Australian National University

Machine Learning in Parkinson's Disease Diagnosis

Bachelor's Thesis

Max Wang

Supervised by:
Dr Deborah Apthorp
Dr Hanna Suominen

Contents

Abstract

Parkinson's disease (PD) is a degenerative neurological disorder, affecting around 1% of the population by the age of 70. There is currently no objective test for PD and studies suggest expert misdiagnosis rates of up to 34%. Hence, there is interest in investigating if machine learning can provide a more reliable and objective diagnosis.

Current machine learning literature test a model's ability to differentiate between already diagnosed PD and control subjects. This setup does not mirror real-life diagnosis as the primary difficulty neurologists face is excluding disorders with similar symptoms and individuals exhibiting minimal symptoms. Most studies are also based on small (<50) datasets which suffer from a tendency to bias and overfitting due to Freedman's paradox. A large dataset of individuals pre-diagnosis to confirmed diagnosis would be optimally be used to assess machine learning in PD.

This thesis investigates the applicability of machine learning in PD diagnosis by testing a model's ability to diagnose PD using symptoms unobservable by a neurologist. Current literature is replicated on the much larger 6,000 participant mPower dataset, consisting of crowdsourced recordings from smartphone sensors. Results showed that the simple models used in current literature are insufficient for a reliable diagnosis using the large and noisy mPower data. More powerful and robust models were developed by consolidating recent ideas EEG and non-linear signal processing, deep learning and computer vision, resulting in improvement from XX% to YY% accuracy.

Results suggest that machine learning can offer a valuable source of information for experts as these models quantify symptoms differently from experts. –more discussion here–

Introduction

This thesis has been written for all audiences, however a background in machine learning may be useful to understand and follow some assumptions in the methodology. The work spans multiple disciplines and I have opted to summarise these fields concisely and provide references to seminal or well-written papers in the area for the reader interested in a more in-depth understanding. These papers will also provide the mathematical formulation of the signal processing and machine learning models which have been abstracted in favour for the intuition behind them.

Throughout the thesis I have used highlights and footnotes 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.

The thesis is organised as follows:

Chapter 1 summarises Parkinson's Disease and relevant prior work in the field of feature extraction and machine learning. The remainder of the thesis is written with the assumption that the reader understands this background, enabling it to be very concise.

Chapter 2 begins with a literature review of relevant works in PD machine learning.

Major contributions are...

[1] *Footnotes* provide contextual background information

1 Background

Parkinson's disease (PD) is a major health problem, affecting around 1% of the population by age 70 [1]. PD 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 have not fully manifested [2]. Studies suggest that motor symptoms only manifest once 20-40% of dopamine[1] producing cells 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 commonly associated with tremor, only around 70% of patients experience resting tremor [73].

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 [4, 5, 6] . Thus, an accurate early diagnosis is crucial to ensuring a higher quality of life later in life.

PD is currently diagnosed with a subjective test by a neurologist [13]. This test gener-

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

ally involves qualifying visible symptoms such as tremor and dysphonia, and assessing the patient's response to Levodopa[2]. As visible symptoms do not manifest until later stages, an early stage diagnosis is rare. There has been research in qualifying minor changes in speech [7, 8], sleep, olfactory and gastrointestinal behaviours [9, 10] as early markers of the disease.

The primary difficulty in diagnosis is differentiating from other Parkinsonism[3] disorders such as Multiple System Atrophy, Supranuclear Palsy and Essential Tremor [11]. Confirmation of diagnosis is generally only possible with an autopsy. As there is no definitive test and symptoms resemble other neurological disorders, misdiagnosis rates are high. Studies suggest a misdiagnosis rate is high, ranging from 9–34% depending on methodology [13, 2, 12].

Highlight* 1.1 *(Diagnosis). PD is diagnosed subjectively by a neurologist. As many disorders have similar symptoms, the misdiagnosis rate is high — up to 34%.

As there is no consensus for PD diagnosis, the search for a more objective measure for PD 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 [13, 14, 15] to discovering more quantifiable biomarkers such as gene expression [10, 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.

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 system. It will be examined in greater depth in section 1.3.

Machine learning presents an objective and low cost solution to diagnosing PD. There has been a large body of work in the field however the applicability all current work is limited due to the cost and difficulties associated with gathering a sizeable dataset. A majority of datasets used in literature consist of fewer than 40 subjects. Reported results are

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

[3] *Parkinsonism* movement disorders are those with similar symptoms to PD.

therefore prone to biases in the dataset, Freedman's paradox[4] [18] and overfitting on cross validation [19]. Thus, it is difficult to empirically compare results of different papers.

***Highlight* 1.2.** It is difficult to compare and evaluate work in PD machine learning due to variation in data and small dataset sizes.

There has been preliminary investigation in the applicability of machine learning in differentiating PD and other Parkinsonism disorders with promising results [20, 21]. However a majority of literature in the field uses machine learning to differentiate between PD and control subjects. This artificial setup simplifies the complexities involved in a neurologist's diagnosis for PD. As patients have already been diagnosed with PD, they likely exhibit noticable symptoms. Neurologists must perform diagnosis in early stages when symptoms are not evident and must consider the possibility of any number of causes for the symptoms. Current methods in machine learning only show its ability to detect the symptoms associated with PD and are difficult to relate to real world diagnosis.

***Highlight* 1.3.** Current research tasks machine learning to differentiate between PD and control subjects. This is a much simpler problem than what is faced by neurologists who have to rule out a number of other possibilities for symptoms.

