs, microphones and cameras, they are a promising tool in *telemedicine* — the remote diagnosis or monitoring of PD. The universal nature of smartphones allows for large PD datasets, with the 8,000 participant mPower [105] dataset used in this paper crowdsourced from smartphone users.

Smartphone studies use features presented in voice and accelerometer research, along with additional tests such as memory or tapping tasks [52]. However, the resolution and accuracy of smartphone sensors greatly varies and introduces significant noise to the data. The influence of smartphone models on results has yet to be investigated, and it is unknown whether generalizing between phones is possible. Smartphone step and motion mode recognition[12] [106, 107] is a similar research area; however, techniques are less applicable as measures are often more coarse.

Little et al. [25] provides evidence that a high quality microphone is not required to classify dysphonia, obtaining good results with phone-call recordings. Brunato et al. [39], Boussios et al. [38] and Arora et al. [19] also manage to obtain good results with simple smartphone IMU-based features. However, all of these models have been tested on small

[12] *Motion mode recognition* involves classifying whether the user has their phone in their pocket, hand, bag

datasets, which are prone to overfitting on cross validation [21] from bias and uninformative predictors [20].

Zhan et al. [22] conducted a smartphone feasibility study on the largest dataset to date — 121 PD and 105 controls. Participants were recruited into the study and asked to perform tasks such as walking, saying /aa/ and alternated tapping [52]. However, Zhan et al. obtained 71 per cent accuracy — especially poor considering that the mean (standard deviation) age of PD subjects was 57.6 (9.4) and control 45.5 (15.5). A similar result may be obtained by a model classifying with age alone. This result is in direct contradiction with previous works such as Arora et al. [19] which reported 98 per cent accuracy on very similar accelerometer features. It is clear that reported results must be taken with a grain of salt.

Neto et al. [108] is the first study based on the 6,000 participant mPower dataset [105]. Neto et al. focused on the data analysis aspect, investigating the impact of medication and the "time of the day" effect on activity performance. The features used in this analysis are not described — the author list and citations suggest that accelerometer features used were an extension of Arora et al. [19].

Neto proposed a method of "collapsing" multiple recordings for a participant into one by taking the median value of each feature over all the recordings. This improves model performance, but may not be entirely valid. In mPower, participants with PD often perform more recordings than other participants, who are likely less serious about the study. When taking the median of feature values, unrealistic combinations of values may arise as they may be from different recordings. The model may be learning to associate PD with these strange combinations of 'median-like' values.

1.2.4 EEG

Electroencephalogram (EEG) signal processing presents an interesting challenge as the characteristics of an EEG signal are less well understood compared to speech and motion. There are specific features for diagnosis PD or Alzheimer's disease[13]; however, this section will focus on the techniques applicable to speech and movement data.

Non-linear dynamical systems theory inspires many EEG signal processing techniques, as EEG signals are believed to be generated by non-linear coupling interactions between neuronal populations [109]. Patients suffering from neurodegenerative disorders often exhibit decreased complexity in EEG patterns, believed to be caused by the a decrease in non-

[13] Alzheimer's disease is a similar neurodegenerative disease. It is characterised by a loss of neurons in the cerebral cortex whereas PD targets dopaminergic neurons.

linear cell dynamics [110]. Features developed with EEG signal processing aim to characterise the dynamic structure of this system. As these features are not related to human senses, they may be promising when applied to the task of measuring the presence of voice or movement symptoms undetectable by a neurologist.

***Highlight* 1.8.** The nature of features used in EEG make them promising for the task of detecting the presence of symptoms undetectable by a neurologist.

Chaos theory is a field analysing systems which are sensitive to initial conditions. One can imagine the generation of a speech signal as a system, where parameters involve the state of components in the vocal tract. Given no change in parameters such as vocal fold tension, regular /aa/ phonation can be modelled with much fewer dimensions in *phase*[14] space [111, 112]. A recurrence plot is commonly used to understand chaotic systems and is depicted in *Figure* 1.2.

The *Lyapunov Exponents* quantify the divergence of two systems with infinitesimally similar initial parameters. The *Largest Lyapunov Exponent* (λ^*) characterises the chaos in a system and is commonly estimated with Rosenstein's algorithm [114] which reconstructs the system's dynamics using a time delay technique. Its inverse, $\frac{1}{\lambda^*}$ is the Lyapunov time, which defines how long the behaviour of a system can be predicted. The λ^* has long been used in the EEG analysis of sleep and as a feature for machine learning [115, 116]. The λ^* have also been applied in the analysis of speech [117, 118], gait and balance [119, 120, 121].

The *fractal dimension* is another measure commonly used in the analysis of EEG and other dynamical systems along with the LLE. It represents the ratio of the log change in detail to log change in scale of a signal[15] [122]. A higher value correlates to a more complex signal, and the fractal dimension of an EEG signal with open vs. closed eyes and normal vs. epileptic states are observably smaller [123, 124]. The fractal properties exhibited in neuronal control are reflected in heartbeat and gait [125] with force plate data from elderly and Parkinsonian subjects showing a significant increase in fractal dimension compared to healthy young subjects [126, 127, 128]. Esteller et al. [129] compares algorithms estimating the fractal dimension of signals.

Fisher Information is a measure relating to the uncertainty of measuring a variable (signal) about the unknown parameters modelling its distribution [130]. It is applicable in quantifying non-linear dynamics [131] and has been applied in the analysis of EEG [132]. General entropy will not differentiate two sequences where the frequency of each variable

[14] The *phase* space represents all possible states of a dynamic system.

[15] The coastline paradox is the observation that as you measure a coastline with increasingly smaller measuring sticks, the measured coastline length will increase. The *fractal dimension* would measure the ratio of change in length as of the 'stick' used to measure the coastline is made shorter.

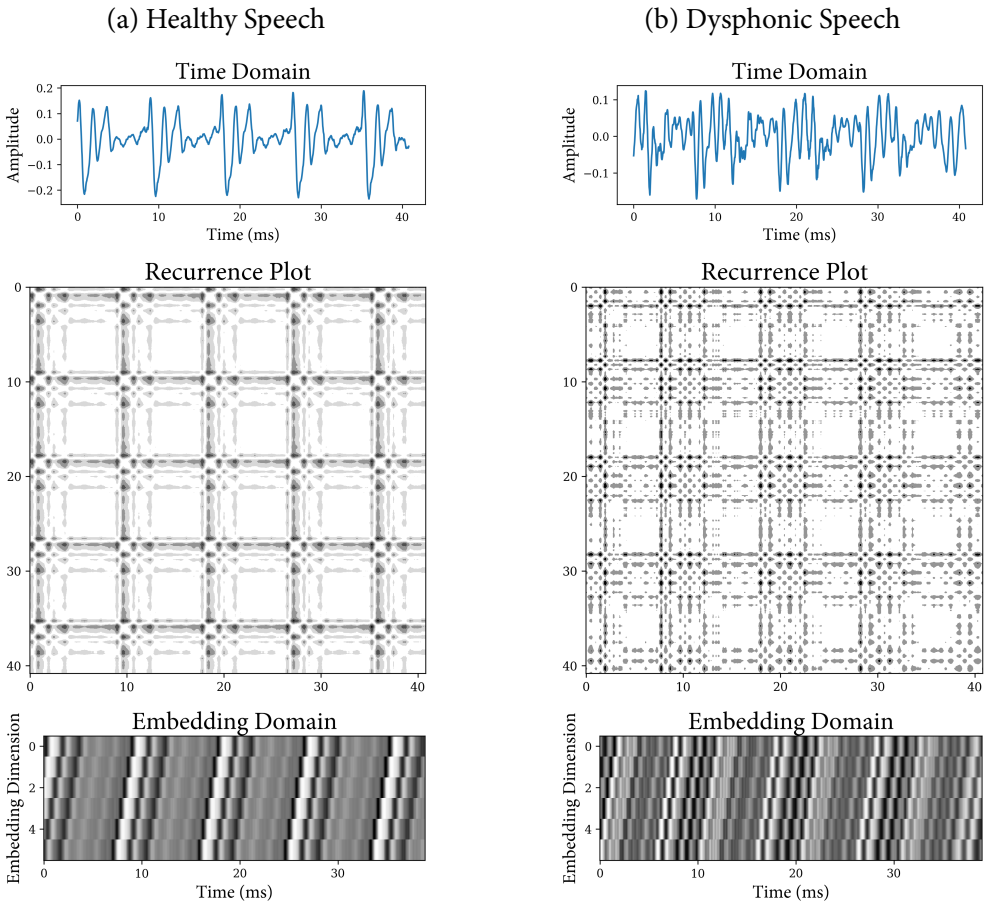

Figure 1.2: A visualisation of speech on various domains. The recurrence plot [113] was developed to visualise the periodicity of a signal, highlighting where the system approximately reaches the same state. The time-lagged embedding domain is used by Rosenstein et al [114] to calculate non-linear measures such as the largest Lyapunov exponent and in measures such as Fisher Information or SVD Entropy.

is the same; however, the sequences 0,0,0,0,1,1,1,1 and 0,1,0,0,1,1,0,1 are clearly generated by different stochastic processes. *Approximate* and *sample entropy* are similar measures to entropy; however, they quantify this unpredictability in a signal [133, 134]. The multi-scale sample entropy [135] is especially powerful tool in the analysis of biological signals [136, 137]. Although approximate and sample entropy are prominently used in EEG signal processing, they are rarely applied to voice and movement analysis.

Although signals may appear to have high information content on the time domain, they may be easier to represent on others. For example, the JPEG image compression format [138] primarily relies on the inability of human vision to perceive high frequency information. Images are compressed by taking a Fourier transform and reducing the amount of high frequency information. *Spectral entropy* measures the information content of the signal in its frequency domain representation. The singular value decomposition (SVD) factorises a matrix M into $U\Sigma V$ where U and V are orthogonal matrices and Σ is a diag-

onal matrix of *singular values*. *SVD entropy* [139] measures the entropy of the singular values obtained when the signal is embedded with Rosenstein's [114] technique.

Like DFA, the *Hurst* exponent characterises the autocorrelation of a signal [140]. DFA is a generalisation of the Hurst exponent for non-stationary processes[16]; however, DFA has been criticised for being biased towards underestimation in short signal recordings, and its ability to address non-stationary signals questioned [142]. For self-similar signals, the Hurst exponent relates directly to the fractal dimension. In general the measures are independent, with the Hurst exponent characterising the global rather than local properties of a signal [143]. A Hurst exponent less than 0.5 characterises the signal 'switching' between high and low values, 0.5 characterises random walk like behaviour, and values greater than 0.5 imply positive autocorrelation. Like fractal dimension, the Hurst exponent is a valuable tool in the analysis of gait and balance [144].

1.2.5 Summary of Features

Tables 1.3 to 1.6 summarise all of the features used in the construction of models presented in this thesis. The libraries and parameters used to implement them are described in *Section* 2.8. Features are organised by field of introduction or by their most applicable field. All relevant general and EEG features are calculated for both the audio and accelerometer models used in this thesis. Additionally, RPDE and DFA are used as a feature in accelerometer based models.

[16] Non-stationary systems are those with properties which evolve over time. A grid-like, repetitive recurrence plot (*Figure 1.2* is indicative of a stationary system, whereas curved patterns are signs of a non-stationary, chaotic system [141]

Table 1.3: Features and techniques which are applicable to any signal processing problem.

General Signal Processing

Feature	Description
Moments	Statistical features — mean, variation, skewness, kurtosis, etc.
Crossing Rate	Rate the signal oscillates around a value — usually zero or the mean.
Information Theoretic	Entropy, mutual information, cross-correlation and related measures based on the information content of signal.
Spectral Flux	Rate at which the power spectrum changes
Fourier	Transforms the signal from time domain to frequency/spectral domain. Quantifies the *power* of a signal at a given frequency.
Wavelet	A variation of the Fourier transform with a different bases, allowing it to quantify both time and frequency
Energy Operators [145]	Quantifies the instantaneous amplitude and frequency of a signal. Common operators are Teager-Kaiser (TKEO) and Squared (SEO)

Table 1.4: Dysphonia signal processing generally quantifies the variation in each glottal cycle during speech production

Voice – Dysphonia

Power Cepstrum	From the inverse Fourier transform. Commonly taken in the Mel-log scale [90], resulting in the MFCC [93]. Minimal interpretability, though it is the primary feature used in speech recognition [94].
Pitch [146]	Although obtainable with a Fourier transform, pitch often refers to estimating the exact duration of each glottal cycle.
Loudness	The volume of a sound in relation to human hearing. Only meaningful if recording setup is strictly controlled.
Formants	The resonance frequencies of an audio sample.
HNR [87, 147]	Measures the ratio of noise in a voiced signal (signal to noise)
Jitter [86]	Measures of the variation between the length of each glottal cycle.
Shimmer [85]	Measures of the variation of amplitude between each glottal cycle.

LPCC [148]	Coefficients of an *autoregressive* model which measures how well a signal can be modelled linearly by its previous values.
GNE [88]	An extension of HNR by Michaelis et al. [88] to improve reliability in dysphonia quantification.
VFER [18]	An further extension of HNR, building upon the theory of GNE.
EMD-ER [149]	Another technique developed based on non-linear speech theory to quantify signal to noise
GQ [18]	Measures standard deviation of duration the glottis is opened vs. closed.
DFA [99, 100]	Detrended Fluctuation Analysis. A generalisation of the Hurst exponent which measures the autocorrelation of a time series.
RPDE [99]	Measures the periodicity of a signal, specifically designed with non-linear speech as the target.
PPE [25]	Measures the stability of pitch in sustained phonation.
Wavelet Measures [26]	A set of 180 measures for dysphonia based on wavelet transforms to the f_0 of speech introduced by Tsanas et al. [150].
GeMAPS [151]	A minimal acoustic feature set of 58 or 87 (eGeMAPS) parameters that performs well in general speech classification [32].
Interspeech ComParE [152]	An exhaustive 6,368 feature set for general speech classification [32]. Feature/dimensionality reduction generally improves performance unless data is plentiful.

Table 1.5: There are few movement specific features, with most based on simple measures of postural sway or irregular gait.

Movement

Fourier Bands	The power in bands such as 3.5hz-7hz compared to 7hz-12hz are the primary features used to detect Parkinsonian tremor.
Jerk [153]	The change in acceleration. The jerk signal may be more effective when combined with certain signal processing methods.

Sway Area	This can be calculated naïvely by multiplying the range of sway in the A/P and M/L directions or by fitting a bounding ellipse in the principal component axis [154]. As A/P and M/L directions are lost in accelerometer data, the bounding ellipse method is used.
Cadence Measures	The steps per minute, variation in time taken for each step, difference between left and right stride times.
Stride Measures	The length of each step and variation in step lengths. This was not measured as leg length is not available in the dataset used [104].

Table 1.6: EEG signal processing is often based on non-linear systems theory. These features may be effective in detecting the presence of symptoms invisible to neurologists.

EEG

Hjorth Parameters [155]	Three simple statistical measurements of a signal which have been used as features in EEG and IMU models [156].
Lyapunov Exponents [157]	Characterises the divergence of systems with close initial conditions. The largest exponent (λ^*) [119] is generally the most informative.
Fractal Dimension [122]	A measure of how the detail in a signal changes with the scale at which it is measured. The Higuchi [158] and Petrosian [159] fractal dimensions are used in this thesis.
Hurst Exponent [140]	Characterises self-similarity. DFA is a generalisation of the Hurst Exponent and is robust to non-stationary signals. The difference in measurements may be informative.
Fisher Info [132]	Quantifies the non-linear dynamics in the system generating a signal.
Ap/Samp Entropy [134]	Approximate and sample entropy quantify the unpredictability of a signal. Multiscale entropy increases information content [135].
Spectral Entropy	Measures the regularity of the spectral (frequency) distribution. A high spectral entropy implies sharp differences in frequencies present in the signal.
SVD Entropy [139]	A measure of complexity. The entropy of the singular values of the signal after applying the time delay embedding method [114].

1.3 Machine Learning

***Highlight* 1.9.** Fundamentally, the goal of machine learning is to use past data to make accurate predictions about new data.

Machine Learning tasks can be classified as classification or regression, and supervised or unsupervised. Classification involves predicting the *class* of a datapoint — for instance, distinguishing PD from control — whereas regression involves predicting a numerical value, such as the UPDRS motor scores. In supervised learning, the data is *labelled* with the ground truth — i.e., whether the subject has PD — whereas an unsupervised model must find patterns in the data without any prior knowledge. This section will focus specifically on *supervised binary classification* (two classes).

Supervised binary classification can be viewed as 'learning' a model which given a set of numerical input features, predicts a class 0 (control) or 1 (PD). This can be visualised as a function $f : \mathbb{R}^d \mapsto \{0, 1\}$ where d is the number of features used in the model. The edge where the f transitions from zero to one is denoted the decision boundary (or 'hyperplane') which partitions the data into the two classes. This is depicted in *Figure* 1.3.

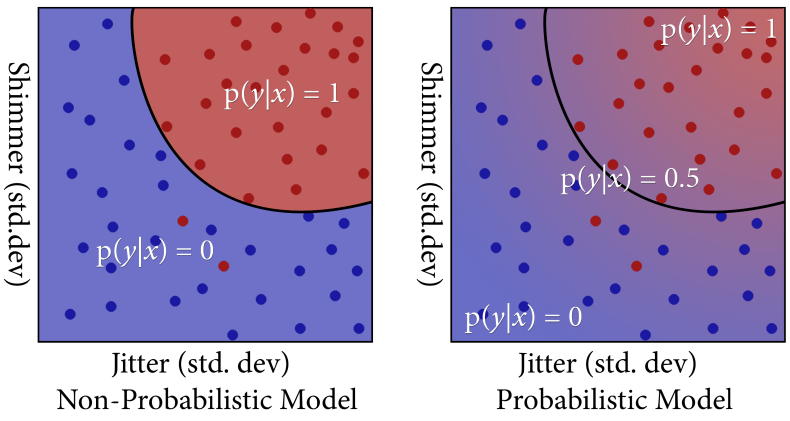

Figure 1.3: A visualisation of binary classification with two features. Data is rarely as 'clean' as this artificial example.

Traditional machine learning models are built on solid statistical and mathematical foundations, and are well understood. However, these models have been developed on assumptions that are rarely satisfied by real-world data. Models such as deep neural networks have started to rise to popularity recently due to their modelling power; however, their behaviour is still poorly understood and they are difficult to analyse [160].

Models have strengths in different areas, and very rarely does a model strictly dominate another. The structure of the data informs the choice of model. For example, models like deep neural networks may perform well when data is plentiful, but in small datasets simple decision (*Section* 1.3.1) trees may outperform neural networks (*Section* 1.3.2).

***Highlight* 1.10.** There is no 'best' model — the choice of model is informed by the data.

The predictive error in any model can be decomposed as *irreducible error*, *bias* and *variance*. Irreducible error occurs when the features used are too noisy[17] or unrelated to accurately predict the data. An optimal model cannot achieve performance beyond this irreducible error. Bias describes a model 'fitting' the data poorly — a model with high bias will have low accuracy. Variance describes how 'unstable' a model is — a model with high variance may score 100 per cent accuracy but generalise poorly to new data. A model with high variance is essentially predicting results by 'memorisation.' Fitting a model with high variance is often known as *overfitting*. The bias-variance tradeoff [161] is a fundamental problem in machine learning, describing the difficulty in reducing bias without increasing variance and vice versa. Bias and variance are depicted in *Figure* 1.4.

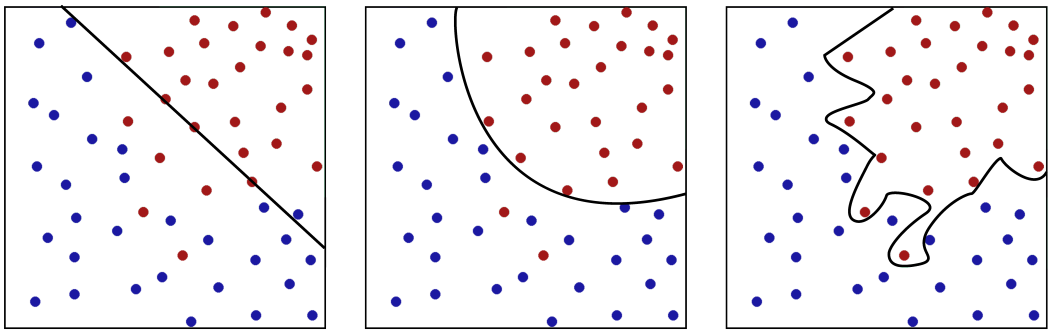

Figure 1.4: Machine learning models and their parameters must be carefully chosen to ensure the optimal fit.

Overfitting is a major issue in machine learning as data is limited and models are often too complex to analyse. Visualising and detecting overfitting may be simple when fitting a function in two dimensions, but it is significantly more difficult when the input has thousands of dimensions. Cross validation is the gold standard in machine learning when it comes to model evaluation and recognising overfitting; however, there are a number of caveats in its application. Cross validation, its caveats, and other techniques used for model evaluation will be discussed in detail in *Section* 1.3.4. Like any statistics-based field, careful analysis of the results is required; unfortunately, this is often neglected in machine learning literature.

The following sections will cover three common models: random forest classifiers, support vector machines, and neural networks. The mathematical formulation of these models is abstracted in favour of intuition behind their behaviour. We refer to Bishop et al. [162] for a formal description of these models.

[17] *Noisy* in the context of machine learning of signal processing relates to the inherent variance of a measure. An inaccurate, low quality sensor can be considered 'noisy'.

1.3.1 Traditional

Traditional models are the approach favoured in current literature [33] due to the small dataset sizes. The two most popular models used are *Random Forests* (of decision trees) and *Support Vector Machines* (*SVM*). Both of these are suitable for small datasets as they are relatively resistant to the curse of dimensionality[18] [163]. Both are also non-probabilistic classifiers[19]. There exists models which are inherently probabilistic such as Gaussian processes [165], but these are often out-performed by decision-boundary based models in binary classification.

Random Forest [166] classifiers are derived on the concept of Bootstrap Aggregation (*bagging*) [167] where the results of multiple models are combined to obtain better performance than any of the constituent models alone. Random forests aggregate *Decision Trees* which are one of the simplest and most common approaches to data mining and machine learning.

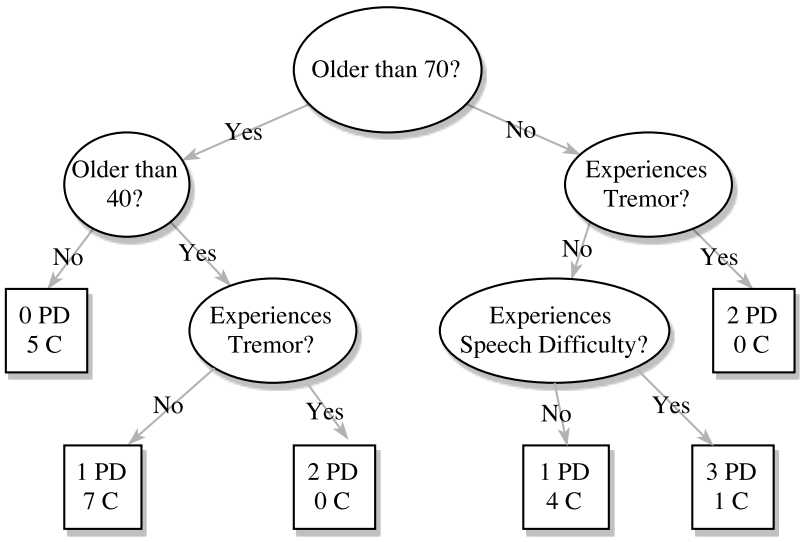

Figure 1.5: A simple Decision Tree with cutoff depth 3. Data is split by rules until a leaf contains only one class exists or a cutoff criterion is satisfied.

Decision Trees (*Figure* 1.5) are robust against high dimensional data and their results are simple to interpret and transparent to human users (compared to SVMs and neural networks). However, selecting the optimal decision rules and cutoff criterion is a NP-complete problem. Decision rules are often developed based on greedy algorithms related to information criterion or search. A deep decision tree is prone to overfitting whereas a shallow one underfits.

[18] The *curse of dimensionality* states that exponentially more training data is often required for each additional feature to ensure a complete and reliable model.

[19] In general. Random forests and SVMs can provide pseudo-probabilistic output [164].

Random Forests correct for the tendency of decision trees to overfit and offer robust and consistent results regardless of hyperparameters. The two hyperparameters are the number of trees to aggregate over and the number of features used in the search to split each branch of the tree. If the number of trees used is greater than the 'complexity' of the problem, additional trees will not affect results [168]. The square root of the number of features for classification is recommended by Breiman [166] and performs well in most cases. Therefore, it is rare to perform hyperparameter tuning on random forests. In general, they are easier to interpret than SVMs and neural networks.

***Highlight* 1.11.** Random forests provide robust and consistent results without the need for hyperparameter tuning.

Support Vector Machines [169] are built on the concept of creating the optimal decision boundary. The motivation is to create decision boundary which maximises the margin[20] between different classes. A Lagrangian can be used to mathematically solve for a linear decision boundary. As most problems are not linear, the *kernel trick* is used to transform the data into a linear space.

Kernels measure the similarity between two data points, and the kernel trick transforms the raw input into the feature space of the kernel[21]. Non-linear kernels enable a SVM to fit a non-linear function; however, selecting the perfect kernel is hard unless the exact non-linearity of the data is known. There are uncountably many kernels, and kernels such as the Radian Basis Function (RBF), Fisher and Polynomial are commonly used[22]. Kernels also have adjustable parameters, such as the degree and constant coefficient for polynomial kernels.

The original SVM algorithm was not able to handle cases where data was not separable. This led to the introduction of slack variables which define a penalty for data beyond the SVM's margins, extending their use to non-separable data [170]. The sum of these slack variables is added to the SVM's Lagrangian equation along with a constant scaling factor C. The parameter C balances the penalty for data beyond the margins with the size of the margin. A small C is incentive to create a large margin whereas a large C is incentive to minimise errors.

The combination of kernels and slack variables resulted in SVMs becoming popular in the machine learning community. A consequence of the improved performance is a reduction in interpretability — making the SVMs a 'black-box' method [171] — limiting its adoption in health care, where establishing user's trust through interpretable and

[20] The *margin* is the smallest distance between the decision boundary and any of the samples
[21] Kernels perform the same role as basis functions in linear regression
[22] There is rich literature in developing new kernels; however, these are rarely applied.

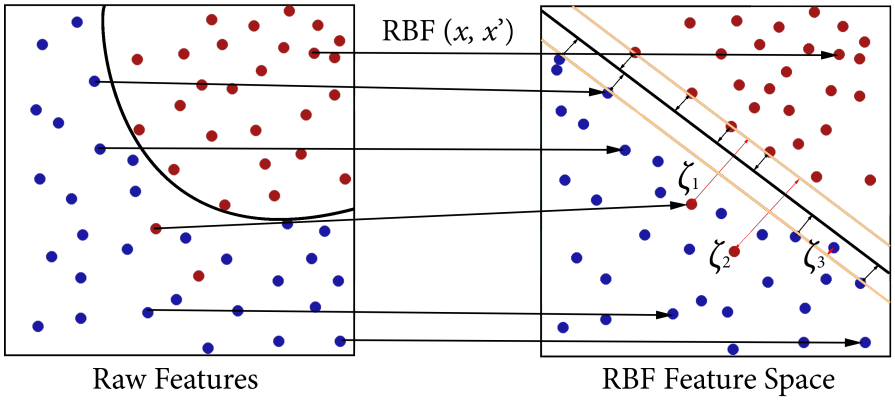

Figure 1.6: A RBF kernel is used to transform the data into a more linearly separable space. ζ_i denote slack variables which lie beyond the margin (depicted by beige lines).

easy-to-understand results is important. Kernels and slack variables also introduce many hyperparameters, such as the scaling factor C and the type of kernel. Although intuition and knowledge of the data can guide kernel and hyperparameter choice, techniques such as grid or random search [172] are often used to fine-tune them. However, hyperparameter tuning increases the risk of overfitting, which will be discussed in *Section* 1.3.4.

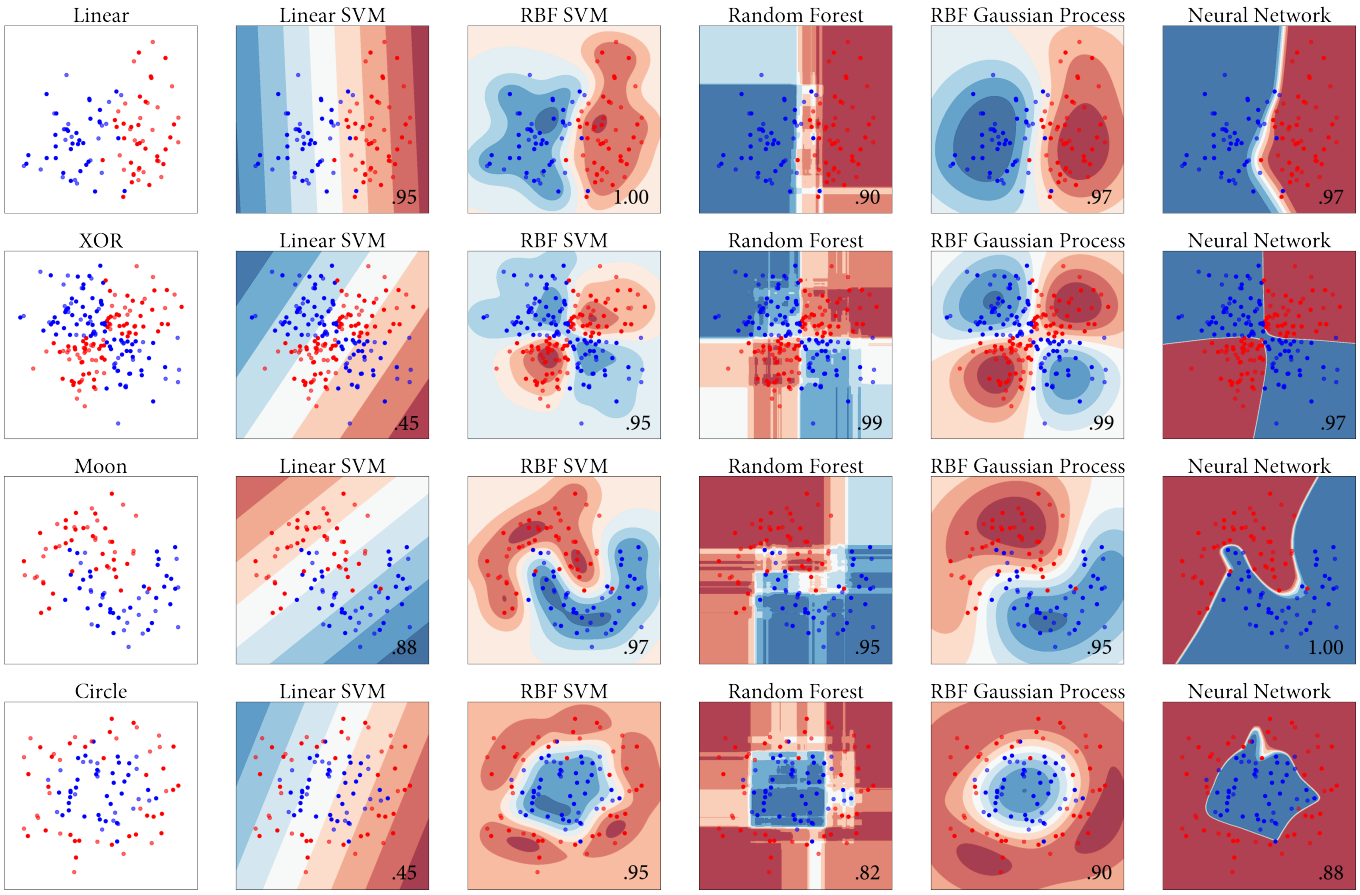

Figure 1.7: A visualisation of various models (x axis) fitting two dimensional distributions of data (y axis). Choosing a model is difficult when the characteristics of the data are not well understood.

1.3.2 Artificial Neural Networks

Traditional machine learning models perform best when data is structurally simple. Most statistical models such are designed to fit a linear function through the data, using predefined basis functions or the kernel trick to imitate non-linearity. Random forests and decision trees are powerful when data is readily available, but they do not model functions of data, and are less suitable when predicting unseen outliers [173]. Neural networks are popular models used when the dataset is reasonably sized and there exists a complex structure between the input features. They are extremely powerful, but very difficult to interpret or debug; they are also highly prone to overfitting.

Although neural networks have only recently risen to the spotlight, their history begins in 1943 with the introduction of a computational model of biological neurons[23] [174]. In 1958, the simple perceptron learning algorithm (*Figure* 1.8) was developed [175], which would become the building blocks of neural networks today. The fundamental concept of a neural network is to connect many perceptrons (acting as neurons) together to simulate the behaviour of a biological brain.

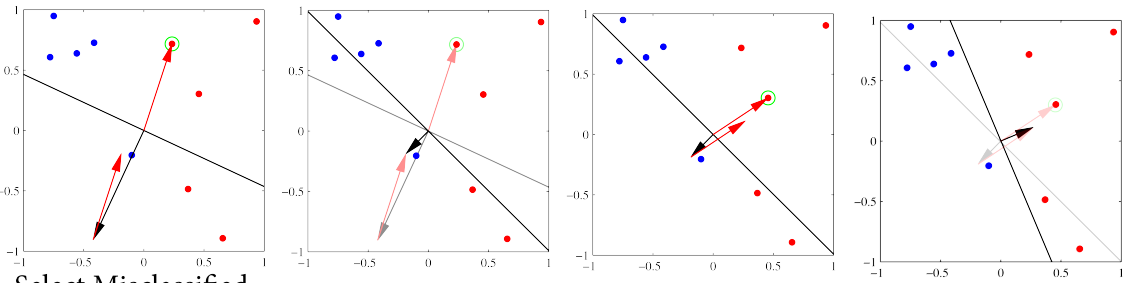

Figure 1.8: The simple perceptron learning algorithm. The original incarnation could not handle inseparable data [175]. Images adapted from Bishop [162].

A perceptron by itself (*Figure* 1.9) is a simple machine learning model, taking input features and outputting a value representing a class or probability. As neurons were thought to have two states — either firing or not — the output was passed through a Heaviside[24] *activation function*. At the time, computational power was limited, and large networks impossible to train. Early works by Minsky and Papert [176] were misinterpreted as stating that perceptrons were incapable of modelling the 'exclusive or' (XOR) function. However, Minsky and Pampert only proved this for a single perceptron and believed that multiple

[23] *Neurons* are cells which transmit information via chemical and electrical signals. They are the fundamental building block of the human brain.

[24] A discontinuous function which outputs either 0 or 1, defined as $H(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$

layers of perceptrons could model the XOR function. In 1989, it was shown that a single layer with enough perceptons can approximate any non-linear continuous function [177].

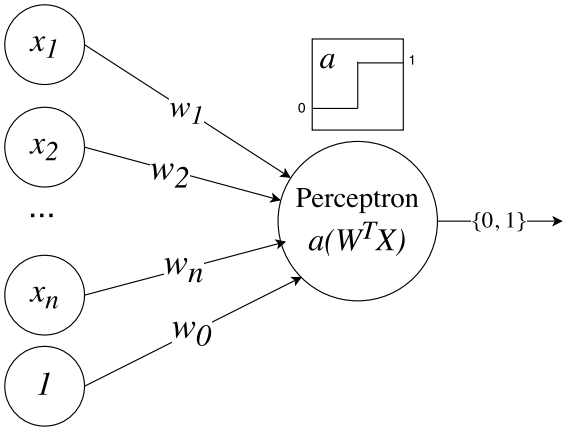

Figure 1.9: A single perceptron node. Takes input X and learns the weight vector W to classify the output with the Heaviside activation function a.

Multiple layers of perceptrons had always been the goal of neural network research; however, training them was not possible until backpropagation, a form of gradient descent[25] was developed [178]. Backpropagation required the activation function to be differentiable, hence the sigmoid[26] replaced the Heaviside activation function. Neural networks (*Figure* 1.10) were now trainable, although computational power would be a bottleneck for a couple of decades. *Deep learning* or *Deep neural networks* are a general term for neural networks with many (generally more than 3) layers.

***Highlight* 1.12.** A neural network's ability to learn complex non-linear relationships provides an advantage over traditional models, where this non-linearity must be defined.

Neural networks are computationally expensive models and training them requires optimising a non-convex function. This is computationally intractable, so current neural networks use gradient descent and backpropagation to find a good local optima [180]. The vanishing gradient problem [181] limited the depth of neural networks until the recent development of batch normalisation [182]. Previously, careful management of gradient flow was required to train deep neural networks [183].

Two major variations of the traditional fully connected structure are convolutional and recurrent neural networks. Convolutional neural networks (*CNNs*) are inspired visual cortex, where neurons are connected to local regions of the visual field. These networks contain

[25] Surprisingly, Werbos' work on backpropagation [178] was lost and would be rediscovered a decade later in 1985 by Rumelhart et al. [179]

[26] The sigmoid function is defined as $\sigma(n) = \frac{1}{1+e^{-x}}$

[27] Typically, neural networks are more complex, consisting of more hidden layers, a variety of activation functions, and partial/structured connections.

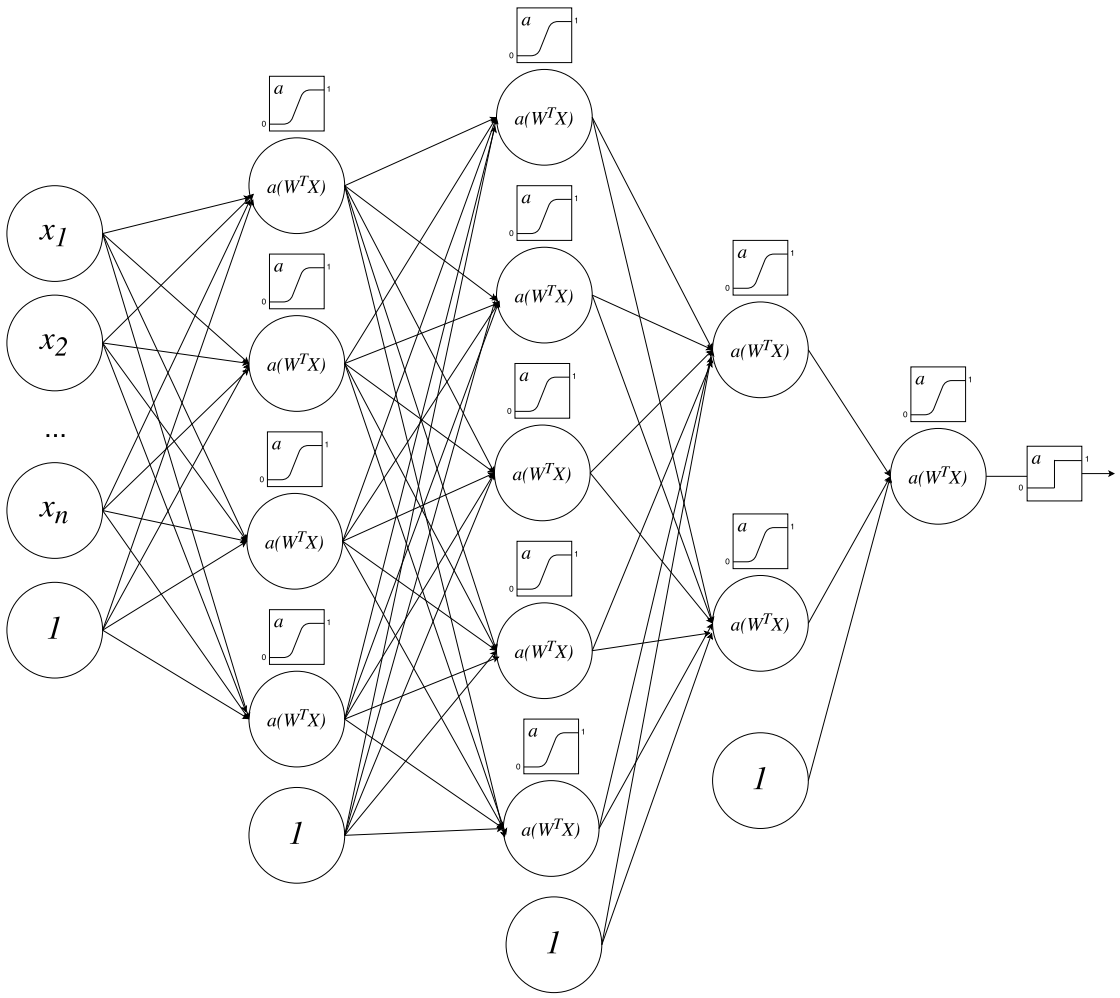

Figure 1.10: A simple[27] neural network with three fully connected hidden layers using sigmoidal activations. By stacking non-linear activation functions, neural networks are able to learn any non-linear function of the input. The '1' nodes represent the bias at each layer.

'convolution' layers where neurons are connected to a small local region of neurons in the previous layer. These convolution layers learn a hierarchy of features (*Figure* 1.11), and can negate the need for feature engineering for certain types of input data. Their power is clear in the task of image recognition, where CNNs have rapidly exceeded the performance of traditional models.

Recurrent neural networks (*RNNs*), on the other hand, have cyclic connections. This simulates an internal state which allows RNNs to base future predictions on past data, therefore better handling temporal information. There are a number of RNN variants, and long short-term memory (LSTM) nodes are the most common in practice as they are more robust to the vanishing gradient problem [185]. Recently, fusion models of LSTM and convolution layers have seen success in EEG classification [186] and multimodal activity recognition [187].

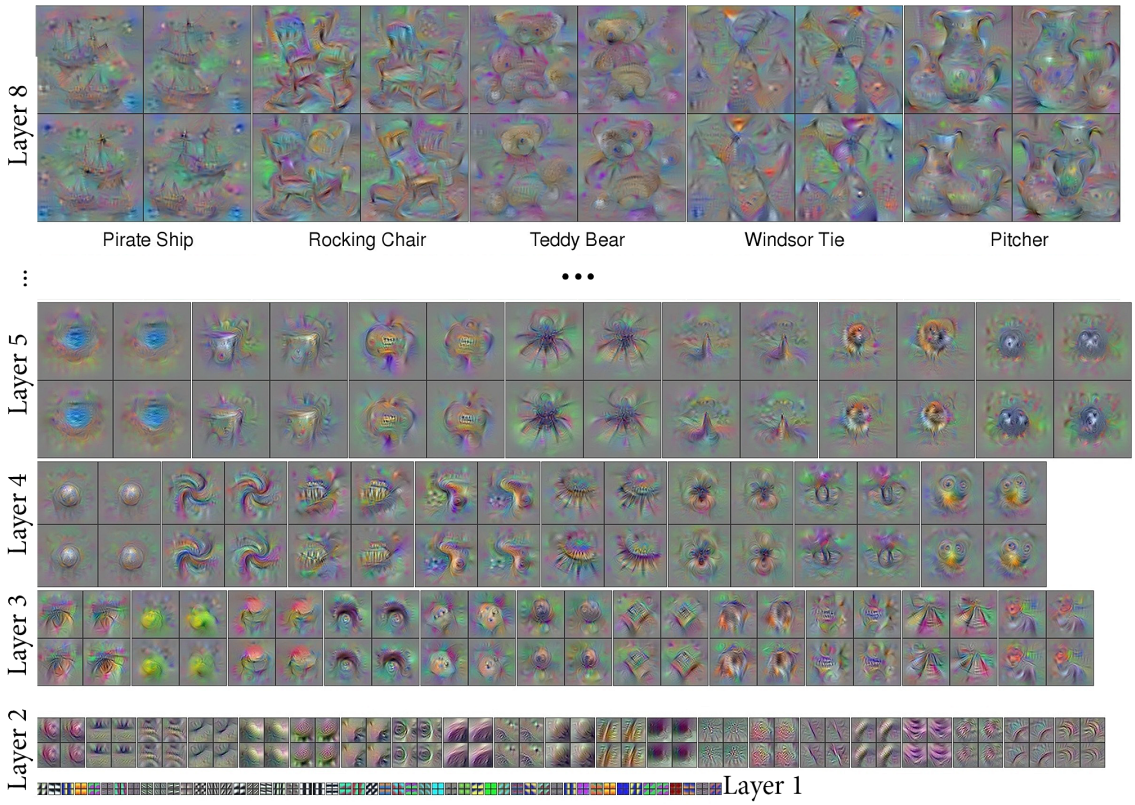

Figure 1.11: A visualisation of CNN activations from Yosinski et al. [184]. Layers capture increasingly complex relationships between pixels and are input features to further layers.

Neural Network Hyperparameters

Neural networks have many hyperparameters — combined with computationally expensive training, hyperparameter tweaking often relies on intuition. This section will provide a basic intuition behind selecting hyperparameters of a neural network.

The most fundamental features are the width and depth of the network. In general, increasing the number of nodes in a network reduces its bias and is effective when data is plentiful. It is thought that networks with many nodes per layer (*width*) are better at memorization, while additional layers (*depth*) are better at generalisation of features [188]. Depth can also be exponentially more valuable than width for modelling the structure of complex non-linear data [189]. There is still no consensus on the balance between number of nodes and layers — this varies significantly problem to problem.

A large neural network tends to overfit by memorising the data. Regularisation is a method of preventing this without reducing the size of the network. In traditional machine learning, l_1 and l_2 weight regularisation is most common. This involves adding a penalty to weights, motivated by Occam's razor where a simpler model (smaller weights) is preferred. When applied to neural networks, weight regularisation slows convergence and complex models can still be learned with a deep enough network. Early stopping and dropout are the most common forms of regularisation in practice.

Early stopping involves stopping training before the optima is reached, at the point where the cross-validation accuracy starts to decrease from overfitting. *Dropout* involves randomly disabling some percentage of nodes on each layer at each iteration of gradient descent [190]. This may appear unintuitive; however, the motivation is to promote redundant feature representations to improve its robustness. Dropout and early stopping are a powerful combination and used in most architectures. There are also variations of dropout such as dropconnect [191] where connnections rather than nodes are zeroed.

A major problem with the sigmoidal activation function is that as the activation approaches either 0 or 1 the gradient approaches zero. This is known as *saturation* and significantly slows the convergence of gradient descent in the training process. Rectified linear units (*ReLUs*) use the activation function $f(x) = \max(0, x)$ which resolve the gradient issue and are believed to be more biologically plausible [192, 193]. One notable characteristic of ReLUs is that once the unit outputs zero, it is essentially 'dead' as the gradient of the rectifier is zero. A number of modifications to ReLU have been proposed such as the leaky/parametric ReLU [194] ($f(x) = \max(\alpha x, x)$ for $\alpha \leq 1$), Maxout [195], noisy ReLU [192] and exponential linear unit [196].

The initialisation of the weights in the network will affect the solution found by gradient descent and the rate of convergence to it. Poor initialisation can result in the death of ReLUs or saturation of sigmoidal and tanh units. Recently, it was shown that initializing the weights according to a Gaussian distribution with variance $2/(n_{\text{in}} + n_{\text{out}})$, where n_{in} is the number of inputs to the node and n_{out} the number of outputs was highly effective in preventing saturation. This is termed Xavier or Glorot initialisation and is effective for networks with sigmoidal or tanh activations [197]. However, Xavier initialisation causes ReLUs to rapidly tend to zero, so small modification was proposed where variance is set to $2/n_{\text{in}}$, termed He initialisation [194].

The method of gradient descent, referred to as the *optimiser* is also a major area of neural network research. Traditional gradient descent often gets stuck at saddle points and local minima as the gradient is zero, as depicted in *Figure* 1.12. Non-linear techniques developed in convex optimisation such as conjugate gradient descent and (Quasi-)Newton methods are powerful, yet rarely applied in practice due to their computational complexity. The most popular optimisers for neural networks incorporate the concept of momentum, where earlier gradients are considered in the descent. Adam [198] is one of the most recent optimisers and combines elements from two powerful optimisers before it, AdaGrad and RMSProp. Nesterov momentum [199] — which has favourable properties in convex optimisation — can also be incorporated into Adam, creating Nadam [200].

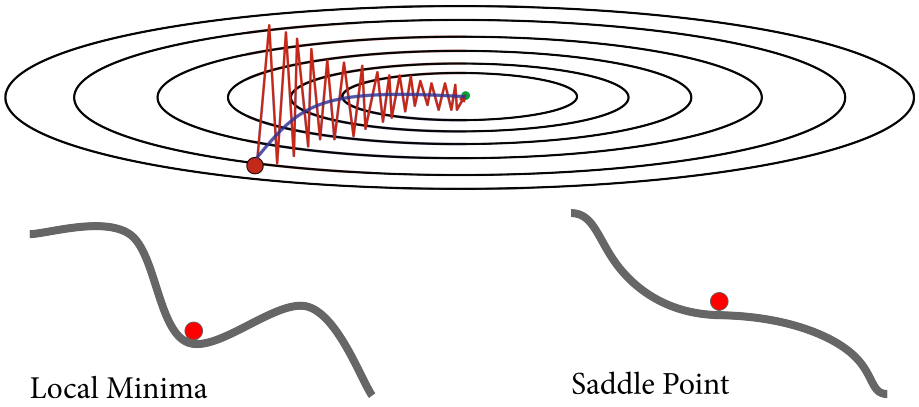

Figure 1.12: Traditional gradient descent (red curve) performs poorly in 'long valleys'. Optimisers generally use momentum to simulate the behaviour of the optimal blue curve and avoid local minima. Diagram adapted from Stanford's CS231n [201].

Each optimiser also has its own hyperparameters, the most major one being the learning rate. As training is stochastic, training multiple models and using them in an ensemble often results in better performance. Loshchilov and Hutter [202] proposed a novel approach where the learning rate is fluctuated during training to create multiple different models to ensemble in one training process.

1.3.3 Feature Selection and Dimensionality Reduction

The general approach to a machine learning problem is to extract as many features as possible then decide which are most relevant. Redundant or highly correlated features reduces the performance of most machine learning algorithms. Simple models like Naïve Bayes rely on the assumption that features are independent and correlated features can disproportionately weigh certain factors. Neural networks are better equipped to handle correlated and redundant features; however, may need more data and training time to do so.

***Highlight* 1.13.** Feature selection techniques aim to eliminate useless features and dimensionality reduction reduces the correlation between features.

Feature selection simplifies the model by selecting a subset of features to use. This increases the interpretability of a model, reduces the probability of cross-validation overfitting [21] and speeds up training. Selecting the optimal subset of features is not a simple task as some features may be uninformative on its own, but useful when combined with others. An computationally intractable exhaustive search would be required to determine the optimal subset. Feature selection algorithms aim to quickly find a good subset and can be categorised as filters, wrappers and embedded methods.

Filters evaluate subsets of features by maximising various criteria such as entropy, similarity and other statistical measures. Evaluation of subsets are fast and results are independent of machine learning model. However, a majority of filters are based on the assumption of linearity, and may not be suitable when complex relationships exist between features. *Wrappers* 'wrap' around existing models, using cross-validation to evaluate a feature subset. Features are therefore better tailored to each model; however, wrappers may be computationally prohibitive and also cater towards the model's tendency to overfit [203]. *Embedded* methods are based on models which inherently perform feature selection during their training, often from strong regularisation.

The performance of these approaches are highly domain-dependent. This thesis employs state of the art supervised feature selection algorithms as depicted in *Table* 1.7. We refer to Li et al. [204] for a thorough description and comparison of chosen techniques.

Table 1.7: Feature selection methods used in this thesis. Implemented with scikit-feature [204]

Filter	Wrapper	Embedded
ReliefF [205]	SVM/Gaussian Process	RFS [211]
Fisher score [206]	Forward/Backwards	ls_l21 [212]
CIFE [207]	Search	
JMI [207]		
ICAP [208]		
MIFS [207]		
MRMR [209]		
CFS [210]		

Rather than eliminating features, *dimensionality reduction* reduces the amount of information required to represent the set of features. This can de-correlate features and improve performance for simpler machine learning models. The two most common forms of dimensionality reduction are the unsupervised principal component analysis (PCA) and supervised linear discriminant analysis (LDA) [162]. Neural networks can also be used to reduce the dimensionality of data by training a network to predict the input, where layers progressively contain fewer nodes. These are termed autoencoder networks and can out-perform PCA; however, they are more difficult to analyse [213].

***Highlight* 1.14.** Feature selection is almost a requirement for small datasets, whereas dimensionality reduction is less commonly applied as it can obfuscate the model.

1.3.4 Model Evaluation and Handling Overfitting

The primary goal of machine learning is to train a model which will generalise well to new data. Accuracy over the entire dataset is evidently not a good metric, as a model which memorises the data (overfit) can appear to have perfect accuracy while failing to generalise to new data. Model selection and evaluation is the field in statistics which handles this.

Cross validation (*CV*) has become the de-facto standard in machine learning. Conceptually, CV is simple — the primary types used in machine learning are *leave-one-out* and *k-fold*. Let there be 100 data points in a dataset. In leave-one-out CV (LOO-CV), 99 data points are used to train a model, and 1 data point to test and evaluate the performance. This is repeated over each data point and the average result taken as the generalization accuracy. K-fold is similar; however, rather than using only one data point, the data is split into k groups, training on $k - 1$ and testing on 1 group. For example, 2-fold CV involves training on group 1 and testing on group 2 then training on group 2 and testing on group 1. CV reduces the risk of overfitting; however, repeated runs of CV over different models increases the probability of overfitting.

Highlight* 1.15 *(Model Selection). Good model selection techniques are becoming increasingly important in machine learning as improvements become more marginal, and are likely to arise from natural variance in an model [21].

In summary, we will be performing 10-fold CV with random stratification[28] repeated 10 times. This results in a set of 100 accuracy values after taking the mean accuracy of each fold of cv for each model. The same stratification sets are used, and Bayes factor [214] is used to test if the mean performance of one model is greater than the other. This decision will be justified in the following section with more background into model selection and hypothesis testing provided.

Model Selection and Hypothesis Testing

K-fold CV and LOO-CV are the de-facto standards in machine learning, and it is rare to look for alternatives. They provide a good estimate for generalisation error, are easy to implement and fast to evaluate. LOO-CV allows almost all the data to be used in training. When the data is clean (high signal-to-noise ratio) LOO-CV performs nearly unbiased estimations [215]. However, LOO-CV has been criticised for preferring models with a high variance and is less computationally feasible for large data sets [216]. Kohavi [216] instead recommends 10 fold CV in the general case. CV variations such as exhaustive and Monte-Carlo CV exist, but are not recommended by statistical literature [217, 215].

[28] *Stratification* involves ensuring there are an equal ratio of classes in each group. In this case, people with and without PD.

There are a number of catches when performing cross validation. Importantly, CV requires validation data to be independent from training data. In medical contexts, it is common to have multiple recordings from a single patient. Recordings from the same patient are likely to share similar attributes and cross-validating naïvely over the whole dataset can easily overfit [108]. Secondly, when performing hyperparameter optimisation the CV score is used as a metric. This introduces the risk of the best model hyperparameters fitting the validation sets by chance [21].

Overfitting on cross-validation is difficult to detect without additional data and is a major issue in small datasets. A common approach is to take a subset of data as the 'test' data which remains unseen in hyperparameter optimisation; however, this is infeasible when there is not enough data to create a test set large enough for results to be meaningful. Ng [21] proposes an algorithm to select from a number of competing hypothesis. Repeating k-fold CV with different division of folds can also reduce the likelihood of overfitting on CV by chance. Bouckaert [218] recommends 10 fold CV repeated 10 times after extensive empirical testing.

Accuracy is the most basic and intuitive measure of performance; however, it has been the subject of a number of criticisms. Firstly, it is susceptible to the false positive paradox[29], and may not be a good representation of a model's effectiveness in difficult tasks. Sensitivity, or recall, is a measure of the proportion of positive classes correctly identified and specificity measures the proportion of correctly identified negative examples. Precision is occasionally used instead of specificity in the machine learning community, measuring the proportion of correctly identified positive examples over all positive predictions. The F_1 score is the harmonic mean of sensitivity and precision and is an more effective measure of model performance when classes are unbalanced.

Secondly, accuracy does not take into consideration the confidence of a model's predictions. The area under the ROC (Receiver Operating Characteristics [219]) curve (*AUROC*) was proposed as a better alternative to accuracy. The ROC curve is created by plotting sensitivity and specificity at all confidence thresholds, and the area under ROC was believed to be a more robust and consistent measure of model performance [220]. However, recent empirical experiments have shown that AUROC favours particular models [221] and it has been criticised for being incoherent [222, 223]. Modifications to AUROC have been proposed [221, 223], but they are uncommon in practice. As a result, accuracy will be the primary performance measures utilised in this thesis as it is interpretable and independent

[29] The false positive paradox occurs when there is a very low incidence of a positive results in the target population. For example, when only 1 per cent of the population suffer from PD, a model which only predicts 'no PD' will be completely uninformative yet perform better than any model which predicts PD sometimes.

of model characteristics. AUROC is more robust when there is a large disparity between classes, and will also be recorded for comparison with prior works such as Neto et al. [108].

Table 1.8: A summary of common performance measures used in machine learning and statistics literature.

Confusion Matrix

	Pred True	Pred False
Real True	True Positive (TP)	False Negative (FN)
Real False	False Positive (FP)	True Negative (TN)

Measure Definitions

Sensitivity /Recall	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{FN+FP}$
Precision	$\frac{TP}{TP+FP}$
F_1 Score	$\frac{2TP}{2TP+FP-FN}$

Hypothesis tests are used to determine if the results obtained in experiments are *statistically significant*. After cross validation, a hypothesis test should be used to determine that the difference in results is not by chance. Paired t-tests are the traditional approach; however, they have been subject to a range of criticisms regarding replicability [224, 225]. Recently, the American Statistical Association has officially endorsed *Bayes factor* [214] (*BF*) as their preferred method of hypothesis testing [226]. Mass-replication studies have shown that almost a half of prior psychological research does not meet the criteria for strong evidence when Bayes factor is applied [227]. Their interpretation is depicted in *Figure* 1.9. The standard two tailed Cauchy distribution is used as a prior in this thesis [214].

Table 1.9: Interpretation of Bayes factors [228]. Bayes factors do not account for model validity and selection bias.

< 1	Evidence for the null hypothesis of $1/BF$
1 to 3	Insignificant evidence
3 to 10	Substantial evidence
10 to 30	Strong evidence
30 to 100	Very strong evidence
≥ 100	Decisive evidence

Consider three models: A which achieves 80 per cent accuracy, B at 81 per cent and C at 81 per cent. B is strictly dominant to A, correctly classifying all of A's successes and additionally some of its failures. C has a very different set of successes and failures to A. Our Bayesian approach to model selection distinguishes the two cases, with B obtaining

the more significant result than *C*, as it is more likely due to an improvement in the model rather than chance.

In statistics, there is no agreed upon method for model selection and evaluation. Penalization based evaluation[30] criteria such as Akaike/Bayesian/General Information Criterion [229, 230] and Minimum Description Length [231] are common model selection techniques. However, these are less suitable for machine learning as it is difficult to quantify the complexity of model such as neural networks. Cross validation is therefore the only feasible technique to compare completely different models.

Overall, it should be clear that results reported in machine learning are highly susceptible to overfitting and should be considered with a grain of salt. Unless standard datasets of an substantial size are developed, methodology rather than results should be the primary concern. Seemingly insignificant details such as stratifying on a per-person or a per-sample level can be the difference between an AUROC of 98 and 45 per cent [108].

Cross validation results should not be interpreted as real-world generalisability — especially in small datasets. Accuracy will undoubtedly increase in the future when more training data becomes available. Furthermore, there is no statistical technique which can reliably account for inherent bias in the dataset or flawed methodology.

[30] *Penalization* based model criteria are inspired by Occam's razor, preferring simple models over complex ones which as they are less likely to overfit. A penalty is attached to the number of free parameters in a model.

2 Our Work

Although there is a rich selection of prior work in PD diagnosis with machine learning, the lack of a standard dataset and methods limits the comparability of different studies. It is apparent that multiple sub-fields exist and research is often confined within its own sub-field. For example, the top papers in the Interspeech 2015 PD speech challenge [54] used methods independent of the dysphonia feature engineering previously done for PD. Research also rarely considers the results of works completed in challenges such Interspeech or Michael J. Fox Foundation Parkinson's data challenges [232]. It is common to find papers failing to cite prior work which performs the same experiments. A major goal of this thesis is to consolidate and distil the techniques applied in prior works in a coherent, sequential format.

***Highlight* 2.1.** Multiple sub-fields exist in PD literature and research is often isolated within a sub-field.

Although prior works have reported excellent results, it is difficult to determine if these results are caused by biases in the dataset or overfitting. Like any field based on empirical statistics, a publication bias exists [224], and there will exist irreproducible results [225]. *Section* 1.3.4 details measures such as significance testing to reliably evaluate and select models; however, their implementation is uncommon in machine learning literature.

The variation of results shines doubt on the replicability of the best performing papers. Arora et al. [19] achieves 98 per cent accuracy using smartphone IMU data from 20 participants. In contrast, Zhan et al. [22] uses all features in Arora et al., in addition to speech and tapping measures; however, only manages 71 per cent accuracy. Furthermore, state of the art techniques in motion mode recognition rarely achieves such results, despite motion mode recognition likely being the 'easier' task [106].

Neto et al. [108] shows that it is possible to "digitally fingerprint" people using accelerometer and voice data, and that failure to split data on a per-participant scale is subject to overfitting. The walk signal was very informative at classifying PD when split at the sample level (AUROC $\approx$ 98 per cent) and uninformative when samples were split at a participant level (AUROC $\approx$ 45 per cent).

2.1 The mPower Dataset

To minimise the likelihood of bias or overfitting, a larger dataset was required. Currently, the only publicly available dataset that satisfies this requirement is mPower [105].

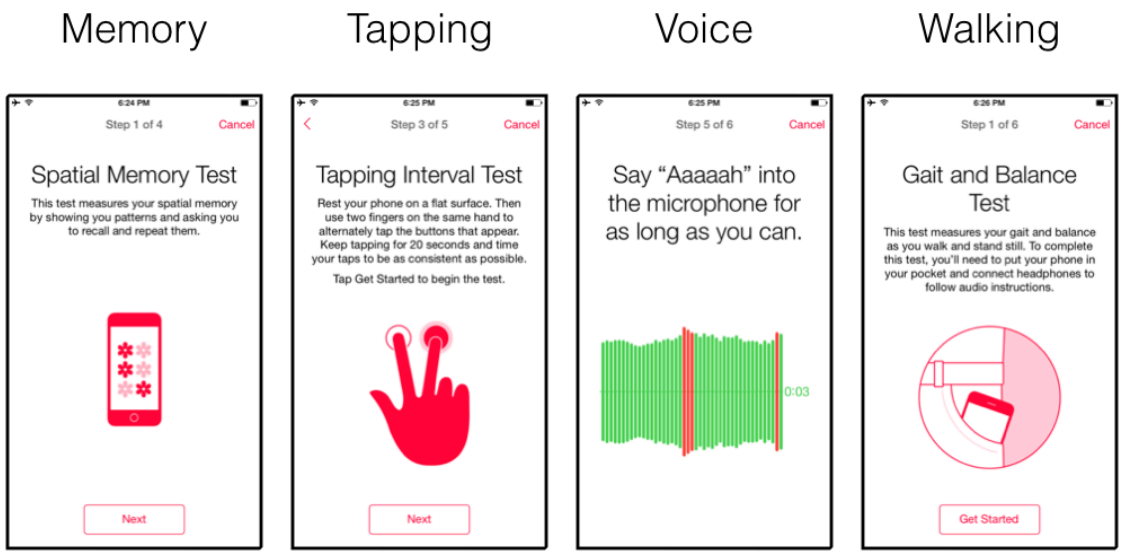

Figure 2.1: The mPower app consists of several tasks to evaluate memory, bradykinesia, voice and gait.

The mPower study began in March 2015, open to people living in the United States who owned an Apple iPhone or iPod released in 2011 or later. Upon downloading the app, the user is presented with the tasks depicted in *Figure* 2.1, along with general demographics and UPDRS questions. Each task/questionnaire was optional and could be completed multiple times. As of writing, there are around 6,500 participants in the study, 1,100 with PD. Users come from diverse backgrounds and may have other illnesses. The mPower dataset also contains a number of cases of young-onset Parkinson's disease[1], which have rarely been studied in a diagnosis context [233, 234]. Age is a bias in the dataset as a majority of the non-PD participants in the study were young adults.

The dataset was released late 2015 and has been used in a few studies. The most significant being the recently published Neto et al. [108], which provides extensive analysis on the data with respect to medication state and time of the day[2]. A major issue with mPower is that the data is 'noisy' — a common problem in any crowdsourcing project without significant precautions [235].

[1] Assuming the participants are honest of their circumstances.

[2] Neto et al. [108] was published in Jun 29 on arXiv and was not discovered until a late stage literature re-review. Unfortunately techniques introduced in Neto such as detangling medication states and "time of the day" effects were not explored due to time limitations.

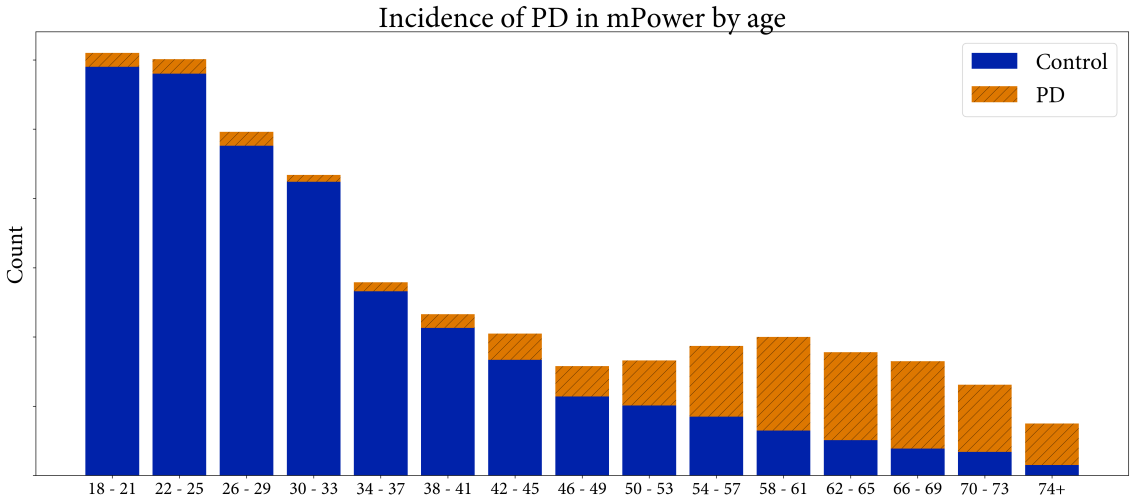

Figure 2.2: Age is a bias in the mPower dataset as most non-PD participants are young. There are also some cases of rare young-onset PD[1].

Preprocessing and Feature Selection.

Vowel phonation was captured with the single channel iPhone/iPod microphone at 44,100 Hz. Initial investigation showed that a substantial number of participants did not complete the task to an acceptable standard. Excessive levels of environmental noise, hesitation, interruptions or pronouncing vowels other than /aa/ were common. Additionally, the distance between the phone and user varies, with some participants speaking directly into the microphone, creating a large amount of 'wind noise'. Preprocessing techniques reduce these interferences exist, but have not been applied to avoid introducing bias to the data [84, 55, 236].

At the time of writing, there were 65,000 phonation samples from 6,000 participants in the mPower dataset (a majority of these from a small number of users). We evaluated 1,600 randomly selected samples for performing the task correctly, rejecting around 25 per cent. Simple metrics such as variance in short time energy and noise prior to recording were used in hand-crafted rules to rank and filter the phonation samples. After filtering, 4,100 users remained, 900 with PD. The highest ranked sample was selected for each of the users[3]. Upon reflection, a valid recording should have been randomly sampled from each participant to avoid the bias of PD participants performing more recordings. Machine learning could optimise this process; however, was avoided due to the possibility of introducing bias to the data.

The *walking* task involves the participant putting their phone in the pocket or bag, walking 20 steps then standing still (balancing) for 30 seconds. During this task, accelerometer and gyroscope is continually sampled at 95 ± 7 Hz. Basic features such as the cadence were

[3] Optimally, all samples should be used to improve robustness, but available processing power was limited.

extracted; however, mPower does not record the parameters necessary (such as leg length) to estimate step length [104]. The characteristics of the signal in-pocket and in-bag would vary significantly, with female users predisposed to having the phone in their bag [237]. For future data collection projects, phone in hand may be preferable as it introduces fewer variables and does not cause information loss, unlike phone in bag.

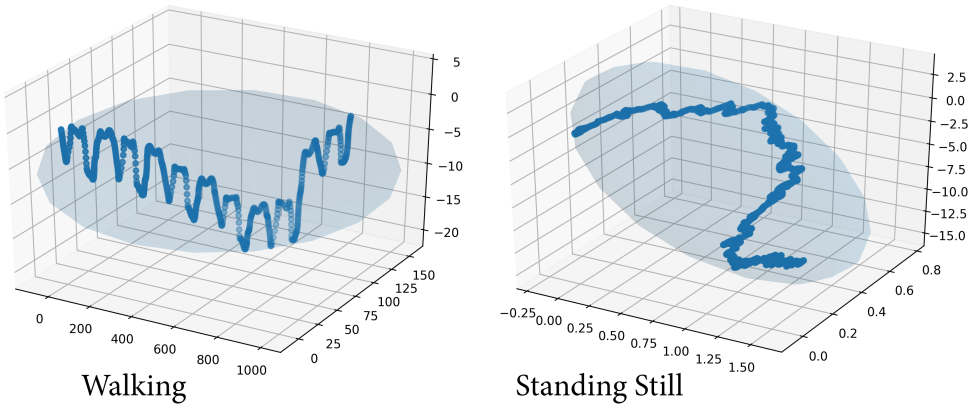

Figure 2.3: A visualisation of position from the raw device accelerometer during walking and balancing. The participant appears to be walking on a slope.

As the device is in the user's pocket or bag, data from the gyroscope would be minimally informative. A rotation matrix was computed with the gyroscope data to align the accelerometer's z axis to the direction of gravity. Another rotation was applied to align the average direction of motion with the x axis.

The balancing task aims to simulate the behaviour of standing on a force plate. Unlike force plate experiments, there is no central reference and the subject was not instructed to stand as still as possible. Many participants showed a significant amount of sway which could be consciously preventable. To map the accelerometer to force plate data, a 10^{th} order zero-phase 1hz Butterworth highpass filter was applied (*Figure* 2.4). The highpass filter removed preventable sway at the cost of valuables information below 1hz [238]. Additionally, the Anteroposterior and Mediolateral directions of sway were lost with this preprocessing.

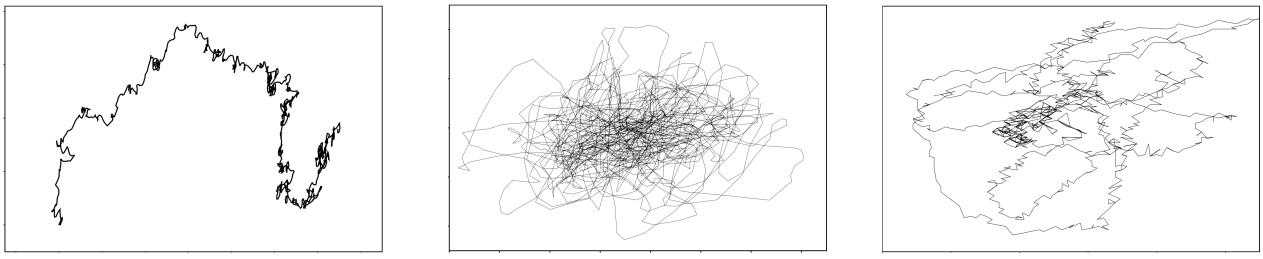

Figure 2.4: The Butterworth filter results in a device path more similar to the centre of pressure; however, low frequency sway information is lost. Note that the device motion recording is 30 seconds long while the force plate is 10 seconds.

A 16 second extract of balance recording between 4s and 20s and the first 10 seconds of the walking task was used for each subject. The choice of these values was solely informed by the nature of the dataset, with the first four seconds of balance data containing significant movement and most recordings of variable length.

Samples which are too short, are corrupted, or have too much variability in recording rate were removed from the set. Approximately 2,300 participants remained with both a valid phonation and movement recording, 600 with PD. The most recent valid sample was selected for each participant, and features specified in *Section* 1.2.5 calculated using the tools and techniques described in *Section* 2.8. Feature engineering was done on both the original and filtered data for the balancing task. Features were computed over the (x, y) and (x, y, z) dimensions for the balancing and walking task respectively.

***Highlight* 2.2.** One recording of each task per user is selected for machine learning. After preprocessing, the phonation task consists of 4,100 users, 900 with PD, and the movement task consists of 2,300 users, 600 with PD.

2.2 Replicating Past Work

The work of Tsanas et al. [18] on diagnosis with vowel phonation, and Arora et al. [19] on smartphone accelerometer data was replicated. SVM results are reported as they consistently performed better than random forests and Gaussian processes.

2.2.1 Vowel Phonation

Tsanas et al. [18] used the National Center for Voice and Speech (NCVS) dataset, which consisted of 33 people with PD, and 10 healthy controls. In total, 263 phonations were recorded in controlled circumstances using a professional grade microphone. HNR, GQ, RPDE, DFA, PPE, GNE, VFER, MFCC and variants of shimmer and jitter were calculated, resulting in a set of 132 features (*Section* 1.2.5).

Figure 2.5: Cross-validation results of Tsanas [18] with a SVM classifier after feature selection. Results reported as mean accuracy ± std accuracy.

LASSO	mRMR	RELIEF	LLBFS
$VFER_{NSR,TKEO}$	2^{nd} MFCC coef	1^{st} MFCC coef	2^{nd} MFCC coef
11^{th} MFCC coef	$Shimmer_{Amplitude, AM}$	11^{th} MFCC coef	11^{th} MFCC coef
$VFER_{NSR,SEO}$	$VFER_{NSR,SEO}$	2^{nd} MFCC coef	9^{th} MFCC coef
4^{th} delta MFCC	$GNE_{NSR,SEO}$	3^{rd} MFCC coef	$VFER_{NSR,TKEO}$
HNR_{mean}	5^{th} delta-delta MFCC	$VFER_{NSR,TKEO}$	$VFER_{entropy}$
GNE_{std}	HNR_{mean}	$VFER_{NSR,SEO}$	$VFER_{NSR,SEO}$
12^{th} MFCC coef	8^{th} MFCC coef	9^{th} MFCC coef	RPDE
RPDE	4^{th} delta MFCC	7^{th} MFCC coef	HNR_{mean}
$OQ_{std\ cycle\ open}$	11^{th} MFCC coef	6^{th} MFCC coef	DFA
2^{nd} MFCC coef	$VFER_{NSR,TKEO}$	8^{th} MFCC coef	4^{th} delta MFCC
94.4 ± 4.4 TP: 97.5 ± 3.4 TN: 86.5 ± 14.3	94.1 ± 3.9 TP: 97.6 ± 3.3 TN: 84.3 ± 13.2	98.6 ± 2.1 TP: 99.2 ± 1.8 TN: 95.1 ± 8.4	97.1 ± 3.7 TP: 99.7 ± 1.7 TN: 89.1 ±13.9

Features were calculated on the 263 phonations and 100 times repeated 10 fold cross validation used to evaluate models. It is unclear whether Tsanas et al. has split the phonations on a per-subject scale — failure to do so substantially increases risk of overfitting, as

phonations from the same subject may appear in both the training and validation set. Random Forests and SVMs were evaluated with hyperparameters selected by gridsearch [172]. As data is limited, feature selection with four common algorithms was performed to improve results. This results in the 10 feature subsets depicted in *Figure* 2.5.

Highlight 2.3. It is unclear whether Tsanas has split the phonations on a per-subject scale and failure to do so presents a substantial risk of overfitting.

We replicated Tsanas on the 3,200PD and 900C phonations selected after preprocessing mPower (see *Section* 2.1). Features were extracted from a 1.5 second window of each audio sample, which is similar to the phonation length used in fundamental frequency estimation datasets [146]. Gridsearch was performed to find (near) optimal SVM hyperparameters. The best performing 10 feature subset in Tsanas et al. (ReliefF) is initially evaluated.

Highlight 2.4 (Hyperparameters). Unless otherwise specified, search over hyperparameter values were performed for all SVMs. The RBF kernel was always the most effective. Probabilistic output SVMs [164] were used for a more accurate AUROC score.

Note the NCVS data used in Tsanas et al. was at a ratio of 33PD:10C, whereas the mPower data is at a ratio of 9PD:32C. We stratified the data by random sampling to simulate NCVS split. On both the NCVS and mPower ratio, the SVM classifier exhibited the false positive paradox — where the most common class is predicted for all inputs. As evident in *Table* 2.1, the replicated results are significantly poorer than the reported results.

Table 2.1: Cross validation results of a SVM using Tsanas' 10 feature ReliefF subset (fig 2.5). Presented as mean ± standard deviation.

Equal stratification (50P:50C)

	Pred PD	Pred C
True PD	28.2 ± 2.6%	21.8 ± 2.6%
True C	13.5 ± 2.4%	36.5 ± 2.4%
Accuracy		64.7 ± 3.0%
Sensitivity		56.4 ± 4.7%
Specificity		73.0 ± 4.4%
AUROC		69.4 ± 4.8%

mPower stratification (9P:32C)

	Pred PD	Pred C
True PD	0 ± 0%	79.2 ± 0%
True C	0 ± 0%	21.8 ± 0%
Accuracy		79.2 ± 0%
Sensitivity		0 ± 0%
Specificity		100 ± 0%
AUROC		72.1 ± 4.8%

The ReliefF [205] feature subset in Tsanas et al. consisted primarily of MFCC coefficients. MFCC is often the primary feature in speech recognition systems; however, the high and low coefficients are known to be rarely informative in speech recognition [239] — the ReliefF feature set contains both the 1^{st} and 11^{th} coefficients. The result suggest that high and low coefficients may be informative when used to detect abnormal speech. MFCC are known for being very sensitive to noise and frequency [95, 240] and another explanation is the data used in Tsanas et al. consisting of controlled recordings, whereas mPower is noisy.

Overfitting is also a possibility. It is ambiguous if Tsanas et al. divided phonations of a per-subject level in cross validation. Naïve CV may result in phonations from same individuals appearing in both the training and validation sets. As MFCCs are sensitive to minor changes in frequency [240], phonations from different individuals may be easily separable in the MFCC space. This is also supported by the disparity of results between the random forest and SVM classifiers on all features (90.2 per cent vs. 97.7 per cent), as the hyperparameters of the RF classifier were not tuned by cross validation and RF is more robust against overfitting.

Table 2.2: Cross validation results of a SVM using all speech features presented in Tsanas et al. [18]. Decisively outperforms 2.1 with a Bayes factor of 10^{17}.

Equal stratification (50P:50C)

	Pred PD	Pred C
True PD	31.5 ± 2.4%	18.5 ± 2.4%
True C	12.2 ± 2.1%	37.8 ± 2.1%
Accuracy		69.3±3.3%
Sensitivity		62.9±4.8%
Specificity		75.6±4.2%
AUROC		75.7 ± 3.4%

mPower stratification (9P:32C)

	Pred PD	Pred C
True PD	3.9 ± 0.8%	16.9 ± 0.8%
True C	1.9 ± 0.6%	77.3 ± 0.6%
Accuracy		81.2 ± 1.1%
Sensitivity		18.7 ± 4.2%
Specificity		97.6 ± 0.9%
AUROC		76.7 ± 2.9%

In our testing, using all measures presented in Tsanas et al. results in improvements over any of the 10 feature subsets presented in *Figure* 2.5. *Table* 2.2 shows the improvement — it is barely enough to escape the false positive paradox.

2.2.2 Movement

Arora et al. [19] used smartphone accelerometer data to distinguish 10 healthy and 10 PD participants. Participants were given a LG Optimus S smartphone and instructed to walk 20 steps, turn around, walk 20 steps then stand upright for 30 seconds. The position of the device was not specified — it is most likely in the participant's pocket. No preprocessing was done to the data, and features engineering included simple statistical and entropy measures, DFA, mean TKEO and the dominant frequency. Arora et al. obtained 98 per cent accuracy on with 100 times repeated 10-fold cross validation. Zhan et al. [22] extended the feature set and performed a similar experiment on 121 PD and 105 control using additional voice and tapping features; however, only achieved 71 per cent accuracy.

This thesis also coincided with the PD Digital Biomarker DREAM Challenge which involved using accelerometer data to classify PD or predict the UPDRS motor score [241]. Sage Bionetworks, sponsor of the mPower dataset and a organiser for the challenge released a baseline feature set [242] which included all features in Arora et al. [19], with additional

jerk based measures and the peak of the Lomb-Scargle periodogram [243]. The work completed as part of this thesis was submitted to the challenge. The challenge evaluated the trained models on an unreleased portion of the mPower data, validating that our models have not overfit.

We evaluate both the features in Arora and the DREAM challenge baseline on the mPower accelerometer data. The DREAM baseline is expected to outperform Arora as it is a superset of the features used in Arora. The DREAM features are assumedly an extension of the original code in Arora et al. [19] (based on the organiser's affiliations), whereas the Arora feature set is calculated with our own implementation. The models were evaluated on a 50PD:50C stratification of the data to better convey performance.

Table 2.3: Cross validation results of a SVM on a 50PD:50C stratification of the mPower accelerometer data. Bayes factor of 1.2 — there is insufficient evidence to conclude the DREAM set outperforms Arora.

Arora Features

	Pred PD	Pred C
True PD	30.0 ± 3.2%	20.0 ± 3.2%
True C	17.5 ± 2.9%	32.5 ± 2.9%

Accuracy	62.5±4.2%
Sensitivity	60.0±6.4%
Specificity	65.0±5.9%
AUROC	72.1 ± 4.8%

DREAM baseline features

	Pred PD	Pred C
True PD	31.3 ± 2.7%	18.7 ± 2.7%
True C	17.5 ± 2.8%	32.5 ± 2.8%

Accuracy	63.8 ± 3.9%
Sensitivity (TP)	62.6 ± 5.4%
Specificity	65.0 ± 5.7%
AUROC (TN)	72.1 ± 4.8%

The results in *Table* 2.3 fall far from those reported in Arora et al. [19], and mirrors the findings of Zhan et al. [22]. It is clear that performing machine learning in small datasets is a perilous task, with reported results at high risk of overfitting on cross validation.

Highlight 2.5. Neither Tsanas' or Arora's results could be replicated on the mPower dataset.

The mediocre performance is not unexpected — it is unlikely that simple statistical and entropy measures are sufficient to accurately distinguish PD from control. There is a significant amount of noise from variables such as whether the phone was in a pocket or bag and large variance in human gait. This noise is difficult to handle with simple statistical and entropy measures which are sensitive to minor changes in the signal. More advanced signal processing techniques which may be more resistant to the noise will be required.

2.3 Novel Features for PD Diagnosis

The features used in current PD literature are insufficient to accurately perform a diagnosis — especially given the natural variance of biological signals. EEG signal processing is a field that faces similar issues, with the characteristics of an EEG signal difficult to define. Machine learning with EEG data has evolved to rely primarily on non-linear signal processing techniques. These are detailed in *Section* 1.2.4 — the relation of some of these measures to PD symptoms will be summarised below.

Non-linear signal processing involves the estimation more abstract characteristics of a signal. Previously, the nonlinear methods DFA, RPDE and PPE have been applied to phonation [99, 25]. DFA and RPDE are measurements quantifying the autocorrelation of a signal, which is expected to be lower in dysphonic phonation due to the variation introduced by turbulent airflow around the incomplete glottal closure. These features may also be applicable to accelerometer data, with the increased tremor and jerking motion from cogwheel rigidity [244] resulting in a less stable signal. PPE is a measure of the stability of the fundamental frequency in speech and is less applicable to accelerometer data.

A wider range of non-linear signal processing has been applied on EEG signals. The largest Lyapunov exponent is a measure of the chaos in a system, which is likely higher in dysphonic speech and dyskinetic movement. The fractal dimension relates to the complexity of a signal, and Fisher information and sample entropy measure the unpredictability of a signal. These are likely to be higher in dysphonic speech and dyskinetic movement due to added information from turbulent airflow or tremor and jerking movements.

***Highlight* 2.6.** The novel features we introduce do not directly relate to human senses and may be more promising at detecting subtle symptoms missed by neurologists.

We believe these novel features are more robust measures of symptoms. These features do not directly relate to the information captured by human senses, and may be applicable in detecting symptoms unnoticeable by an expert. We repeated earlier accelerometer and speech experiments on the mPower dataset using the novel features introduced in *Section* 1.2.4, with the results recorded in *Table* 2.4.

The novel features are stronger differentiators of PD than features used in current literature, but not sufficiently to robustly classify PD and control. Bayes Factor analysis shows novel features alone for accelerometer performs better than the DREAM baseline with a $BF = 5$. This value is substantial in statistics, but in cross validation implies that the classifiers are succeeding on fairly different examples. Combining both novel features and Tsanas' feature set improves performance greatly on the speech data.

Table 2.4: Cross validation results for a SVM when including the non-linear features. Adding the novel features improves model performance on the accelerometer and phonation data with a Bayes factor of 10^7 and 10^{12} respectively.

Accelerometer (50PD:50C stratification)

	DREAM	Novel Only	Both
Accuracy [%]	63.8 ± 3.9	66.0 ± 4.0	68.8 ± 3.7
Sensitivity [%]	62.6 ± 5.4	61.8 ± 5.7	67.2 ± 5.3
Specificity [%]	65.0 ± 5.7	70.2 ± 1.9	70.5 ± 5.4
AUROC [%]	72.1 ± 4.8	71.3 ± 4.5	74.8 ± 4.1

Phonation (50P:50C stratification)

	Tsanas	Novel Only	Both
Accuracy [%]	69.3 ± 3.3	62.9 ± 3.5	72.2 ± 3.1
Sensitivity [%]	62.9 ± 4.8	48.5 ± 5.0	69.7 ± 4.3
Specificity [%]	75.6 ± 4.2	77.2 ± 4.9	74.8 ± 4.7
AUROC [%]	75.7 ± 3.4	66.6 ± 3.9	78.3 ± 2.9

Highlight 2.7 (Novel Features). The novel features are stronger differentiators of PD; however, there is still more to be desired in terms of performance.

2.4 Visualising The Features

In this section, we explore and visualise why models replicated earlier perform well below the reported level in prior works. This disparity is possibly caused by differences in dataset quality — particularly in ensuring that the task is performed correctly or consistently. It is equally likely that introducing a greater diversity of subjects increases the problem difficulty due to the natural variances in speech and gait. No individual feature could achieve greater than 60% classification accuracy on an equal stratification of the data. Note that only a fraction of the data is visualised to improve clarity — more formal analysis of the features is performed in *Section* 2.7.1

Highlight 2.8. No single feature was able to differentiate Parkinson's disease with a greater than 60% accuracy.

2.4.1 Speech

Many dysphonia features rely on precise measurements of the length of each glottal cycle. The SWIPE [245] fundamental frequency estimation algorithm was used to obtain these measurements. Most f_0 algorithms are sensitive to changes and noise in the signal, and are not yet suitable to handle the noisy mPower data. Issues with f_0 extraction would invalidate the measurements of a number of features. A simple investigation shows that the standard deviation of f_0 (*Figure* 2.6) exceeds 10Hz for 347 subjects, a value that indicates a failure of the algorithm or a poorly executed recording.

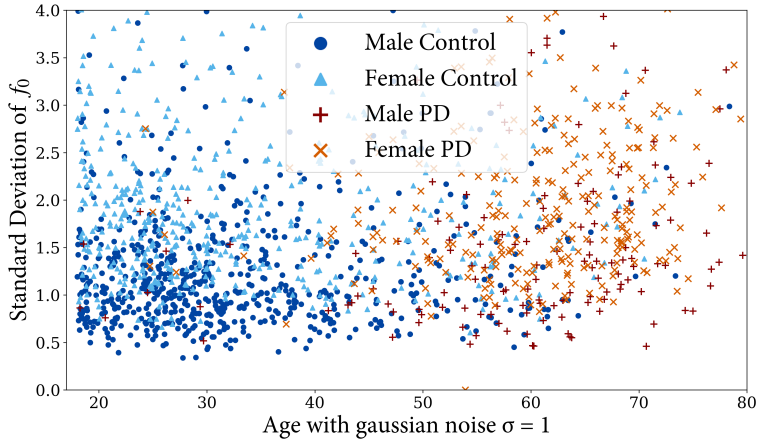

Figure 2.6: Standard Deviation of the fundamental frequency during phonation. Females and older individuals exhibit notably higher variations in f_0, whereas the distinction between age matched PD and control participants is less clear.

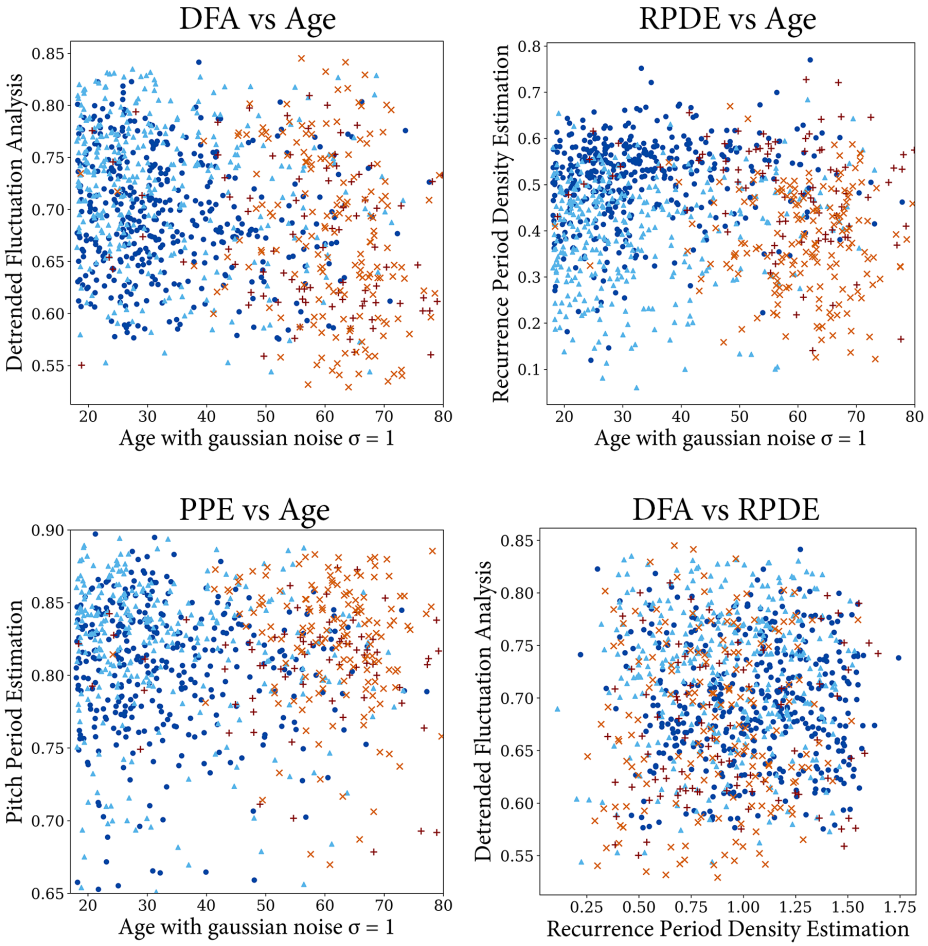

Figure 2.7: PD subjects exhibit a lower DFA and RPDE, and a higher PPE. However, these features are hardly separable. A combination of DFA and RPDE significantly enhances the separability of female participants with PD. See *Figure* 2.6 for legend.

Little et al. [99, 25] introduced three measures to distinguish dysphonia — DFA, RPDE and PPE. DFA and RPDE are measures of the autocorrelation of a signal. As evident in *Figure* 2.7, people with PD exhibited a lower autocorrelation than age matched control

subjects, indicative of a more chaotic and variable speech signal. People with PD also show an increase in PPE to age matched control, which is evidence of fluctuations in pitch above healthy speech production. However, distinguishing dysphonic speech is not easy due to the natural variance in speech production.

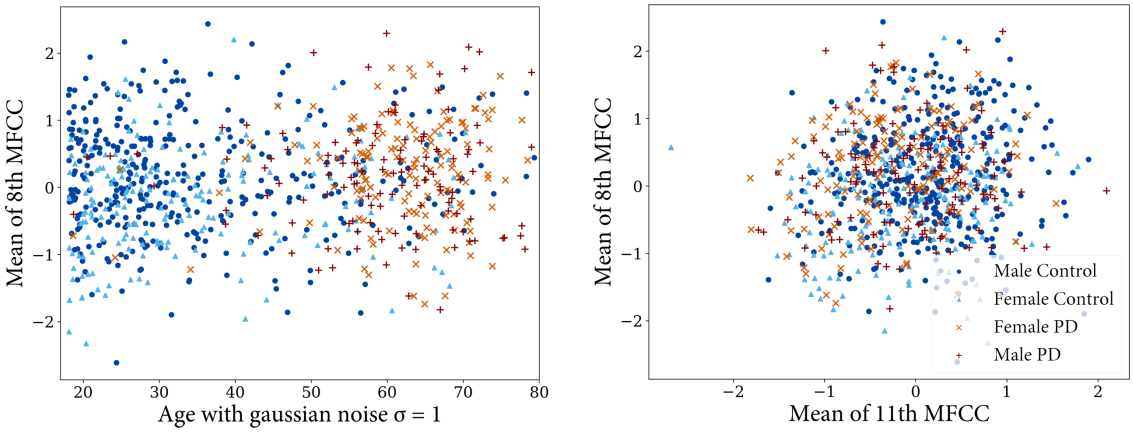

Figure 2.8: Unlike Benba [246], the 8th MFCC coefficient is not a notable for distinguishing dysphonia. A combination of multiple MFCC adds marginally more information.

Benba et al. [246] distinguished PD with a 82 per cent success rate primarily with MFCC, which are also present in most of the feature subsets derived by Tsanas et al. [18]. Being the primary feature in most speech recognition systems, it is not surprising MFCC are strong features in detecting dysphonia. However, *Figure* 2.8 suggests that there is a very minimal correlation between MFCC and PD. The MFCC are very sensitive to changes in the signal, and more advanced measures than the mean and standard deviation may be required to fully utilise them. An example is using a LSTM neural network, which is common in speech recognition [247].

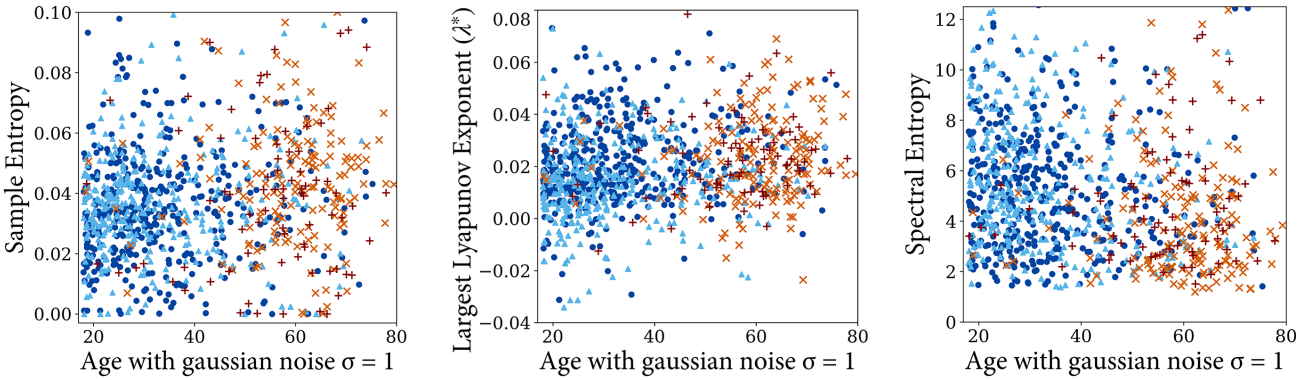

Figure 2.9: A visualisation of some nonlinear speech features. Sample entropy and λ^* (measures of unpredictability) are higher in subjects with PD, whereas spectral entropy is lower. See *Figure* 2.8 for legend.

The novel features introduced in *Section* 1.2.4 are strong differentiators of PD. As expected, the sample entropy and largest Lyapunov exponent (λ^*) are greater in participants

with Parkinson's Disease, indicative of a more chaotic/unpredictable signal. The lower spectral entropy is indicative of a more even distribution of the frequencies, which can be visualised in the 'blurry' nature of the spectrogram in *Figure 1.1*.

The largest Lyapunov exponent was mostly positive during the phonation of /aa/, with female participants with PD in particular exhibiting a greater λ^*. This is also a significant result for literature analysing the non-linearity of speech with λ^* as it was conducted on a very large dataset. This provides evidence for results indicating a slightly positive λ^* in vowel phonation [118, 248] rather than close to zero [117, 249].

2.4.2 Movement

Understanding the features extracted for the movement data is a trickier task. Like speech, a significant amount of variance will exist based on physical factors (such as leg length) and environmental factors (such as depth/tightness of pocket). The mPower application asked the user to place their phone in either their pocket or bag, and there likely exists a tendency for females to complete the activity with their phone in a bag [237]. This choice of experimental design may limit machine learning potential, as information will be lost from the padding and pendulum motion of shoulder bags and padding. A better setup would be to ask participants to hold the device in their hand — although additional variability is introduced, this could be filtered whereas lost information cannot be recovered.

Accelerometer signals contain three dimensions of data (x, y, z). Rotations have been performed such that the z axis is aligned with gravity, and the average direction of the walk is aligned with the x axis. Features are visualised as the mean value over all dimensions.

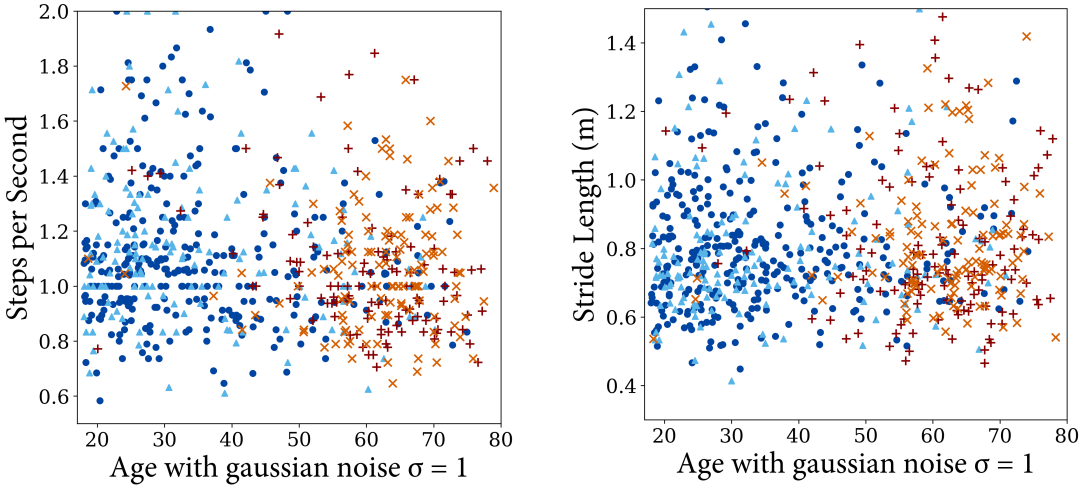

Figure 2.10: The average cadence and stride length of participants with PD is only marginally lower than age matched control.

Basic walking features were calculated from the provided pedometer data, likely extracted from the implementation in Apple ResearchKit [105]. The estimation of steps and

especially distance may not be accurate, accounting for additional variance in results. This is visualised in *Figure* 2.10 and it is clear that these simple features are barely informative in diagnosis.

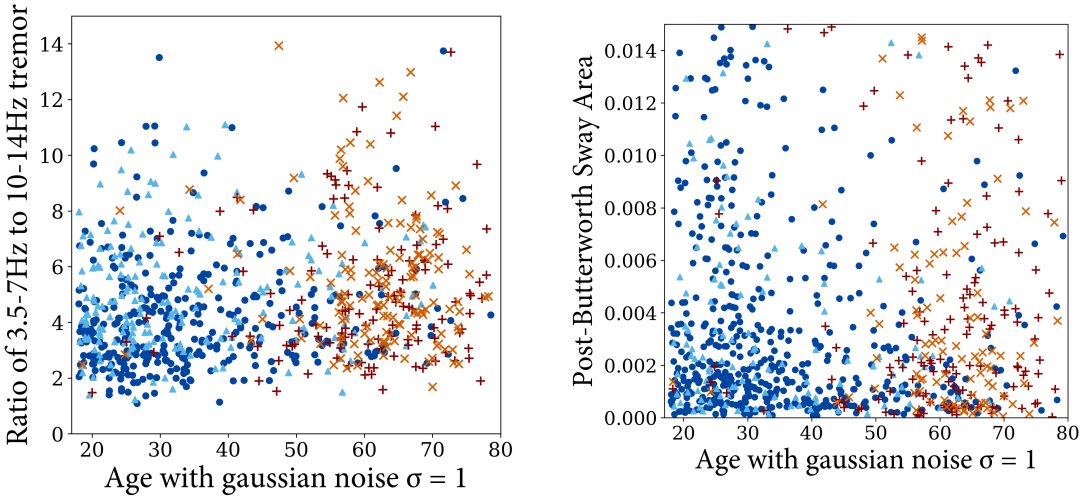

Figure 2.11: PD participants exhibit a definite increase in 3.5-7Hz tremor. The variance of sway area is high, and PD participants only exhibit a minor increase in sway area compared to age matched control.

The most common features used diagnosis of force plate and IMU data are tremor and postural sway. These are depicted in *Figure* 2.11. Visualising the ratio of tremor in the 3.5–7 Hz bands (most commonly associated with PD) and 10–14 Hz bands, there is a definite increase in Parkinsonian tremor. Post-Butterworth sway area (by bounding ellipse) is less informative and suffers from high variance.

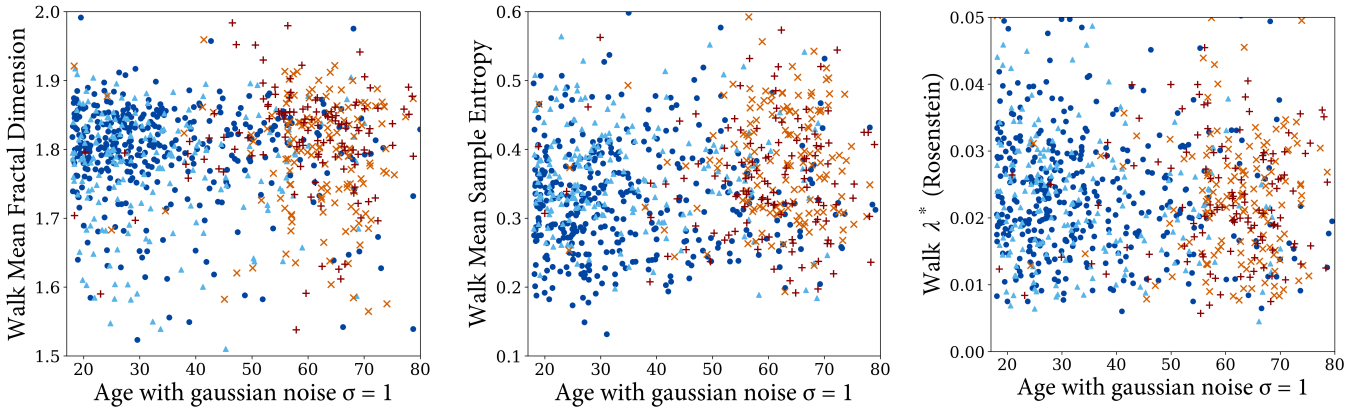

Figure 2.12: The Higuchi fractal dimension and sample entropy are strong differentiators of PD, whereas λ^* is uninformative.

The novel features introduced in *Section* 1.2.4 are more suited at quantifying Parkinsonian gait. Three of them are visualised in *Figure* 2.12. Fractal dimension and sample entropy are measures of the detail and complexity of a signal, and tend to be higher participants

with PD — especially females. This was a surprising result as we believed that the use of handbags would hinder the differentiability of PD.

As each gait cycle is significantly longer than a glottal cycle (speech), we would expect lower predictability in the movement task. The largest Lyapunov exponent is always positive; however, we find no statistically significant correlation between λ^* and PD on both balance and walk data. This result contrasts with Howcraft et al. [120] which showed a lower λ^* in people with diabetic neuropathy (NP). This may suggest a possibility distinguish characteristics of PD and NP gait, but could also be attributed to the low sample rate of phone accelerometers which is unsuitable for estimating the Lyapunov Exponents [114].

2.4.3 Conclusions and Recommendations

The wide variation in natural speech production makes it difficult to distinguish dysphonic speech. Many features are not invariant to the fundamental frequency of the speaker, as evident in differences between the male and female groups. Normalising these features with respect to f_0 may improve the applicability of these features.

Many features are sensitive to minor fluctuations in the signal and their value can change dramatically depending on the segment used, as will be shown in *Section* 2.5. The large fluctuations in value reduces their effectiveness as simple predictors; however, they may be valuable when the feature is calculated over short-time windows. Models such as CNNs or LSTMs (*Section* 1.3.2) can utilise the additional short-time temporal information to make more robust predictions. This would also solve the issue of some features being dependent on data length. Unfortunately this will not be explored due to computational limitations.

***Highlight* 2.9.** Making sense of each individual feature is a challenging task. Machine learning models are required to interpret the uncertainty of these features to obtain a reliable result.

As no individual feature can achieve good classification accuracy, machine learning models must combine multiple features to reach a reasonable level of performance. Traditional machine learning models are best suited to modelling features which are independent. They often perform worse when two features are correlated, despite there being additional information. Feature selection [204] can be used in traditional machine learning models to reduce the impact of this. Models such as neural network are able to model non-linear relationships between features; however, training data may not be plentiful enough. In the following section, we will explore typical approaches to improving model performance in machine learning.

2.5 Improving Performance

Three standard machine learning techniques for improving performance will be investigated: data augmentation, feature selection, and ensemble models. These techniques have been shallowly applied to demonstrate their effectiveness — additional computational resources is required to investigate the limits of these techniques. Data augmentation is applied to the phonation task, and feature selection and ensembling applied to the movement task.

2.5.1 Data Augmentation for Phonation

An quick examination of the features used for the phonation data showed that many were unstable, with around 10 per cent of them varying by at least 0.5 standard deviations when computed over the same audio signal offset by different amounts. It was also clear that some features were not length independent, though resolving the length-invariance of features will not explored in this thesis.

We extracted seven 1.5 second sample from the recording, starting at the 1.5 to 4.5 second mark with a 0.5 second step size. The applicability of three simple methods of utilising the extra information is explored in *Table* 2.5. *Augmentation* involves using the additional samples as extra cross validation data. *Merging* combined the 4 $\mathbb{R}^d$ vectors of features into a $\mathbb{R}^{4d}$ vector of features. Finally, *Meanify* involves taking the mean of the features computed over the 4 samples. Note that this differs from 'collapsing' by median in Neto et al. [108] as the number of samples is fixed so bias from PD participants performing more recordings is avoided. Ensemble models are alternative techniques of utilising the extra data, and will be explored in *Section* 2.5.2.

Table 2.5: Cross validation results on full vocal feature set using a SVM on a 50PD:50C stratification of the data with various data augmentation techniques.

	Original	Augmented	Merged	Mean
Accuracy [%]	72.2 ± 3.1	76.9 ± 1.0	70.3 ± 2.8	73.3 ± 3.0
Sensitivity [%]	69.7 ± 4.3	73.5 ± 1.5	69.6 ± 4.7	69.9 ± 4.9
Specificity [%]	74.8 ± 4.7	80.3 ± 1.4	71.1 ± 4.5	76.9 ± 4.3
AUROC [%]	78.3 ± 2.9	84.0 ± 1.1	75.9 ± 0.3	80.4 ± 2.7

Augmentation and mean exceed the performance of the original feature set at a Bayes factor of 10^{23} and 10^3 respectively. As expected, merging does not perform well with a SVM and shows the applicability of feature selection — even in models as robust as SVMs. Data augmentation strategies evidently provide a viable solution to overcoming the low data quality in the mPower dataset.

***Highlight* 2.10.** Data augmentation was extremely effective, implying that performance can be significantly improved with more data.

Recurrent and convolutional networks can utilise additional temporal and variance information to improve predictions. However, the neural network architectures we applied achieved similar performance to the augmented SVM. Likely, more than seven segments are required to fully utilise the benefits of these models.

2.5.2 Feature Selection and Ensembles for Accelerometer

There are 431 numerical features used to classify accelerometer data, and many are likely redundant. Most signal processing measures are only well defined for one dimensional signals and the accelerometer provides three (x, y, z) dimensions of data. Although information is gained from computing measures over all dimensions, it is difficult for most machine learning models to utilise this information. It is likely that the additional variability in the axis, and the correlated nature of the data will reduce the performance of most classifiers. The effect of taking the Euclidian norm on classification accuracy is depicted in *Table* 2.6.

Table 2.6: Results of taking the norm for features extracted over (x, y, z) dimensions with a SVM. The normed features perform worse than the full feature set ($BF = 41$).

	DREAM	Novel Only	Both (full)	Both (normed)
Accuracy [%]	63.8 ± 3.9	66.0 ± 4.0	68.8 ± 3.7	67.7 ± 4.0
Sensitivity [%]	62.6 ± 5.4	61.8 ± 5.7	67.2 ± 5.3	65.4 ± 6.1
Specificity [%]	65.0 ± 5.7	70.2 ± 1.9	70.5 ± 5.4	70.0 ± 5.2
AUROC [%]	72.1 ± 4.8	71.3 ± 4.5	74.8 ± 4.1	74.0 ± 4.0

Somewhat unexpectedly, the normed feature set performs on par or worse than the full feature set. This implies that a SVM classifier can take advantage of the additional data provided by the individual dimensions.

Features were extracted over the raw and high-passed balancing signal. Likely, only one of these will be redundant. Many measures such as the Higuchi and Petrosian fractal dimension are also correlated, and performance can be improved in some models by removing these redundant features. We evaluated a number of state of the art feature selection methods (*Table* 1.7) and with a RBF SVM and sparse Gaussian process (GP)[4] optimised over the 204 feature normed set (hyperparameter search for each feature subset was

[4]Kernel based models are computationally infeasible for large datasets. *Sparse* methods involve learning over a representative set of the full data [250]. SVMs are inherently sparse, and sparse GPs [251] are an approximation of a full GP.

computationally prohibitive and the parameters of the normed model would be a good midway for small and large feature sets).

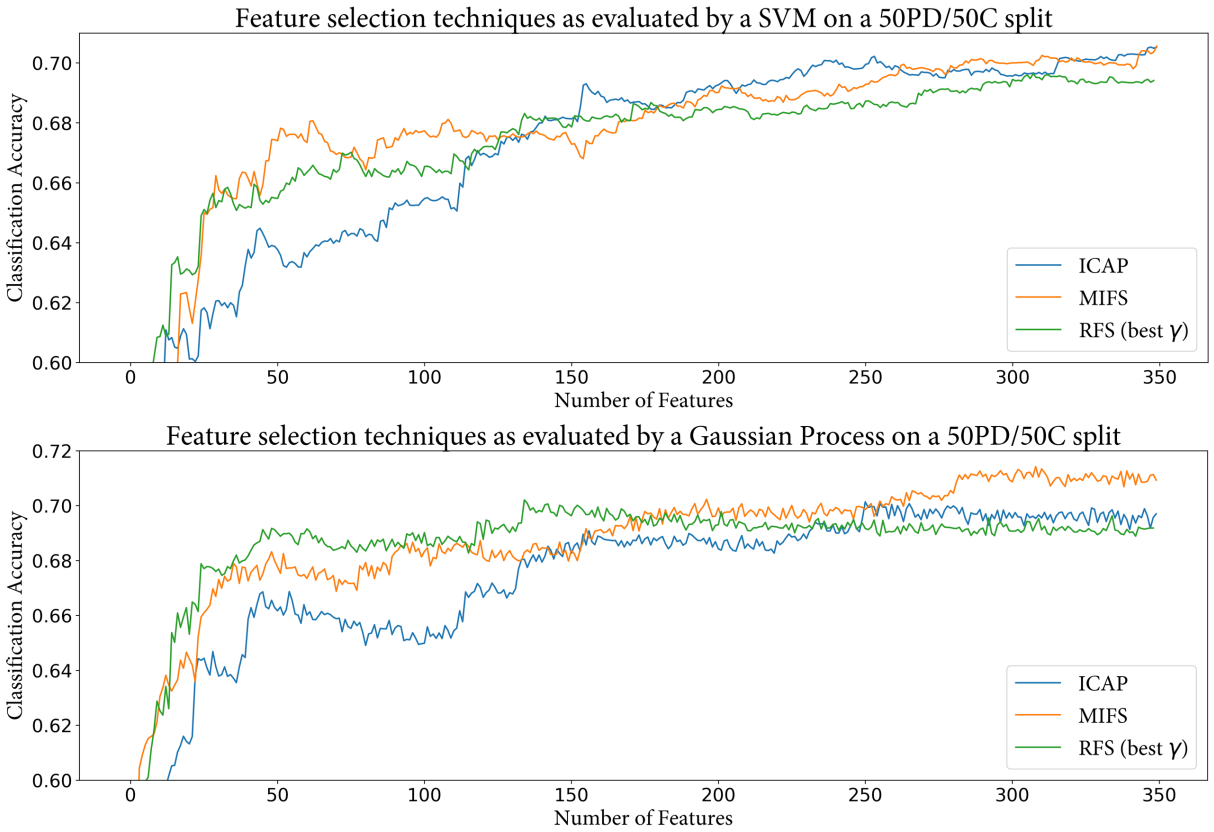

Figure 2.13: Most notable 3 feature selection techniques evaluated with pre-optimised SVM and sparse GP on the full accelerometer feature set. Information theoretic filters generally performed best along with $l_{2,1}$ norm minimisation based RFS.

The best feature subsets exhibited a correlation between performance and the number of features used. This is likely due to sparse GPs and SVMs being reasonably robust to redundant and correlated features (unlike naïve Bayes), and the natural uncertainty of the features as predictors. With the information theoretic based feature selection techniques, there are 'spikes' in classification accuracy — suggesting that there exist features which appear uninformative, but are informative when combined with other features.

The performance of the best 150–350 feature subsets appeared to exceed the performance over all features (*Table* 2.6), despite the lack of hyperparameter search. This could be an artefact of using 3-fold cross validation rather than 10-fold. We evaluated the final MIFS 350 feature subset with our standard methodology, finding no significant performance improvement over the full feature set. This reinforces the need for robust model evaluation before accepting a result.

***Highlight* 2.11.** The feature reduction techniques explored were not very effective. This is primarily due to the inherently high uncertainty and correlation of most features.

An interesting observation is that the GP and SVM differ in performance on the same feature subsets. Although ICAP performed well with a SVM, CIFE was more suitable for the GP. Similarly, the small 50 feature subsets selected by RFS performed exceptionally with the GP whereas they were average when used by the SVM. This suggests that each model uses a different set of information to make classifications, and performance could potentially improve with a more complex model.

Noting this observation, we investigated the performance of an ensemble model consisting of seven independent predictors: A SVM, Gaussian process, random forest, k-nearest neighbour (with $k = 3$) classifier, and neural networks with three, five and seven hidden layers. All models chosen were suited to the task, diagnosing PD within ± 2 per cent accuracy of each other.

The goal of an ensemble is the combine multiple machine learning models to create a more robust one. Ensembling techniques such as bagging and boosting [252] are more suitable for ensembling 'weak' learners [253]. When strong predictors are used, it is more appropriate to aggregate their predictions, either by averaging (*voting*) or using another model (*stacking*) [254].

Building on stacking, *Feature-Weighted Linear Stacking* (*FWLS*) assigns a weight to each feature and model combination [255]. The motivation of FWLS is that in complex feature spaces, models make predictions with different sets features. This hypothesis is likely, due to the varying performance of the SVM and GP on different subsets of features. FWLS was the technique used to ensemble the winning model of the prestigious "Netflix prize" [256].

A RBF Gaussian process was chosen to aggregate the models in stacking and FWLS based ensembles. Gaussian processes are inherently probabilistic classifiers and are suitable for making decisions in situations of high uncertainty. Although the Netflix prize was won using gradient boosted trees [257], Gaussian processes worked better in our problem. The results of these ensembling techniques are presented in Table 2.7.

FWLS ensembles are the stronger option, showing an improvement over the SVM and stacked models with a Bayes factor of 10^{15} and 10^{2} respectively. The ensemble models appear to have a higher variability in results — especially with sensitivity and specificity. A quick analysis of the results showed that the ensemble was more robust – most repetitions of k-fold had better accuracy than the SVM model. However, each individual fold showed a large variance in sensitivity and specificity, a primary influence of the neural networks which alternated between biasing for sensitivity and specificity.

Table 2.7: The results of various ensembling techniques over a combination of SVM, Gaussian process, random forest, k-nearest neighbour, and a three, five and seven layer neural network.

	SVM Only	Voting	Stacked	FWLS
Accuracy [%]	68.8 ± 3.7	70.0 ± 4.5	70.9 ± 5.0	72.3 ± 5.4
Sensitivity [%]	67.2 ± 5.3	66.9 ± 7.5	66.3 ± 7.8	69.4 ± 7.6
Specificity [%]	70.5 ± 5.4	73.3 ± 7.1	75.6 ± 6.0	75.1 ± 7.6
AUROC [%]	74.8 ± 4.1	76.2 ± 5.1	78.1 ± 5.1	79.6 ± 4.9

Highlight 2.12 (Ensembling). Ensemble models significantly improved results, combining the advantages of multiple models. However, they are more difficult to interpret.

Training ensemble models is more computationally expensive; however, making a prediction from a pre-trained model will take a negligible amount of time. Although the current cost of training is negligible, larger datasets required to train robust models used in the medical context may be computationally intractable. SVMs and Gaussian processes in particular require polynomially more computational resources for each data-point, and quickly become intractable upon exceeding tens of thousands of data points. Localised approximations of these models which run in linear time exist [258], though ensembles of neural networks may be more appropriate for large datasets. Ensembles also suffer from the 'black-box' effect, where their results can be difficult to interpret.

2.6 Automatic Feature Engineering: Neural Networks

A notable weakness of feature engineering is that information is lost, as it is difficult for features to perfectly describe a signal. In computer vision and EEG signal processing [186], CNN and LSTM based neural network architectures present a viable solution, automatically 'learning' the best features from the raw signal.

The biggest tradeoff in using automatic feature engineering is that understanding the features engineered is a bigger investment than developing the model. Trust must be placed on the model and that it has not conveniently overfit on the dataset. However, diagnosing PD is a difficult enough task such that many features are already difficult to understand. As explored in *Section* 2.4.3, no individual feature is a good quantifier of PD and a similarly difficult to interpret model must be used to fit these uncertain features.

***Highlight* 2.13.** Structured neural networks are able to extract features from the raw signal data, and are extremely effective in speech recognition and EEG signal processing.

We have decided to focus on automatic feature engineering for accelerometer signals. This was motivated by two factors: primarily, architectures based on accelerometer data are far more lacking compared to the abundance of speech recognition models [247, 259]. Neural networks have only been recently applied to accelerometer signals in the field of human activity recognition [187]. Secondly, the three dimensions (x, y, z) of data presents an additional challenge.

Training deep neural networks is fraught with the possibility of overfitting — especially as multiple epochs of gradient descent are used in training, effectively resulting in different models for the data at each epoch. The ten times repeated 10-fold cross validation performed on previous models would not be as feasible given the available computational resources (no GPU). Conveniently, the PD Digital Biomarker DREAM Challenge [241] coincided with the timeline for this thesis. The challenge involved predicting PD based on the mPower motion data and is evaluated on a previously unreleased set of the data — perfect for unbiased evaluation!

We propose two neural network architectures to extract features from the accelerometer data, inspired by the current state of the art in speech recognition and computer vision.

2.6.1 Background/Inspiration

Although neural networks can extract information from raw signals, transformations can guide machine learning models to the goal. From *Section* 1.2.1 recall that audio signals are

one-dimensional vectors, representing the amplitude of the sound wave over time, generally sampled at around 44,100Hz with phone microphones.

Traditional speech recognition systems were based on Hidden Markov Models which performed best when paired with the information dense MFCC [260]. MFCC and other transforms can result in the loss of valuable information; thus it is preferable to work with either the raw frequency or spectral information. Recent research has shown that deep neural networks are better equipped to handle information coming directly from a short-time Fourier transform [73].

Recently, novel architectures have been proposed which are suitable for processing the raw signal. *Wavenet* [259] is one of the most significant developments, proposed as a method of generating speech. Wavenet stacks *dilated* convolutional layers to create a large 'receptive field' of data with few input layers. Residual and skip connections were used to enable training deeper networks. Wavenet inspired our first model.

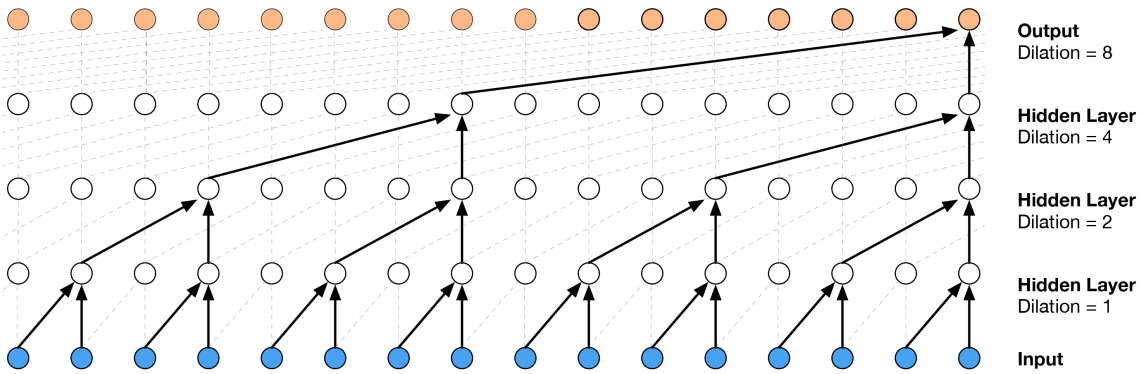

Figure 2.14: Stacked dilated convolutional layers have a larger receptive field than normal convolutional layers. Image from Oord et al. [259].

It is a fusion model consisting of two convolutional layers which feed into a number of LSTM layers. Short time spectral features from the 40 dimensional Mel-scale filterbank are input to both the convolutional and LSTM layers. The CLDNN based model shows a fair improvement over prior models; however, it is evident that the behaviour of these fusion architectures are not yet well understood. Individually, the behaviour of CNNs and LSTMs are difficult to interpret [184, 160] and when combined their behaviour is more unpredictable.

2.6.2 Automatic Feature Engineering for PD Diagnosis

Our problem space is quite different from prior works in the field . Although mPower is the largest current dataset of PD data, its size is not comparable to computer vision and speech recognition datasets. For instance, the ImageNet [261] dataset for computer vision consists of 14.2 million images and CLDNN [247] was trained on 2,000 hours of speech. There are

many (30,000) walk recordings, but a majority of these are from a small subset of users. To avoid the 'digital fingerprinting' effect[5] biasing the data, a maximum of 20 recordings were taken from any participant, resulting in a final dataset size of 12,000.

Overfitting was a major issue, and the problem changed from designing the most complex 'deep' architecture capable of modelling the data, to designing an architecture which extracts useful features before severely overfitting. Simple, highly regularised models performed best in this task as they were less likely to memorise the data.

Accelerometer signals contain three dimensions of information (x, y, z) and are sampled at a much lower rate: approximately 100 Hz in the mPower data (vs. the 44,100 Hz of audio). Experiments with various neural network architectures showed that using the normed acceleration signal performed better than using all three individual dimensions. Although using all dimensions provides more information, the lack of training data resulted in the neural architectures quickly overfitting before extracting useful features.

Another difference is that diagnosing PD is a binary classification task and does not require intermediary predictions — unlike Wavenet which was designed for speech generation, and CLDNN for speech recognition. There are also two distinct tasks (walking and balancing) involved in the mPower dataset, hence the final architecture would be an ensemble of at least two smaller architectures.

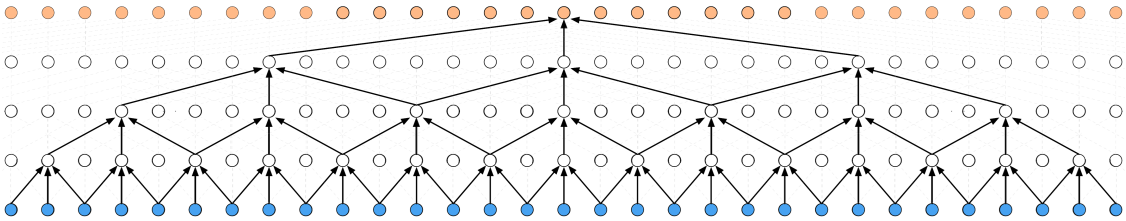

Figure 2.15: Bi-directional connections in Wavenet performed better in our task.

Out first architecture was inspired by Wavenet. As Wavenet is designed for speech generation, it only considers data from the 'past'. We modified the Wavenet architecture, adding a dependency to 'future' data. This was inspired by bidirectional RNNs which perform better in classification problems by utilising both past and future information. We term this *Bi-Wavenet*, which exhibited improved convergence speed and peak validation accuracy compared to Wavenet. Following CLDNN, we explored a number of designs for the LSTM+CNN fusions. Layers of LSTMs followed by convolution layers seemed to be the most effective at extracting features. We denote this architecture *LSTMConv*.

[5]Neto et al. [108] discovered that allowing mPower participants to appear in both train and test sets would result in an AUROC of 98 per cent for models trained on the walk sample. Stratifying by participant would reduce the AUROC to 45 per cent.

Transforming the input data showed significant improvements to model performance. This is especially true with a limited amount of data, as models were less likely to overfit on the more interpretable data format. A short-time Fourier transform was used to map the input into the frequency domain. For the balancing data, a 3.2 second window with 1 second step length was the most effective input for both architectures. The walking data was more complicated as a long Fourier transform would remove valuable temporal information, and models trained on the raw signal would quickly overfit before extracting useful features.

A compromise was made, training a model on both the raw and spectral signal with a 0.5 second window and 0.2 second step length. The outputs of each model were concatenated and a final neural network used to reduce the dimensionality of the data. Features were extracted by taking the activations of the penultimate layer of the network.

The LSTMConv architecture was vastly superior to Bi-Wavenet when handling spectral input. On the raw walking signal, Bi-Wavenet was noticeably more effective as LSTMConv based models would overfit on the walking signal before extracting useful features in validation. The final neural network architecture trained and used in the feature submission to the DREAM biomarker challenge is depicted in Figure 2.16.

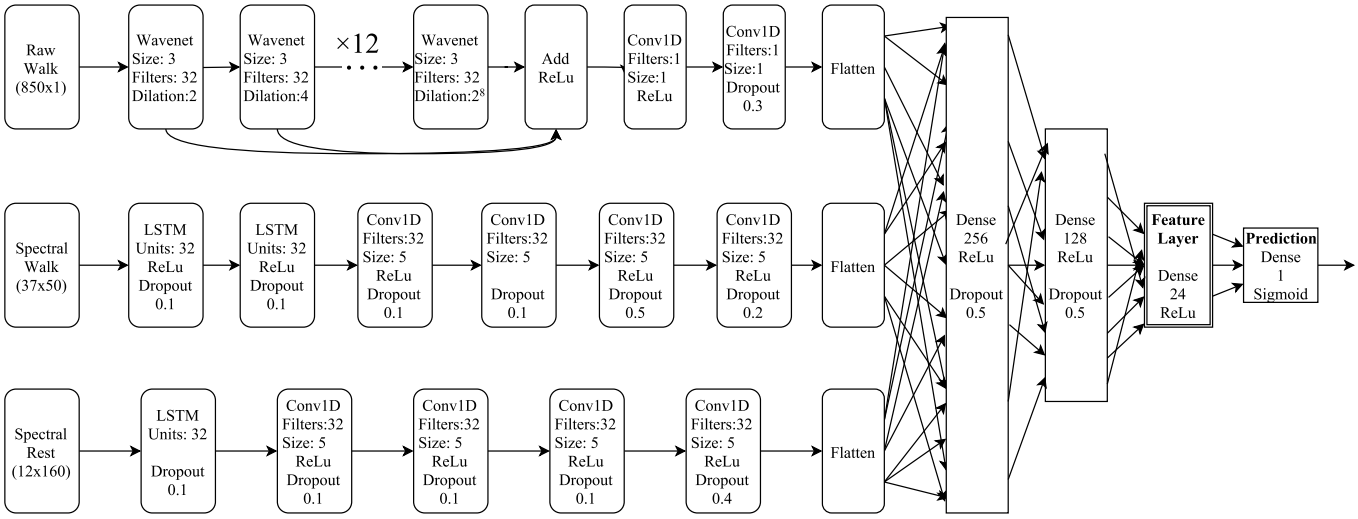

Figure 2.16: The final neural network architecture used to extract features submitted to the DREAM challenge.

Using the DREAM evaluation methodology on the features we submitted, we obtained a cross validation AUROC of 71 per cent with all traditionally engineered features, and a cross validation AUROC of 80 per cent when the neural network based features were added. The results on the unseen testing dataset will be announced on the 31st of October at synapse.org/digitalbiomarkerchallenge.

2.7 Humans vs. Machines: A Discussion

Now that we have shown the applicability of a range of techniques for diagnosing Parkinson's disease with raw sensor data, let us answer the question of what machine learning can provide to real-world diagnosis. The raw classifier performance on the mPower dataset is insufficient to replace neurologists; however, the results are impressive — considering the diagnosis is solely based on a 4 second phonation of /aa/ with environmental noise, or a low quality accelerometer recording of walking.

***Highlight* 2.14.** Can machine learning differentiate between healthy individuals, and those with PD, but no speech difficulties?

In this section, we investigate if machine learning is able to detect symptoms imperceptible to a neurologist. Prior to training any models, we removed a set of participants with PD who declared themselves as having no speaking difficulties on the mPower UPDRS survey. We are aware that this relied on the honesty of participants and may not necessarily correlate to a neurologist's UPDRS assessment. There were 220/900 participants who had responded as having zero speaking difficulties and had filled in an answer other than the default (zero) for one other UPDRS criteria after the speech question. Of these participants, we listened to each audio sample and evaluated them for both quality and whether dysphonia was evident[6]. This resulted in a final set of 86 participants, which were removed from the dataset in all earlier models prevent overfitting from model tweaking.

Table 2.8: Results of the FWLS ensemble on the 86 unseen PD participants with no speaking difficulties and the 28 PD participants in Sakar et al. [57]. Data augmentation was not applied due to limited /aa/ phonation time in recordings from Sakar et al.

	Cross Validation	No Speech Difficulties	Sakar
Accuracy [%]	75.4 ± 2.7	73.1 ± 1.9	85.7
Sensitivity [%]	71.1 ± 4.1	73.1 ± 1.9	85.7
Specificity [%]	79.7 ± 4.6	–	–
AUROC [%]	82.5 ± 2.6	78.3 ± 1.7	90.6
Total Subjects	1716	86	28

We applied the most powerful model we had available, the FWLS ensemble (*Section* 2.5.2) to this task with our standard ten times repeated 10-fold cross validation. The set of 86 participants were divided into ten groups of 8-9 individuals, and the ensemble trained

[6]Optimally, a neurologist would be involved in this process. The resources were not available. Note that the evaluator had listened to 1,600 speech samples from PD and control participants to develop the filtering criteria and is familiar with the characteristics of dysphonia.

on all except one of these groups. To confirm results and ensure that the models able to generalise outside of the mPower dataset, we test our model's performance on the speech samples of 28 subjects with PD released by Sakar et al. [57] on the UCI Machine Learning repository. The best /aa/ phonation for each participant of 1.5 seconds of length was extracted as testing data. The results are presented in *Table* 2.8.

***Highlight* 2.15.** Machine learning is able to diagnose participants with PD without speech difficulties.

Surprisingly, there is no statistically significant difference between the specificity when diagnosing participants with perceptible and imperceptible dysphonia. This implies that there must exist some features, or a combination of features which enable the models to distinguish participants with PD and no speech difficulties and control. Some of these features will be explored in the next section. The result on Sakar et al. [57] provides strong evidence that the mPower trained models are not overfitting, and generalise well to unseen data.

2.7.1 Human Hearing vs. Signal Processing

As the ensemble model obtains similar classification accuracies on both the participants with and without perceptible speech difficulties, combinations of features must exist which distinguish Parkinsonian and normal speech. Some features quantify a characteristic which should be discernible with human hearing, whereas others may measure more abstract qualities.

***Highlight* 2.16.** There must be a subset of features which differentiates participants with PD and no speech difficulties to healthy individuals.

We investigate how some of the features we have introduced relate to human senses. If individuals with PD and and *no speech difficulties* (NSD)[7] are distributed similarly, and NSD and control differently, this implies that the feature is likely *inaudible* to the human ear. If the distributions of NSD and control are similar and control and PD different, the feature is likely to be *audible*. Otherwise, the distribution of NSD may be similar to neither PD nor control. This may imply that the feature is audible to the human ear; however, not directly associated with Parkinson's disease.

These hypothesis will be tested with two approaches. As the features are generally not distributed in a Gaussian fashion, we employ the non-parametric Kruskal-Wallis [262] ANOVA test, using the Dwass-Steel-Critchtlow-Flinger method to control for repeated

[7] In this section, *NSD* denotes the set of participants with Parkinson's disease and no speech difficulties and *PD* denotes the set of participants with Parkinson's disease and speech difficulties.

pairwise comparisons [263]. Two groups are considered sampled from different distributions if $p < 0.05$, denoted with a star* and $p < 0.001$ denoted by three stars***.

We are also interested in the result that two distributions are similar; however, the assumptions underlying p-value hypothesis testing only allow us to conclude that there is insufficient evidence to show the distributions are different [264]. Bayesian tests such as the Bayes factor ANOVA test can quantify evidence in support of two distribution means being the same (the null hypothesis) [214]. However, these tests are parametric, requiring a prior assumption over the distributions. We apply the Aligned Rank Transform [265] to transform the data into a format suitable for Bayes factor analysis. Two distribution means are considered similar if $p > 0.25$ and $BF < \frac{1}{3}$, denoted by a dagger†. The Aligned Rank Transform is less suitable when repeated comparisons are performed [266], hence the Kruskal-Wallis test is preferred when testing for differences in the distribution (rejecting the null hypothesis).

Note that the results presented in this section are influenced by selection bias, as the NSD group have been verified as performing the task properly. The other groups may have more outliers from performing the task incorrectly or excessive background noise. This should be considered when interpreting conclusions — 'bad' samples often appear as outliers, but this may not be the case for all features.

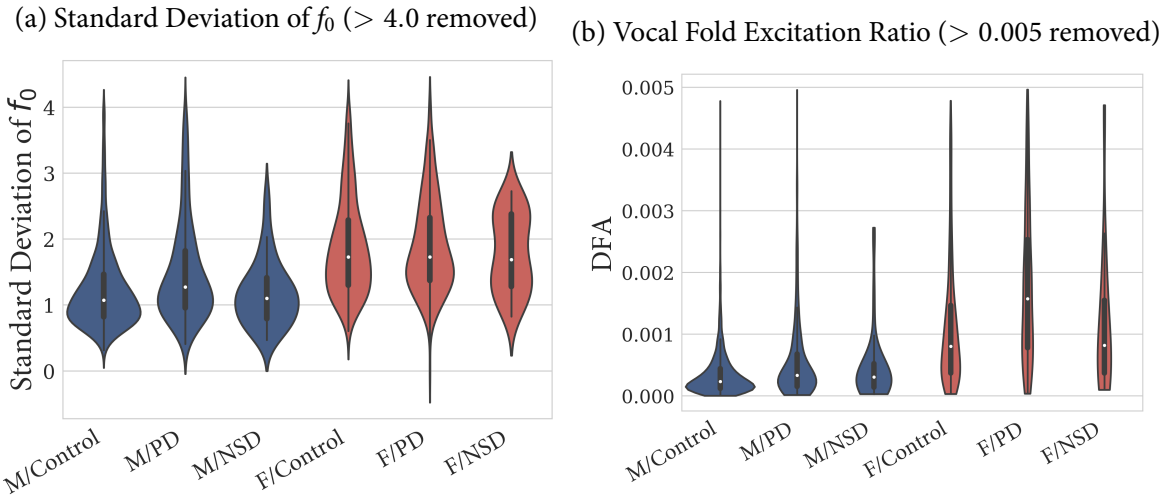

Figure 2.17: Violin plots of features over all samples used in the machine learning models with extreme outliers removed. (a) Standard deviation of f_0 in males is likely audible to the human ear as control and NSD are distributed similarly †, whereas control and PD are different***. There is no significant variance in f_0 between females†. (b) The VFER of female speech is likely audible to humans as control and NSD have similar distributions† while control and PD are different***. Additional data is required for males.

The human ear is sensitive to changes in pitch, and Parkinsonian speech has greater fluctuations in frequency. Therefore it is expected that the NSD group has a lesser standard

deviation of f_0 than the PD. The amount of noise (related to breathiness) should also be perceivable with human hearing. This feature is quantified by the harmonics to noise ratio (HNR); however, the HNR was not a feature strong enough to differentiate normal and dysphonic speech; thus, the Vocal Fold Excitation Ratio, an extension of the HNR was used instead [91]. The differences between PD participants with and without speech difficulties are presented in *Figure* 2.17.

Based on the evidence in *Figure* 2.17, we can conclude that there is substantial evidence the variations in f_0 in males are audible to humans. The f_0 is estimated with the SWIPE algorithm [245] and it is possible that the shorter pitch periods of female phonation cause a larger error margin in the f_0 estimates [146], blurring the difference between dysphonic and normal female phonation. The evidence that VFER is audible to humans is less substantial, but it is not a strong measure for differentiating control and NSD groups.

***Highlight* 2.17.** Features such as the standard deviation in the fundamental frequency and the signal to noise ratio are not useful at distinguishing individuals with invisible PD symptoms.

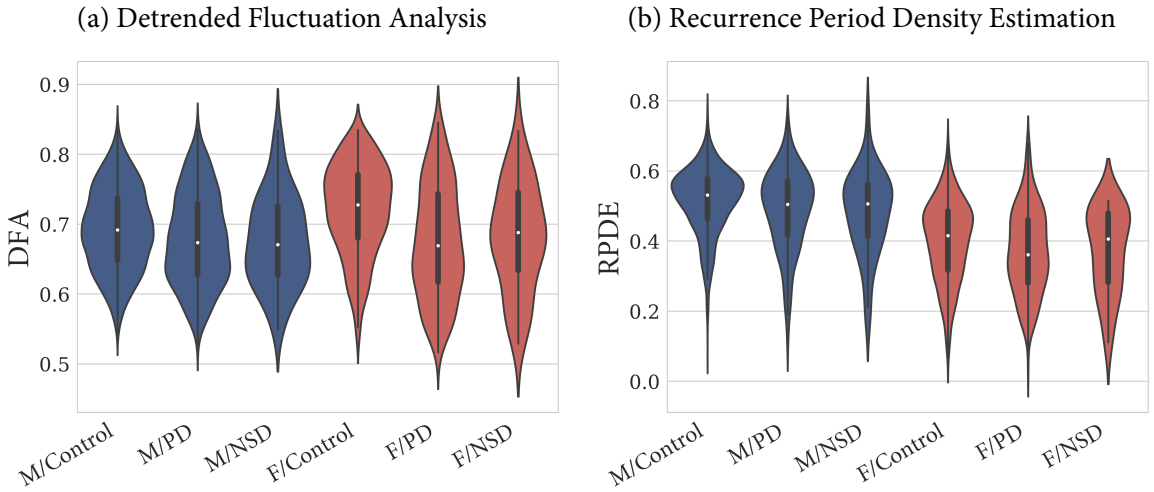

Figure 2.18: (a) DFA is likely inaudible in both males and females as PD and NSD are distributed similarly† and control and NSD differently*. (b) RPDE likely isn't audible in males as PD and NSD are distributed similarly† and control and NSD differently*. For females, evidence in insufficient to conclude whether NSD is sampled from a different distribution to control or PD.

DFA and RPDE are two of the stronger non-linear measures used in prior work, which detect Parkinsonian dysphonia [25]. DFA is a measure of the autocorrelation of a signal, and RPDE is a measure of periodicity. Like VFER, these features attempt to quantify the amount of disturbance in the speech signal from turbulent noise. However, VFER relies on DYPSA [92] to measure the fundamental frequency whereas DFA and RPDE are non-linear methods and may be more robust. The results are presented in *Figure* 2.18.

Although DFA and RPDE both quantify a similar characteristic to VFER (breathiness), there is more more evidence that DFA and RPDE measure inaudible qualities in dysphonic speech. A possible explanation is that DFA and RPDE measure the more abstract notion of recurrence whereas VFER are a more direct measure of the turbulent noise.

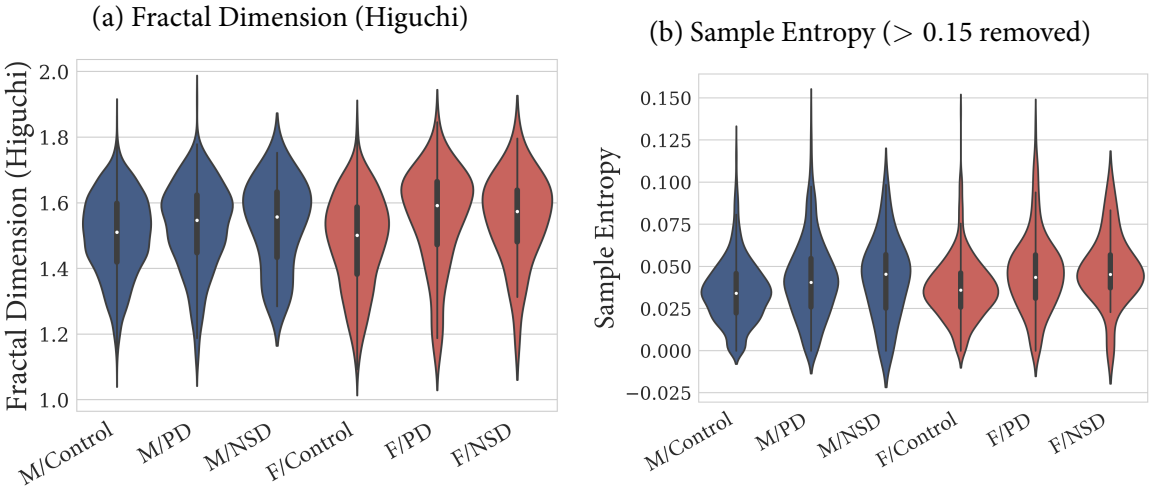

Figure 2.19: (a) The fractal dimension is likely inaudible, as PD and NSD are distributed similarly† whereas control and NSD are different* (b) The sample entropy is also likely inaudible, as PD and NSD are distributed similarly† and control and NSD differently in males* and females***.

The fractal dimension and sample entropy are two features introduced from EEG signal processing. These features measure the signal's complexity and uncertainty, and are less biologically motivated than DFA and RPDE. In particular, these features often relate to detailed high frequency information, to which the human auditory system is less sensitive to [267]. We expect that the non-linear features introduced in *Section* 1.2.4 are some of the better differentiators of NSD and control as they quantify similarly abstract qualities of a signal.

As shown in *Figure* 2.19, fractal dimension and sample entropy of a signal are unlikely to be audible qualities of a signal. It is a combination of these features which enable machine learning to differentiate healthy individuals and those with PD but no audible speech difficulties.

***Highlight* 2.18.** The novel features introduced in *Section* 1.2.4 were effective at differentiating individuals with invisible PD symptoms.

2.7.2 Significance of Findings

From *Section* 2.4, it is clear that making sense of each individual feature is a challenging task. Machine learning models are required to interpret the uncertainty of the features for a robust diagnosis. Prior models in this thesis have been constructed solely based on the characteristics of either a speech or walking signal. Including demographics information will significantly improve results. This is not solely due to the differences in PD amongst different genders and races [268, 269, 97] — demographics information also provides valuable information about the expected speech and gait characteristics.

Younger subjects were filtered out to create an approximate 50/50 split of the data and simulate more realistic diagnosis scenarios. Using the rule age $\geq$ 35 on phonation task, a set of 1591 individuals, 803 with PD and 75 with PD and no speaking difficulties remain. The rule age $\geq$ 34 on the movement task results in a set of of 775 individuals, 389 with PD. Our FWLS model with all novel features and demographics features achieves 82.1 $\pm$ 2.7 per cent classification accuracy on the phonation task and 81.7 $\pm$ 3.3 per cent on the walking task. These models did not use data augmentation or neural network based feature engineering, which are likely to further improve accuracy.

Although the performance of these machine learning models are insufficient to substitute neurologist diagnosis, they provide an effectively zero cost tool for neurologists to validate their diagnosis. Previously, we have shown that these machine learning models can detect dysphonia better than most humans (*Section* 2.7) and it is likely that this is true for irregular gait. A positive result may be an sign to perform further checks on the patient. The results are especially good, considering many samples present in the mPower data may not be of suitable quality for diagnosis by a trained neurologist.

***Highlight* 2.19.** Although the current performance is not sufficient to replace neurologists, machine learning can be a valuable aid in the diagnosis process.

Additional data and cleaner data will always improve a classifier's performance, and we are far from reaching the limits of machine learning. In all models tested, over half of the mPower data was removed to allow for the 50PD:50C stratification. This left 1,700 subjects in the phonation task and 1,200 subjects in the walking task — a sufficiently large sample for a good representation of the population; however, insufficient for such a complex task. This is evident in the significant improvements in results from simple data augmentation techniques in *Section* 2.5.

These diagnoses can be performed without the installation of any application on the user's phone. As a demonstration, we developed a web application which used the Javascript

web audio and device motion APIs to record microphone and accelerometer data. The recordings were submitted to a basic (2GB RAM/single core) server which could perform diagnosis within 30 seconds running unoptimised code. The predictions are high variance — taking the mean of multiple diagnosis on segments of each task (like data augmentation in *Section* 2.5) greatly improved robustness at the cost of extra computation.

***Highlight* 2.20.** It is extremely easy to place a trained machine learning model on the cloud — making diagnosis or monitoring extremely cheap and easy.

It is clear machine learning can be a valuable tool for providing insights to neurologists. There is also enormous potential for growth — as more data becomes available to machine learning models, their power and robustness will vastly improve.

2.8 Implementation

We would like to extend our thanks to all open-source libraries, and academics making their code publicly available. Without these, development would have been a significantly slower process.

The project was primarily implemented in Python, acting as an interface between a number of libraries written in Matlab (Matlab Engine), R (rpy2) and C. The code for this project is published online at https://github.com/maxwg/parkinsons-mpower.

Wherever possible, reliable standard libraries or implementations used in previous research were preferred to maximise reproducibility and reliability. Standard speech features used in Interspeech were calculated using the official openSMILE [270] program, which uses the sub-harmonic summation method of f_0 estimation [271]. Most dysphonia-specific features were calculated using the toolbox published by Tsanas [91] which uses the SWIPE [245, 146] f_0 estimation algorithm.

Following Tsanas [91], 120hz and 190hz were used as the mean healthy f_0 for males and females respectively. Tsanas and Little et al. [99] used an embedding dimension of 4 for non-linear features such as RPDE; however, our data suggested that 6 was more suitable. This was confirmed with features extracted in the embedding dimension of 6 being more informative in our classifiers.

EEG and non-linear signal analysis was performed using the PyREM library [272], which builds upon PyEEG [273], correcting some implementation flaws. The false nearest neighbour implementation in the pypsr [274] library was used to calculate the embedding dimension for the relevant non-linear signal processing algorithms. The embedding dimension used in both speech and accelerometer was 6 and the embedding delay was chosen per sample as the first minimum of the approximate mutual information [114].

PyREM does not provide an implementation for Lyapunov Exponents and some experimentation with artificially constructed signals showed that PyEEG's implementation is severely wrong. The nolds [275] library was used to compute the Lyapunov exponents, using Rosenstein's [114] algorithm for λ^* and Eckmann's algorithm [157] for $\lambda_1 - \lambda_6$.

Jerk based accelerometer features were considered [153]; however, performance was generally worse than using the raw acceleration features over all computed measures. The DREAM challenge baseline features [242] were used in conjunction with our own implementation of the methods used in Arora et al [19]. The primary differences likely result from our preprocessing of the data, which rotated all accelerometers relative to the x axis.

Traditional machine learning algorithms were mostly based on the standard scikit-learn [276] implementation, with Gaussian processes implemented with GPy [277]. Hyperopt [278] was used to aid in finding the optimal hyperparameters for some models. Scikit-feature [204] was used to implement the filter and embedded feature selection methods. Stacked regressions [254] and Feature-weighted linear stacking [255] were implemented with the stacked_generalization library [279].

Neural networks were primarily implemented in Keras [280]. In general, ReLu [192] was chosen as the activation and Nadam [200] used as the optimiser. Batch Normalisation [182] was applied whenever relevant and dropout [190] used to regularise the models. A sigmoidal node with a binary cross-entropy loss function used in binary classification. For regression of UPDRS scores, the Huber loss [281] was selected.

3 Conclusion

Diagnosing Parkinson's disease with machine learning is a task far more difficult than what is suggested by prior literature. Replicating prior work on PD diagnosis with smartphones achieves just above 60 per cent performance — a large decrease from the reported 98 per cent. This reveals the dangers of machine learning in small datasets, with their results caused by either bias in the data and/or overfitting on cross validation.

We identify that results obtained in small datasets may not extrapolate outside of the dataset, and utilise the much larger mPower dataset. The mPower dataset consists of crowdsourced audio recordings of /aa/ phonation and a gait and balance task from 6,500 users, 1,100 with PD. We apply a number of novel features commonly used in EEG signal processing to the task, showing a noticable improvement. We then visualise and analyse some of these features, showing how noisy they are, and the necessity of machine learning to make sense of them.

Following a better understanding of the features, we then investigate the applicability of a number of strategies for improving machine learning performance: data augmentation, features selection and model ensembling. Data augmentation and model ensembling are shown to be highly effective, whereas feature selection selection is not.

We then investigate the concept of automatic feature engineering with neural networks on the accelerometer data. Automatic feature engineering has been successfully applied on image and speech data, but applications to accelerometer data have been more limited [187]. We develop an architecture inspired by recent developments in speech recogntition [259] and show that it is an extremely effective tool on top of traditional feature engineering. We did not have the resources to perform our stringent model evaluation strategy, so we took advantage of an unseen mPower test set released as part of the DREAM PD biomarker challenge.

After developing the infrastructure for a reliable model to diagnose Parkinson's disease, we refer to our original question: "is it possible for machine learning to detect symptoms imperceptible by humans?" We had a hold-out set of 86 participants who marked

themselves as having zero speech difficulty, verified by the researcher. We discovered that machine learning could diagnose these individuals as accurately as the other participants with more noticeable speech disabilities. The features enabling machine learning to achieve this were examined — in particular, most novel measures introduced were strong differentiators of PD and inaudible. This is a very promising result for machine learning, and we finish with some recommendations on how to progress and ideas that we developed but did not explore.

3.1 Where to? Recommendations for Future Research

This thesis may be very information dense, yet covers only the beginning for what must be completed before machine learning can be applied in a medical context. Trust of machine learning is the biggest barrier. This can be developed by *understanding* the behaviour of these machine learning models better, and recognising their strengths and weaknesses.

We have introduced many of features to diagnose Parkinson's disease from voice and gait, and only perform a preliminary analysis of a subset of these features. A thorough analysis on a cleaner dataset is desirable — it may be the case that some features are good enough indicators of PD once the recording setup is controlled, and may be interpretable without machine learning. We have also noted that a number of features are not length-invariant — developing length invariant alternatives or using a time stepped augmentation approach would greatly improve robustness.

As we have shown in *Section* 2.6, neural networks are a powerful tool in the feature engineering process. However, there remains a lot to be desired in terms of understanding and interpretation of these networks. The behaviour of CNNs and LSTMs have individually been visualised [184, 160] and there is still a significant amount of work to be done in understanding either architecture. Fusion Architectures combining both LSTM and convolution layers are even less understood [282], but they perform exceedingly well on time series data [247].

Although machine learning may not be at a stage where it can be directly applied in a medical context, its ability to detect symptoms of Parkinsonian speech indistinguishable to humans is a very promising result for the field. This thesis only delves into a basic exploration of the features which allow machine learning to do this, and a more comprehensive analysis with more reliable data is required. It may be possible to visualise some inaudible features in the form of an oscilloscope or spectrogram.

Telemonitoring is one of the most promising applications of machine learning as small errors in predictions are tolerable. Setting up a smartphone telemonitoring platform would

also provide a good source of data to train future machine learning models. Although telemonitoring has not been discussed in this thesis, the methods used in feature engineering are interchangeable. The biggest change would be the machine learning model, with Gaussian processes and neural networks more suitable for regression tasks. The work in this thesis has also been adapted and submitted to the second challenge of the DREAM PD biomarker competition [241], which involves predicting UPDRS scores from smartwatch accelerometer data.

There were many ideas developed in completing this project, and only a small subset could be explored. The diagnosis of Parkinson's disease commonly involves observing a subject's response to medication such as Levodopa [8]. The mPower dataset has fields describing the medication state of participants with PD. Utilising this additional information and developing features measuring the differences before and after medication could significantly improve diagnosis accuracy.

Parkinson's disease motor symptoms are often presented asymmetrically, with studies showing 86 per cent of patients exhibit different symptoms on different sides of their body [71]. The biological mechanisms are not well understood [283]; however, asymmetry in symptoms may be able to help differentiate Parkinson's disease from other similar disorders. With smartphone based data, measuring asymmetry would require repeated tests while holding the smartphone with either hand — or like the second PD DREAM subchallenge, having smartwatches on both wrists.

We explored some feature engineering work in the Interspeech challenges, which covers tasks ranging from emotion recognition to speaker trait recognition [284]. The features and approaches to these challenges are undoubtedly relevant to PD diagnosis via speech. In our limited exploration of the Geneva [151] and ComParE [152] feature sets, significant improvements were observed; however, these feature sets were not analysed in depth due to prioritising dysphonia specific and our novel features.

There is much more work to be done, and it is important that it is done in an open manner. Progress should be measured on standardised datasets, allowing for empirical comparisons between techniques. The actual numbers should not matter — only that a performance increase has been observed. One of the biggest advantages of machine learning is that many basic methods can be ensembled into a high performing model with minimal effort [256]. Releasing code publicly so results can be easily replicated or extended will also be important in accelerating the field. Methods are becoming more complex and it can be very difficult to concisely summarise them in research paper format.

3.2 A Finishing Note

We advocate that measures such as accuracy and AUROC should be used for the purposes of model selection and not interpreted as the ability of a model to generalise to the real-world or to different datasets. Machine learning for Parkinson's disease is still in its initial stages, and controlling for dataset quality and ensuring a representative sample of the population at this stage would slow progress.

The importance of significance testing and careful cross validation to prevent overfitting should also be evident (*Section* 1.3.4). As machine learning models continually improve, the benefits from improvements will become less significant. As improvements become more marginal, testing that an increase in performance is due to the improvement rather than natural variance in the machine learning model's predictive power will become necessary.

The reader may have noticed that we discarded more than half of the mPower dataset to obtain a 50PD/50C stratification in the models we have trained. We have also demonstrated the power of automatic feature engineering, though have not applied it to our results. The goal of this thesis is not to present the best possible results, but rather to demonstrate the applicability of each individual approach. Machine learning enables us to ensemble the results of many individual improvements to develop more powerful models. Every little improvement matters, as evident in the Netflix prize, where all top teams involved large ensembles of smaller, simpler models [256].

Bibliography

[1] J. M. Savitt, V. L. Dawson, and T. M. Dawson, "Diagnosis and treatment of Parkinson disease: Molecules to medicine," *The Journal of Clinical Investigation*, vol. 116, no. 7, pp. 1744–1754, 2006.

[2] D. J. Brooks, "Parkinson's disease: Diagnosis," *Parkinsonism & Related Disorders*, vol. 18, pp. S31–S33, 2012.

[3] H. Bernheimer, W. Birkmayer, O. Hornykiewicz, K. Jellinger, and F. Seitelberger, "Brain dopamine and the syndromes of Parkinson and Huntington clinical, morphological and neurochemical correlations," *Journal of the Neurological Sciences*, vol. 20, no. 4, pp. 415–455, 1973.

[4] F. Wilkins, "What is Parkinson's disease?," *WPF*, 2011.

[5] S. Pålhagen, E. Heinonen, J. Hägglund, T. Kaugesaar, O. Mäki-Ikola, R. Palm, S. P. S. Group, *et al.*, "Selegiline slows the progression of the symptoms of Parkinson's disease," *Neurology*, vol. 66, no. 8, pp. 1200–1206, 2006.

[6] A. L. Whone, R. L. Watts, A. J. Stoessl, M. Davis, S. Reske, C. Nahmias, A. E. Lang, O. Rascol, M. J. Ribeiro, P. Remy, *et al.*, "Slower progression of Parkinson's disease with ropinirole versus levodopa: The REAL-PET study," *Annals of Neurology*, vol. 54, no. 1, pp. 93–101, 2003.

[7] S. Fahn, P. S. Group, *et al.*, "Does levodopa slow or hasten the rate of progression of Parkinson's disease?," *Journal of Neurology*, vol. 252, no. 4, pp. iv37–iv42, 2005.

[8] E. Tolosa, G. Wenning, and W. Poewe, "The diagnosis of Parkinson's disease," *The Lancet Neurology*, vol. 5, no. 1, pp. 75–86, 2006.

[9] D. R. Williams and I. Litvan, "Parkinsonian syndromes," *Continuum: Lifelong Learning in Neurology*, vol. 19, no. 5 Movement Disorders, p. 1189, 2013.

[10] A. J. Hughes, S. E. Daniel, Y. Ben-Shlomo, and A. J. Lees, "The accuracy of diagnosis of Parkinsonian syndromes in a specialist movement disorder service," *Brain*, vol. 125, no. 4, pp. 861–870, 2002.

[11] N. Quinn, "Parkinsonism–recognition and differential diagnosis.," *British Medical Journal*, vol. 310, no. 6977, p. 447, 1995.

[12] J. Jankovic, A. H. Rajput, M. P. McDermott, and D. P. Perl, "The evolution of diagnosis in early Parkinson's disease," *Archives of Neurology*, vol. 57, no. 3, pp. 369–372, 2000.

[13] S. Daniel and A. Lees, "Parkinson's Disease Society Brain Bank, London: Overview and research.," *Journal of Neural Transmission. Supplementum*, vol. 39, pp. 165–172, 1993.

[14] A. J. Hughes, S. E. Daniel, L. Kilford, and A. J. Lees, "Accuracy of clinical diagnosis of idiopathic Parkinson's disease: A clinico-pathological study of 100 cases.," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 55, no. 3, pp. 181–184, 1992.

[15] C. R. Scherzer, A. C. Eklund, L. J. Morse, Z. Liao, J. J. Locascio, D. Fefer, M. A. Schwarzschild, M. G. Schlossmacher, M. A. Hauser, J. M. Vance, *et al.*, "Molecular markers of early Parkinson's disease based on gene expression in blood," *Proceedings of the National Academy of Sciences*, vol. 104, no. 3, pp. 955–960, 2007.

[16] M. A. Nalls, N. Pankratz, C. M. Lill, C. B. Do, D. G. Hernandez, M. Saad, A. L. DeStefano, E. Kara, J. Bras, M. Sharma, *et al.*, "Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease," *Nature Genetics*, vol. 46, no. 9, pp. 989–993, 2014.

[17] Z. Hong, M. Shi, K. A. Chung, J. F. Quinn, E. R. Peskind, D. Galasko, J. Jankovic, C. P. Zabetian, J. B. Leverenz, G. Baird, *et al.*, "Dj-1 and α-synuclein in human cerebrospinal fluid as biomarkers of Parkinson's disease," *Brain*, vol. 133, no. 3, pp. 713–726, 2010.

[18] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.

[19] S. Arora, V. Venkataraman, S. Donohue, K. M. Biglan, E. R. Dorsey, and M. A. Little, "High accuracy discrimination of Parkinson's disease participants from healthy controls using smartphones," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 3641–3644, IEEE, 2014.

[20] L. S. Freedman and D. Pee, "Return to a note on screening regression equations," *The American Statistician*, vol. 43, no. 4, pp. 279–282, 1989.

[21] A. Y. Ng, "Preventing overfitting of cross-validation data," in *Proceedings of the 14th International Conference on Machine Learning (ICML-97)*, vol. 97, pp. 245–253, 1997.

[22] A. Zhan, M. A. Little, D. A. Harris, S. O. Abiola, E. Dorsey, S. Saria, and A. Terzis, "High frequency remote monitoring of Parkinson's disease via smartphone: Platform overview and medication response detection," *arXiv preprint arXiv:1601.00960*, 2016.

[23] P. Esser, H. Dawes, J. Collett, M. G. Feltham, and K. Howells, "Assessment of spatiotemporal gait parameters using inertial measurement units in neurological populations," *Gait & Posture*, vol. 34, no. 4, pp. 558–560, 2011.

[24] L. Ai, J. Wang, and R. Yao, "Classification of Parkinsonian and essential tremor using empirical mode decomposition and support vector machine," *Digital Signal Processing*, vol. 21, no. 4, pp. 543–550, 2011.

[25] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, L. O. Ramig, *et al.*, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1015–1022, 2009.

[26] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests," *IEEE transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 884–893, 2010.

[27] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel, *et al.*, "Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): Scale presentation and clinimetric testing results," *Movement Disorders*, vol. 23, no. 15, pp. 2129–2170, 2008.

[28] J. Cancela, S. V. Mascato, D. Gatsios, G. Rigas, A. Marcante, G. Gentile, R. Biundo, M. Giglio, M. Chondrogiorgi, R. Vilzmann, *et al.*, "Monitoring of motor and non-motor symptoms of Parkinson's disease through a mHealth platform," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pp. 663–666, IEEE, 2016.

[29] N. F. Troje, "Decomposing biological motion: A framework for analysis and synthesis of human gait patterns," *Journal of Vision*, vol. 2, no. 5, pp. 2–2, 2002.

[30] J. M. Hausdorff, M. E. Cudkowicz, R. Firtion, J. Y. Wei, and A. L. Goldberger, "Gait variability and basal ganglia disorders: Stride-to-stride variations of gait cycle timing in Parkinson's disease and Huntington's disease," *Movement Disorders*, vol. 13, no. 3, pp. 428–437, 1998.

[31] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, "Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task," *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2171–2181, 2013.

[32] F. Eyben, *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*. Springer, 2015.

[33] C. Ahlrichs and M. Lawo, "Parkinson's disease motor symptoms in machine learning: A review," *arXiv preprint arXiv:1312.3825*, 2013.

[34] S. Bind, A. K. Tiwari, and A. K. Sahani, "A survey of machine learning based approaches for Parkinson's disease prediction," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 2, pp. 1648–1655, 2015.

[35] C. Duval, A. Sadikot, and M. Panisset, "The detection of tremor during slow alternating movements performed by patients with early Parkinson's disease," *Experimental Brain Research*, vol. 154, no. 3, pp. 395–398, 2004.

[36] A. Salarian, H. Russmann, C. Wider, P. R. Burkhard, F. J. Vingerhoets, and K. Aminian, "Quantification of tremor and bradykinesia in Parkinson's disease using a novel ambulatory monitoring system," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 2, pp. 313–322, 2007.

[37] L. Palmerini, L. Rocchi, S. Mellone, F. Valzania, and L. Chiari, "Feature selection for accelerometer-based posture analysis in Parkinson's disease," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 3, pp. 481–490, 2011.

[38] C. Boussios, J. Greenbaum, B. Ieong, F. Kokkotos, S. Kokkotos, and M. Zalesak, "The construction of a novel statistical algorithm to objectively diagnose Parkinson's disease using smartphone data," *Michael J Fox Foundation*, 2013.

[39] M. Brunato, R. Battiti, D. Pruitt, and E. Sartori, "Supervised and unsupervised machine learning for the detection, monitoring and management of Parkinson's disease from passive mobile phone data," *Michael J Fox Foundation, Available at: https://kaggle2.blob.core. windows.net/prospectorfiles/1117/958625cf-3514-4e64-b0e7-13ebd3cf9791/kaggle.pdf, Last accessed Oct 2017*, 2013.

[40] L. Rocchi, L. Chiari, A. Cappello, and F. B. Horak, "Identification of distinct characteristics of postural sway in Parkinson's disease: A feature selection procedure based on principal component analysis," *Neuroscience Letters*, vol. 394, no. 2, pp. 140–145, 2006.

[41] T.-Z. Chen, G.-J. Xu, G.-A. Zhou, J.-R. Wang, P. Chan, and Y.-F. Du, "Postural sway in idiopathic rapid eye movement sleep behavior disorder: A potential marker of prodromal Parkinson's disease," *Brain Research*, vol. 1559, pp. 26–32, 2014.

[42] M. F. Gago, V. Fernandes, J. Ferreira, H. Silva, L. Rocha, E. Bicho, and N. Sousa, "Postural stability analysis with inertial measurement units in Alzheimer's disease," *Dementia and Geriatric Cognitive Disorders Extra*, vol. 4, no. 1, pp. 22–30, 2014.

[43] R. Begg and J. Kamruzzaman, "Neural networks for detection and classification of walking pattern changes due to ageing," *Australasian Physical & Engineering Science in Medicine*, vol. 29, no. 2, pp. 188–195, 2006.

[44] R. d. M. Roiz, E. W. A. Cacho, M. M. Pazinatto, J. G. Reis, A. Cliquet Jr, and E. Barasnevicius-Quagliato, "Gait analysis comparing Parkinson's disease with healthy elderly subjects," *Arquivos de Neuro-psiquiatria*, vol. 68, no. 1, pp. 81–86, 2010.

[45] A. Khorasani and M. R. Daliri, "HMM for classification of Parkinson's disease based on the raw gait data," *Journal of Medical Systems*, vol. 38, no. 12, p. 1, 2014.

[46] J. Barth, J. Klucken, P. Kugler, T. Kammerer, R. Steidl, J. Winkler, J. Hornegger, and B. Eskofier, "Biometric and mobile gait analysis for early diagnosis and therapy monitoring in Parkinson's disease," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pp. 868–871, IEEE, 2011.

[47] V. Renaudin, M. Susi, and G. Lachapelle, "Step length estimation using handheld inertial sensors," *Sensors*, vol. 12, no. 7, pp. 8507–8525, 2012.

[48] B. Sijobert, M. Benoussaad, J. Denys, R. Pissard-Gibollet, C. Geny, and C. A. Coste, "Implementation and validation of a stride length estimation algorithm, using a single basic inertial sensor on healthy subjects and patients suffering from Parkinson's disease," *Electronic Healthcare*, vol. 7, no. 6, pp. 704–714, 2015.

[49] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, and M. Faundez-Zanuy, "Decision support framework for Parkinson's disease based on novel handwriting markers," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 23, no. 3, pp. 508–516, 2015.

[50] C. Kotsavasiloglou, N. Kostikis, D. Hristu-Varsakelis, and M. Arnaoutoglou, "Machine learning-based classification of simple drawing movements in Parkinson's disease," *Biomedical Signal Processing and Control*, vol. 31, pp. 174–180, 2017.

[51] S. Das, L. Trutoiu, A. Murai, D. Alcindor, M. Oh, F. De la Torre, and J. Hodgins, "Quantitative measurement of motor symptoms in Parkinson's disease: A study with full-body motion capture data," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pp. 6789–6792, IEEE, 2011.

[52] R. Nakamura, H. Nagasaki, and H. Narabayashi, "Disturbances of rhythm formation in patients with Parkinson's disease," *Perceptual and Motor Skills*, vol. 46, no. 1, pp. 63–75, 1978.

[53] H. Hazan, D. Hilu, L. Manevitz, L. O. Ramig, and S. Sapir, "Early diagnosis of Parkinson's disease via machine learning on speech data," in *Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of*, pp. 1–4, IEEE, 2012.

[54] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The Interspeech 2015 computational paralinguistics challenge: Nativeness, Parkinson's & eating condition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[55] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Nöth, "Voiced/unvoiced transitions in speech as a potential biomarker to detect Parkinson's disease," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[56] L. Cnockaert, J. Schoentgen, P. Auzou, C. Ozsancak, L. Defebvre, and F. Grenez, "Low-frequency vocal modulations in vowels produced by Parkinsonian subjects," *Speech Communication*, vol. 50, no. 4, pp. 288–300, 2008.

[57] B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun, "Collection and analysis of a Parkinson's speech dataset with multiple types of sound recordings," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 4, pp. 828–834, 2013.

[58] R. Armañanzas, C. Bielza, K. R. Chaudhuri, P. Martinez-Martin, and P. Larrañaga, "Unveiling relevant non-motor Parkinson's disease severity symptoms using a machine learning approach," *Artificial Intelligence in Medicine*, vol. 58, no. 3, pp. 195–202, 2013.

[59] K. N. R. Challa, V. S. Pagolu, G. Panda, and B. Majhi, "An improved approach for prediction of Parkinson's disease using machine learning techniques," *arXiv preprint arXiv:1610.08250*, 2016.

[60] G. S. Babu and S. Suresh, "Parkinson's disease prediction using gene expression–a projection based learning meta-cognitive neural classifier approach," *Expert Systems with Applications*, vol. 40, no. 5, pp. 1519–1529, 2013.

[61] C. Salvatore, A. Cerasa, I. Castiglioni, F. Gallivanone, A. Augimeri, M. Lopez, G. Arabia, M. Morelli, M. Gilardi, and A. Quattrone, "Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy," *Journal of Neuroscience Methods*, vol. 222, pp. 230–237, 2014.

[62] D. A. Morales, Y. Vives-Gilabert, B. Gómez-Ansón, E. Bengoetxea, P. Larrañaga, C. Bielza, J. Pagonabarraga, J. Kulisevsky, I. Corcuera-Solano, and M. Delfino, "Predicting dementia development in Parkinson's disease using Bayesian network classifiers," *Psychiatry Research: NeuroImaging*, vol. 213, no. 2, pp. 92–98, 2013.

[63] K. J. Stam, D. L. Tavy, B. Jelles, H. A. Achtereekte, J. P. Slaets, and R. W. Keunen, "Non-linear dynamical analysis of multichannel EEG: Clinical applications in dementia and Parkinson's disease," *Brain Topography*, vol. 7, no. 2, pp. 141–150, 1994.

[64] R. Soikkeli, J. Partanen, H. Soininen, A. Pääkkönen, and P. Riekkinen, "Slowing of EEG in Parkinson's disease," *Electroencephalography and Clinical Neurophysiology*, vol. 79, no. 3, pp. 159–165, 1991.

[65] M. G. Cersosimo, G. B. Raina, C. Pecci, A. Pellene, C. R. Calandra, C. Gutiérrez, F. E. Micheli, and E. E. Benarroch, "Gastrointestinal manifestations in Parkinson's disease: Prevalence and occurrence before motor symptoms," *Journal of Neurology*, vol. 260, no. 5, p. 1332, 2013.

[66] M. A. Thenganatt and J. Jankovic, "Parkinson disease subtypes," *JAMA Neurology*, vol. 71, no. 4, pp. 499–504, 2014.

[67] L. O. Ramig, C. Fox, and S. Sapir, "Speech treatment for Parkinson's disease," *Expert Review of Neurotherapeutics*, vol. 8, no. 2, pp. 297–309, 2008.

[68] L. Hartelius and P. Svensson, "Speech and swallowing symptoms associated with Parkinson's disease and multiple sclerosis: A survey," *Folia Phoniatrica et Logopaedica*, vol. 46, no. 1, pp. 9–17, 1994.

[69] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients," *Journal of Speech and Hearing Disorders*, vol. 43, no. 1, pp. 47–57, 1978.

[70] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease," *Behavioural Neurology*, vol. 11, no. 3, pp. 131–137, 1999.

[71] M. J. Barrett, S. A. Wylie, M. B. Harrison, and G. F. Wooten, "Handedness and motor symptom asymmetry in Parkinson's disease," *Journal of Neurology, Neurosurgery & Psychiatry*, 2010.

[72] C. J. Cellucci, A. M. Albano, and P. E. Rapp, "Statistical validation of mutual information calculations: Comparison of alternative numerical algorithms," *Physical Review E*, vol. 71, no. 6, p. 066208, 2005.

[73] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, *et al.*, "Recent advances in deep learning for speech research at microsoft," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 8604–8608, IEEE, 2013.

[74] S. Takaki and J. Yamagishi, "A deep auto-encoder based low-dimensional feature extraction from fft spectral envelopes for statistical parametric speech synthesis," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5535–5539, IEEE, 2016.

[75] H. Herzel, D. Berry, I. R. Titze, and M. Saleh, "Analysis of vocal disorders with methods from nonlinear dynamics," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 5, pp. 1008–1019, 1994.

[76] M. Little, "Biomechanically informed, nonlinear speech signal processing.," *Diss. University of Oxford*, 2007.

[77] I. R. Titze, "Nonlinear source-filter coupling in phonation," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 1902–1915, 2008.

[78] I. Titze, "Summary statement: Workshop on acoustic voice analysis," *National Center for Voice and Speech*, pp. 26–30, 1995.

[79] K. M. Rosen, R. D. Kent, A. L. Delaney, and J. R. Duffy, "Parametric quantitative acoustic analysis of conversation produced by speakers with dysarthria and healthy speakers," *Journal of Speech, Language, and Hearing Research*, vol. 49, no. 2, pp. 395–411, 2006.

[80] S. U. Hahm and J. U. Wang, "Parkinson's condition estimation using speech acoustic and inversely mapped articulatory data," in *INTERSPEECH*, vol. 2015, International Speech and Communication Association, 2015.

[81] T. Grósz, R. Busa-Fekete, G. Gosztolya, and L. Tóth, "Assessing the degree of nativeness and Parkinson's condition using gaussian processes and deep rectifier neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[82] J. R. Williamson, T. F. Quatieri, B. S. Helfer, J. Perricone, S. S. Ghosh, G. Ciccarelli, and D. D. Mehta, "Segment-dependent dynamics in predicting Parkinson's disease," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[83] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, J. Vargas-Bonilla, J. D. Arias-Londoño, and E. Nöth, "Automatic detection of Parkinson's disease from continuous speech recorded in non-controlled noise conditions," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[84] J.-C. Wang, C.-H. Yang, J.-F. Wang, and H.-P. Lee, "Robust speaker identification and verification," *IEEE Computational Intelligence Magazine*, vol. 2, no. 2, pp. 52–59, 2007.

[85] Y. Horii, "Jitter and shimmer differences among sustained vowel phonations," *Journal of Speech, Language, and Hearing Research*, vol. 25, no. 1, pp. 12–4, 1982.

[86] J. Schoentgen and R. De Guchteneere, "Time series analysis of jitter," *Journal of Phonetics*, vol. 23, no. 1, pp. 189–201, 1995.

[87] E. Yumoto, "The quantitative evaluation of hoarseness: A new harmonics to noise ratio method," *Archives of Otolaryngology*, vol. 109, no. 1, pp. 48–52, 1983.

[88] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal-to-noise excitation ratio – a new measure for describing pathological voices," *Acta Acustica united with Acustica*, vol. 83, no. 4, pp. 700–706, 1997.

[89] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gómez-Vilda, M. Blanco-Velasco, and F. Cruz-Roldán, "The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders," *Journal of Voice*, vol. 24, no. 1, pp. 47–56, 2010.

[90] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83.*, vol. 8, pp. 93–96, IEEE, 1983.

[91] A. Tsanas, "Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning," *Diss. University of Oxford*, 2012.

[92] A. Kounoudes, P. A. Naylor, and M. Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, pp. I–349, IEEE, 2002.

[93] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern Recognition and Artificial Intelligence*, vol. 116, pp. 374–388, 1976.

[94] A. A. Dibazar and S. Narayanan, "A system for automatic detection of pathological speech," in *Conference Signals, Systems, and Computers, Asilomar, CA*, 2002.

[95] V. Tyagi and C. Wellekens, "On desensitizing the mel-cepstrum to spurious spectral components for robust speech recognition," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 1, pp. I–529, IEEE, 2005.

[96] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, *et al.*, "The Interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Fourteenth Annual Conference of the International Speech Communication Association*, 2013.

[97] S. K. Van Den Eeden, C. M. Tanner, A. L. Bernstein, R. D. Fross, A. Leimpeter, D. A. Bloch, and L. M. Nelson, "Incidence of Parkinson's disease: Variation by age, gender, and race/ethnicity," *American Journal of Epidemiology*, vol. 157, no. 11, pp. 1015–1022, 2003.

[98] N. Dahodwala, A. Siderowf, M. Xie, E. Noll, M. Stern, and D. S. Mandell, "Racial differences in the diagnosis of Parkinson's disease," *Movement Disorders*, vol. 24, no. 8, pp. 1200–1205, 2009.

[99] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *Biomedical Engineering Online*, vol. 6, no. 1, p. 23, 2007.

[100] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, "Mosaic organization of DNA nucleotides," *Physical Review E*, vol. 49, no. 2, p. 1685, 1994.

[101] M. Mancini, P. Carlson-Kuhta, C. Zampieri, J. G. Nutt, L. Chiari, and F. B. Horak, "Postural sway as a marker of progression in Parkinson's disease: A pilot longitudinal study," *Gait & posture*, vol. 36, no. 3, pp. 471–476, 2012.

[102] N. Giladi, D. McMahon, S. Przedborski, E. Flaster, S. Guillory, V. Kostic, and S. Fahn, "Motor blocks in Parkinson's disease," *Neurology*, vol. 42, no. 2, pp. 333–342, 1992.

[103] S. Kimmeskamp and E. M. Hennig, "Heel to toe motion characteristics in Parkinson's patients during free walking," *Clinical Biomechanics*, vol. 16, no. 9, pp. 806–812, 2001.

[104] E. M. Diaz and A. L. M. Gonzalez, "Step detector and step length estimator for an inertial pocket navigation system," in *Indoor Positioning and Indoor Navigation (IPIN), 2014 International Conference on*, pp. 105–110, IEEE, 2014.

[105] B. M. Bot, C. Suver, E. C. Neto, M. Kellen, A. Klein, C. Bare, M. Doerr, A. Pratap, J. Wilbanks, E. R. Dorsey, *et al.*, "The mPower study, Parkinson disease mobile data collected using ResearchKit," *Scientific Data*, vol. 3, 2016.

[106] M. Elhoushi, J. Georgy, A. Noureldin, and M. J. Korenberg, "A survey on approaches of motion mode recognition using sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 7, pp. 1662–1686, 2017.

[107] M. Li, V. Rozgica, G. Thatte, S. Lee, A. Emken, M. Annavaram, U. Mitra, D. Spruijt-Metz, and S. Narayanan, "Multimodal physical activity recognition by fusing temporal and cepstral information," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 18, no. 4, pp. 369–380, 2010.

[108] E. C. Neto, T. M. Perumal, A. Pratap, B. M. Bot, L. Mangravite, and L. Omberg, "On the analysis of personalized medication response and classification of case vs control patients in mobile health studies: The mPower case study," *arXiv preprint arXiv:1706.09574*, 2017.

[109] J. Jeong, "EEG dynamics in patients with Alzheimer's disease," *Clinical Neurophysiology*, vol. 115, no. 7, pp. 1490–1505, 2004.

[110] B. Jelles, J. Van Birgelen, J. Slaets, R. Hekster, E. Jonkman, and C. Stam, "Decrease of non-linear structure in the EEG of alzheimer patients compared to healthy controls," *Clinical Neurophysiology*, vol. 110, no. 7, pp. 1159–1167, 1999.

[111] R. Hegger, H. Kantz, and L. Matassini, "Denoising human speech signals using chaoslike features," *Physical Review Letters*, vol. 84, no. 14, p. 3197, 2000.

[112] I. Steinecke and H. Herzel, "Bifurcations in an asymmetric vocal-fold model," *The Journal of the Acoustical Society of America*, vol. 97, no. 3, pp. 1874–1884, 1995.

[113] J.-P. Eckmann, S. O. Kamphorst, and D. Ruelle, "Recurrence plots of dynamical systems," *EPL (Europhysics Letters)*, vol. 4, no. 9, p. 973, 1987.

[114] M. T. Rosenstein, J. J. Collins, and C. J. De Luca, "A practical method for calculating largest Lyapunov exponents from small data sets," *Physica D: Nonlinear Phenomena*, vol. 65, no. 1-2, pp. 117–134, 1993.

[115] A. Babloyantz, J. Salazar, and C. Nicolis, "Evidence of chaotic dynamics of brain activity during the sleep cycle," *Physics Letters A*, vol. 111, no. 3, pp. 152–156, 1985.

[116] N. F. Güler, E. D. Übeyli, and I. Güler, "Recurrent neural networks employing Lyapunov exponents for EEG signals classification," *Expert systems with applications*, vol. 29, no. 3, pp. 506–514, 2005.

[117] M. Banbrook, S. McLaughlin, and I. Mann, "Speech characterization and synthesis by nonlinear methods," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 1, pp. 1–17, 1999.

[118] I. Kokkinos and P. Maragos, "Nonlinear speech analysis using models for chaotic systems," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1098–1109, 2005.

[119] J. B. Dingwell and J. P. Cusumano, "Nonlinear time series analysis of normal and pathological human walking," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 10, no. 4, pp. 848–863, 2000.

[120] J. D. Howcroft, E. D. Lemaire, J. Kofman, and W. E. McIlroy, "Analysis of dual-task elderly gait using wearable plantar-pressure insoles and accelerometer," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pp. 5003–5006, IEEE, 2014.

[121] K. Liu, H. Wang, J. Xiao, and Z. Taha, "Analysis of human standing balance by largest lyapunov exponent," *Computational Intelligence and Neuroscience*, vol. 2015, p. 20, 2015.

[122] B. B. Mandelbrot, "How long is the coast of Britain?," *Science*, vol. 156, no. 3775, pp. 636–638, 1967.

[123] A. Accardo, M. Affinito, M. Carrozzi, and F. Bouquet, "Use of the fractal dimension for the analysis of electroencephalographic time series," *Biological Cybernetics*, vol. 77, no. 5, pp. 339–350, 1997.

[124] A. Babloyantz and A. Destexhe, "Low-dimensional chaos in an instance of epilepsy," *Proceedings of the National Academy of Sciences*, vol. 83, no. 10, pp. 3513–3517, 1986.

[125] C.-K. Peng, J. M. Hausdorff, A. Goldberger, and J. Walleczek, "Fractal mechanisms in neuronal control: Human heartbeat and gait dynamics in health and disease," in *Nonlinear Dynamics, Self-organization and Biomedicine*, pp. 66–96, Cambridge University Press, 2000.

[126] T. L. Doyle, E. L. Dugan, B. Humphries, and R. U. Newton, "Discriminating between elderly and young using a fractal dimension analysis of centre of pressure," *International journal of medical sciences*, vol. 1, no. 1, p. 11, 2004.

[127] Y. Manabe, E. Honda, Y. Shiro, K. Kenichi, I. Kohira, K. Kashihara, T. Shohmori, and K. Abe, "Fractal dimension analysis of static stabilometry in Parkinson's disease and spinocerebellar ataxia," *Neurological research*, vol. 23, no. 4, pp. 397–404, 2001.

[128] J. W. Baszczyk and W. Klonowski, "Postural stability and fractal dynamics," *Acta Neurobiol. Exp*, vol. 61, pp. 105–112, 2001.

[129] R. Esteller, G. Vachtsevanos, J. Echauz, and B. Litt, "A comparison of waveform fractal dimension algorithms," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 48, no. 2, pp. 177–183, 2001.

[130] C. Gourieroux and A. Monfort, *Statistics and econometric models*, vol. 1. Cambridge University Press, 1995.

[131] M. Martin, J. Perez, and A. Plastino, "Fisher information and nonlinear dynamics," *Physica A: Statistical Mechanics and its Applications*, vol. 291, no. 1, pp. 523–532, 2001.

[132] C. J. James and D. Lowe, "Extracting multisource brain activity from a single electromagnetic channel," *Artificial Intelligence in Medicine*, vol. 28, no. 1, pp. 89–104, 2003.

[133] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, pp. H2039–H2049, 2000.

[134] S. M. Pincus, I. M. Gladstone, and R. A. Ehrenkranz, "A regularity statistic for medical data analysis," *Journal of Clinical Monitoring and Computing*, vol. 7, no. 4, pp. 335–345, 1991.

[135] M. Costa, A. L. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of biological signals," *Physical Review E*, vol. 71, no. 2, p. 021906, 2005.

[136] M. Costa, C.-K. Peng, A. L. Goldberger, and J. M. Hausdorff, "Multiscale entropy analysis of human gait dynamics," *Physica A: Statistical Mechanics and its Applications*, vol. 330, no. 1, pp. 53–60, 2003.

[137] H. M. Al-Angari and A. V. Sahakian, "Use of sample entropy approach to study heart rate variability in obstructive sleep apnea syndrome," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 10, pp. 1900–1904, 2007.

[138] G. K. Wallace, "The JPEG still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.

[139] S. J. Roberts, W. Penny, and I. Rezek, "Temporal and spatial complexity measures for electroencephalogram based brain-computer interfacing," *Medical and Biological Engineering and Computing*, vol. 37, no. 1, pp. 93–98, 1999.

[140] H. E. Hurst, "Long-term storage capacity of reservoirs," *Transactions of the American Society of Civil Engineers*, vol. 116, pp. 770–808, 1951.

[141] A. Facchini, H. Kantz, and E. Tiezzi, "Recurrence plot analysis of nonstationary data: The understanding of curved patterns," *Physical Review E*, vol. 72, no. 2, p. 021915, 2005.

[142] R. Bryce and K. Sprague, "Revisiting detrended fluctuation analysis," *Scientific Reports, Available at: https://www.nature.com/articles/srep*00315, *Last accessed Oct 2017*, vol. 2, 2012.

[143] T. Gneiting and M. Schlather, "Stochastic models that separate fractal dimension and the hurst effect," *SIAM review*, vol. 46, no. 2, pp. 269–282, 2004.

[144] M. Duarte and V. M. Zatsiorsky, "On the fractal properties of natural human standing," *Neuroscience letters*, vol. 283, no. 3, pp. 173–176, 2000.

[145] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pp. 381–384, IEEE, 1990.

[146] A. Tsanas, M. Zañartu, M. A. Little, C. Fox, L. O. Ramig, and G. D. Clifford, "Robust fundamental frequency estimation in sustained vowels: Detailed algorithmic comparisons and information fusion with adaptive Kalman filtering," *The Journal of the Acoustical Society of America*, vol. 135, no. 5, pp. 2885–2901, 2014.

[147] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences*, vol. 17, pp. 97–110, Amsterdam, 1993.

[148] D. O'Shaughnessy, "Linear predictive coding," *IEEE Potentials*, vol. 7, no. 1, pp. 29–32, 1988.

[149] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 454, pp. 903–995, The Royal Society, 1998.

[150] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," *Journal of the Royal Society Interface*, vol. 8, no. 59, pp. 842–855, 2011.

[151] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[152] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: What speech, music, and sound have in common," *Frontiers in Psychology*, vol. 4, 2013.

[153] W. Hamäläinen, M. Järvinen, P. Martiskainen, and J. Mononen, "Jerk-based feature extraction for robust activity recognition from acceleration data," in *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*, pp. 831–836, IEEE, 2011.

[154] M. Duarte and V. M. Zatsiorsky, "Effects of body lean and visual information on the equilibrium maintenance during stance," *Experimental Brain Research*, vol. 146, no. 1, pp. 60–69, 2002.

[155] B. Hjorth, "EEG analysis based on time domain properties," *Electroencephalography and Clinical Neurophysiology*, vol. 29, no. 3, pp. 306–310, 1970.

[156] A. J. A. Majumder, F. Rahman, I. Zerin, W. Ebel Jr, and S. I. Ahamed, "iPrevention: Towards a novel real-time smartphone-based fall prevention system," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pp. 513–518, ACM, 2013.

[157] J.-P. Eckmann, S. O. Kamphorst, D. Ruelle, and S. Ciliberto, "Liapunov exponents from time series," *Physical Review A*, vol. 34, no. 6, p. 4971, 1986.

[158] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Physica D: Nonlinear Phenomena*, vol. 31, no. 2, pp. 277–283, 1988.

[159] A. Petrosian, "Kolmogorov complexity of finite sequences and recognition of different preictal EEG patterns," in *Computer-Based Medical Systems, 1995., Proceedings of the Eighth IEEE Symposium on*, pp. 212–217, IEEE, 1995.

[160] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and understanding recurrent networks," *arXiv preprint arXiv:1506.02078*, 2015.

[161] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Computation*, vol. 4, no. 1, pp. 1–58, 1992.

[162] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[163] J. H. Friedman, "On bias, variance, 0/1—loss, and the curse-of-dimensionality," *Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 55–77, 1997.

[164] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[165] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*, vol. 1. MIT press Cambridge, 2006.

[166] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[167] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[168] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?," in *International Conference on Machine Learning and Data Mining*, pp. 154–168, Springer, 2012.

[169] V. Vapnik and A. Chervonenkis, "A note on one class of perceptrons," *Automation and Remote Control*, vol. 25, no. 1, p. 103, 1964.

[170] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[171] A. Navia-Vázquez and E. Parrado-Hernández, "Support vector machine interpretation," *Neurocomputing*, vol. 69, no. 13, pp. 1754–1759, 2006.

[172] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, pp. 2951–2959, 2012.

[173] S. Kramer, N. Lavrač, and P. Flach, "Propositionalization approaches to relational data mining," in *Relational data mining*, pp. 262–291, Springer, 2001.

[174] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[175] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain.," *Psychological Review*, vol. 65, no. 6, p. 386, 1958.

[176] M. Minsky and S. Papert, *Perceptrons.* MIT press, 1969.

[177] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[178] P. J. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral sciences," *Doctoral Dissertation, Applied Mathematics, Harvard University, MA*, 1974.

[179] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," tech. rep., California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[180] D. P. Bertsekas, A. Nedi, A. E. Ozdaglar, *et al.*, *Convex analysis and optimization*. Athena Scientific, 2003.

[181] S. Hochreiter, "Untersuchungen zu dynamischen neuronalen netzen," *Diploma, Technische Universität München*, vol. 91, 1991.

[182] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, pp. 448–456, 2015.

[183] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

[184] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *arXiv preprint arXiv:1506.06579*, 2015.

[185] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[186] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," *arXiv preprint arXiv:1511.06448*, 2015.

[187] F. J. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[188] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, *et al.*, "Wide & deep learning for recommender systems," in *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, pp. 7–10, ACM, 2016.

[189] R. Eldan and O. Shamir, "The power of depth for feedforward neural networks," in *Conference on Learning Theory*, pp. 907–940, 2016.

[190] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[191] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus, "Regularization of neural networks using dropconnect," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 1058–1066, 2013.

[192] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814, 2010.

[193] Y. Bengio, D.-H. Lee, J. Bornschein, T. Mesnard, and Z. Lin, "Towards biologically plausible deep learning," *arXiv preprint arXiv:1502.04156*, 2015.

[194] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer Vision*, pp. 1026–1034, 2015.

[195] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," *arXiv preprint arXiv:1302.4389*, 2013.

[196] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[197] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

[198] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[199] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $o(1/k2)$," in *Soviet Mathematics Doklady*, vol. 27, pp. 372–376, 1983.

[200] T. Dozat, "Incorporating nesterov momentum into adam," *ICLR 2016 Workshop*, 2016.

[201] F.-F. Li, J. Johnson, and S. Yeung, *Cs231n: Convolutional neural networks for visual recognition*. Stanford University, 2017.

[202] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[203] J. Loughrey and P. Cunningham, "Overfitting in wrapper-based feature subset selection: The harder you try the worse it gets," *Research and Development in Intelligent Systems XXI*, pp. 33–43, 2005.

[204] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *arXiv preprint arXiv:1601.07996*, 2016.

[205] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1-2, pp. 23–69, 2003.

[206] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Advances in Neural Information Processing Systems*, pp. 507–514, 2006.

[207] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection," *Journal of Machine Learning Research*, vol. 13, no. Jan, pp. 27–66, 2012.

[208] A. Jakulin, *Machine learning based on attribute interactions*. PhD thesis, Univerza v Ljubljani, 2005.

[209] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 02, pp. 185–205, 2005.

[210] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu, *Advancing feature selection research – ASU Feature Selection Repository*. Arizona State University, 2010.

[211] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint ℓ2, 1-norms minimization," in *Advances in Neural Information Processing Systems*, pp. 1813–1821, 2010.

[212] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient ℓ2, 1-norm minimization," in *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pp. 339–348, AUAI Press, 2009.

[213] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[214] J. N. Rouder, P. L. Speckman, D. Sun, R. D. Morey, and G. Iverson, "Bayesian t tests for accepting and rejecting the null hypothesis," *Psychonomic Bulletin & Review*, vol. 16, no. 2, pp. 225–237, 2009.

[215] S. Arlot, A. Celisse, *et al.*, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010.

[216] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, vol. 14, pp. 1137–1145, Stanford, CA, 1995.

[217] L. Breiman and P. Spector, "Submodel selection and evaluation in regression. the x-random case," *International Statistical Review*, pp. 291–319, 1992.

[218] R. R. Bouckaert, "Choosing between two learning algorithms based on calibrated tests," in *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 51–58, 2003.

[219] F. J. Provost, T. Fawcett, *et al.*, "Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions.," in *KDD*, vol. 97, pp. 43–48, 1997.

[220] C. X. Ling, J. Huang, and H. Zhang, "Auc: A statistically consistent and more discriminating measure than accuracy," in *IJCAI*, vol. 3, pp. 519–524, 2003.

[221] A. T. Peterson, M. Papeş, and J. Soberón, "Rethinking receiver operating characteristic analysis applications in ecological niche modeling," *Ecological Modelling*, vol. 213, no. 1, pp. 63–72, 2008.

[222] J. M. Lobo, A. Jiménez-Valverde, and R. Real, "Auc: A misleading measure of the performance of predictive distribution models," *Global Ecology and Biogeography*, vol. 17, no. 2, pp. 145–151, 2008.

[223] D. J. Hand, "Measuring classifier performance: A coherent alternative to the area under the roc curve," *Machine Learning*, vol. 77, no. 1, pp. 103–123, 2009.

[224] P. J. Easterbrook, R. Gopalan, J. Berlin, and D. R. Matthews, "Publication bias in clinical research," *The Lancet*, vol. 337, no. 8746, pp. 867–872, 1991.

[225] O. S. Collaboration *et al.*, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, p. aac4716, 2015.

[226] R. L. Wasserstein and N. A. Lazar, "The asa's statement on p-values: Context, process, and purpose," *The American Statistician*, vol. 70, no. 2, pp. 129–133, 2016.

[227] B. Aczel, B. Palfi, and B. Szaszi, "Estimating the evidential value of significant results in psychological science," *PloS one*, vol. 12, no. 8, p. e0182651, 2017.

[228] H. Jeffreys, *The theory of probability*. Oxford University Press, 1998.

[229] K. P. Burnham and D. R. Anderson, "Multimodel inference: Understanding AIC and BIC in model selection," *Sociological Methods & Research*, vol. 33, no. 2, pp. 261–304, 2004.

[230] Y. Zhang, R. Li, and C.-L. Tsai, "Regularization parameter selections via generalized information criterion," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 312–323, 2010.

[231] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *The Annals of Statistics*, pp. 416–431, 1983.

[232] "Michael J. Fox Foundation launches $10,000 Parkinson's data challenge." Available at: https://www.michaeljfox.org/foundation/publication-detail.html?id=325, Last accessed Nov 2017, 2013. Michael J. Fox Foundation.

[233] N. Quinn, P. Critchley, and C. D. Marsden, "Young onset Parkinson's disease," *Movement Disorders*, vol. 2, no. 2, pp. 73–91, 1987.

[234] L. I. Golbe, "Young-onset Parkinson's disease a clinical review," *Neurology*, vol. 41, no. 2 Part 1, pp. 168–168, 1991.

[235] D. R. Karger, S. Oh, and D. Shah, "Iterative learning for reliable crowdsourcing systems," in *Advances in Neural Information Processing Systems*, pp. 1953–1961, 2011.

[236] M. N. Schmidt, J. Larsen, and F.-T. Hsiao, "Wind noise reduction using non-negative sparse coding," in *Machine Learning for Signal Processing, 2007 IEEE Workshop on*, pp. 431–436, IEEE, 2007.

[237] F. Ichikawa, J. Chipchase, and R. Grignani, "Where's the phone? a study of mobile phone location in public spaces," *IEE Mobility Conference 2005*, p. 142, 2005.

[238] R. Soames and J. Atha, "The spectral characteristics of postural sway behaviour," *European Journal of Applied Physiology and Occupational Physiology*, vol. 49, no. 2, pp. 169–177, 1982.

[239] A. V. Oppenheim and R. W. Schafer, "From frequency to quefrency: A history of the cepstrum," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 95–106, 2004.

[240] S. Ravindran, D. V. Anderson, and M. Slaney, "Improving the noise-robustness of mel-frequency cepstral coefficients for speech processing," *Reconstruction*, vol. 12, p. 14, 2006.

[241] "Parkinson's disease digital biomarker DREAM challenge." Available at: http://dream challenges.org/project/parkinsons-disease-digital-biomarker-dream-challenge/, Last accessed Oct 2017, 2017. DREAM Challenges.

[242] "Feature extraction toolkit for mPower modules." Available at: https://github.com/Sage-Bionetworks/mpowertools/blob/master/FeatureDefinitions.md, Last accessed Oct 2017, 2017. Sage Bionetworks.

[243] T. Ruf, "The Lomb-Scargle periodogram in biological rhythm research: Analysis of incomplete and unequally spaced time-series," *Biological Rhythm Research*, vol. 30, no. 2, pp. 178–201, 1999.

[244] J. W. Lance, R. S. Schwab, and E. A. Peterson, "Action tremor and the cogwheel phenomenon in Parkinson's disease," *Brain*, vol. 86, no. 1, pp. 95–110, 1963.

[245] A. Camacho, *SWIPE: A sawtooth waveform inspired pitch estimator for speech and music*. University of Florida Gainesville, 2007.

[246] A. Benba, A. Jilbab, and A. Hammouch, "Voice analysis for detecting persons with Parkinson's disease using MFCC and VQ," in *The 2014 International Conference on Circuits, Systems and Signal Processing*, pp. 23–25, 2014.

[247] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4580–4584, IEEE, 2015.

[248] A. Kumar and S. Mullick, "Nonlinear dynamical analysis of speech," *The Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 615–629, 1996.

[249] S. S. Narayanan and A. A. Alwan, "A nonlinear dynamical systems analysis of fricative consonants," *The Journal of the Acoustical Society of America*, vol. 97, no. 4, pp. 2511–2524, 1995.

[250] A. J. Smola and B. Schölkopf, "Sparse greedy matrix approximation for machine learning," in *Seventeenth International Conference on Machine Learning*, pp. 911–918, Morgan Kaufmann, 2000.

[251] L. Csató and M. Opper, "Sparse on-line gaussian processes," *Neural Computation*, vol. 14, no. 3, pp. 641–668, 2002.

[252] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European Conference on Computational Learning Theory*, pp. 23–37, Springer, 1995.

[253] D. W. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *Journal of Artificial Intelligence Research (JAIR)*, vol. 11, pp. 169–198, 1999.

[254] L. Breiman, "Stacked regressions," *Machine learning*, vol. 24, no. 1, pp. 49–64, 1996.

[255] J. Sill, G. Takács, L. Mackey, and D. Lin, "Feature-weighted linear stacking," *arXiv preprint arXiv:0911.0460*, 2009.

[256] Y. Koren, "The Bellkor solution to the Netflix grand prize," *Netflix Prize Documentation*, vol. 81, pp. 1–10, 2009.

[257] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[258] I. W. Tsang, J. T. Kwok, and P.-M. Cheung, "Core vector machines: Fast SVM training on very large data sets," *Journal of Machine Learning Research*, vol. 6, no. Apr, pp. 363–392, 2005.

[259] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[260] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[261] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[262] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[263] C. E. Douglas and F. A. Michael, "On distribution-free multiple comparisons in the one-way analysis of variance," *Communications in Statistics-Theory and Methods*, vol. 20, no. 1, pp. 127–139, 1991.

[264] R. A. Fisher, *The design of experiments*. Oliver And Boyd; Edinburgh; London, 1937.

[265] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins, "The aligned rank transform for nonparametric factorial analyses using only anova procedures," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 143–146, ACM, 2011.

[266] C. Fan and D. Zhang, "Rank repeated measures analysis of covariance," *Communications in Statistics-Theory and Methods*, vol. 46, no. 3, pp. 1158–1183, 2017.

[267] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proceedings of the IEEE*, vol. 81, no. 10, pp. 1385–1422, 1993.

[268] S. Diamond, C. Markham, M. Hoehn, F. McDowell, and M. Muenter, "An examination of male-female differences in progression and mortality of Parkinson's disease," *Neurology*, vol. 40, no. 5, pp. 763–763, 1990.

[269] B. Scott, A. Borgman, H. Engler, B. Johnels, and S. Aquilonius, "Gender differences in Parkinson's disease symptom profile," *Acta Neurologica Scandinavica*, vol. 102, no. 1, pp. 37–43, 2000.

[270] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 1459–1462, ACM, 2010.

[271] D. J. Hermes, "Measurement of pitch by subharmonic summation," *The Journal of the Acoustical Society of America*, vol. 83, no. 1, pp. 257–264, 1988.

[272] Q. Geissmann, "PyREM: Package for sleep staging from EEG data." Available at: https://github.com/gilestrolab/pyrem, Last accessed Oct 2017, 2017. GitHub repository.

[273] F. S. Bao, X. Liu, and C. Zhang, "PyEEG: An open source python module for EEG/MEG feature extraction," *Computational Intelligence and Neuroscience*, vol. 2011, 2011.

[274] H. Harrison, "Phase space reconstruction: pypsr." Available at: https://github.com/hsharrison/pypsr, Last accessed Oct 2017, 2017. GitHub repository.

[275] C. Schölzel *et al.*, "Nonlinear measures for dynamical systems (nolds)." Available at: https://github.com/CSchoel/nolds, Last accessed Oct 2017, 2017. GitHub repository.

[276] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[277] "GPy: A Gaussian process framework in python." Available at: http://github.com/SheffieldML/GPy, Last accessed Oct 2017, since 2012. Github Repository.

[278] J. Bergstra, D. Yamins, and D. D. Cox, "Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms," in *Proceedings of the 12th Python in Science Conference*, pp. 13–20, 2013.

[279] R. Fukatani *et al.*, "Library for machine learning stacking generalization." Available at: https://github.com/fukatani/stacked_generalization, Last accessed Oct 2017, 2016. GitHub repository.

[280] F. Chollet *et al.*, "Keras." Available at: https://github.com/fchollet/keras, Last accessed Oct 2017, 2015. GitHub repository.

[281] P. J. Huber *et al.*, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[282] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2625–2634, 2015.

[283] R. Djaldetti, I. Ziv, and E. Melamed, "The mystery of motor asymmetry in Parkinson's disease," *The Lancet Neurology*, vol. 5, no. 9, pp. 796–802, 2006.

[284] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.