To precisely compare the effectiveness of machine learning to neurologist diagnosis, a large *longitudinal dataset* following subjects pre-diagnosis to a confirmed diagnosis would be required. Such a dataset would be very costly and logistically difficult to collect. To advocate the collection of such a dataset, some evidence of machine learning's applicability to PD diagnosis will be required. This thesis will investigate methods of assessing machine learning's applicability to Parkinson's disease without such a dataset.

Another proposed application for machine learning for PD is telemonitoring [22, 23]. A patient's progression of PD is monitored with a scale, the most common being the MDS-UPDRS [24] which quantifies the extent of 27 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 PD patients visit a specialist every 4-6 months to track progression — this is costly and inconvenient. Machine learning offers the opportunity for patients to track their progress at home with their smartphone or other wearables [25]. Monitoring is a viable avenue for machine learning given current datasets, however will not be explored in this thesis as the primary focus is diagnosis.

[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.

Highlight* 1.4 *(UPDRS). The MDS-UPDRS [24] scale quantifies the extent of 44 motor and non-motor symptoms on a scale of 0-4. It is currently assessed by a neurologist.

The machine learning process for classification can generally be divided into two steps:

1. *Feature extraction* — From the raw input data from devices such as Accelerometers or microphones, features such as pitch and amplitude are quantified.
2. *Feature and Model selection* - A machine learning model is selected and its hyperparameters tweaked to to best suit the problem. The set of features used by the model is often reduced using feature selection [26] and dimensionality reduction [27, 28] due to the curse of dimensionality[5] [29].

1.2 Feature Extraction and Signal Processing

Feature extraction is the process of converting raw input data (*signals*) into meaningful numerical values[6]. For example, with sensors such as microphones, features such as pitch and volume may be extracted. Features should relate to the machine learning task as most machine learning models perform poorly as more unrelated features are added. Understanding raw input data and extracting useful features is a primary component in the field of *digital signal processing*.

Movement related problems are the primary manifestation of symptoms considered by a neurologist when diagnosing PD. Human vision is very advanced and captures and processes 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. However, our ability to differentiate between forms of irregular gait is more limited [11]. Although sensors such as IMUs can only a capture a fraction of the information of human eyes, it is possible that they are better at detecting the differences between forms of irregular gait [32].

***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 [24] scale,

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

[6] This is not required for all sensors data (e.g cameras and MRIs) however is generally required for any time-series sensor.

it has received a great deal of attention in machine learning. There is evidence that speech is one of the earliest indicators of PD [8] and there already exists a large body of work in the field of speech feature extraction [33]. Furthermore, microphones are able to capture a similar level of information as human ears — there is much less information loss compared to sensors used to measure movement[7].

Table 1.2 summarises prior work in feature extraction 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 diagnosing PD with movement or voice features and currently more research focuses on movement [30, 31].

Table 1.2: 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[8] [34, 35, 36] Smartphones [37, 38, 39]	Words and sentences [7, 53, 54]	Demographics
Postural Sway Force Plates [40] IMUs [41, 36]	Sustained vowel phonation [22, 55, 56]	UPDRS Patient Questionnaire [57, 58]
Gait Force Walkways [42, 43, 44] Video [43] Multiple IMUs [45, 46, 47]		Physical Changes Gene Expression [10, 59] MRI [60, 61, 62] EEG [63, 64] Olfactory [57] REM sleep [57, 58] Cerebrospinal Fluids [58] Gastrointestinal [65]
Handwriting [48, 49]		
Motion Capture [50]		
Tapping [51, 52]		

There is evidence that PD is heterogeneous and symptoms are present in distinct subsets [66, 67], however the underlying reasons not well understood. Studies have reported speech dysfunction present in 74-94% of patients with PD [68, 69, 70, 71]. Tremor is reported in 70% of patients [72] and Akinesia in 80% [73]. 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.

[7] Excepting motion capture, which we will cover in section 1.2.3.

[7] 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.

***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 extraction in each of these areas are equally as 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 X.X

Section 1.2.5 summarises all features that will be used in this paper. Feature extraction 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 extraction step and extract information from raw representations of data. These will be covered in section 1.3.2 and their applicability investigated in section X.X.

1.2.1 General Signal Processing

This thesis will focus on signal processing for time-series sensor data. The 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 of 100Hz (phones) to 4000Hz. *Noise* refers to deviations between the measured and true values, generally 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.

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 [137].

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. Given an accelerometer signal, the Fourier transform can be used to determine the amount of tremor in each frequency band - for example, PD tremor is exhibited as an increase in tremor in the 3.5-7Hz bands [34]. The *short time*[8] *Fourier transform* (STFT) is often used when modelling evolving signals such as those generated during speech and walking.

1.2.2 Voice

PD diagnosis with vocal features is a promising option for diagnosis with machine learning as microphones are readily available and capture a comparable level of information to ears. Little et al. (2009) [22] shows that audio from a phone is of sufficient quality to perform diagnosis with reasonable accuracy. This gives rise to the possibility of self diagnosis with a smartphone. However current feature extraction algorithms are sensitive to noise so robustness must be improved or bad recordings detected and filtered.

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 is expelled from the lungs builds pressure below the glottis. The imbalance of pressure below and above the glottis causes it to oscillate, producing sound. Muscles in the vocal folds enable adjustment 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 sounds producible.

The vocal tract comprises the components between the larynx and lips such as the

[8] *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.

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*[9] and *time invariant*[10]. This assumption greatly simplifies the the analysis of speech and grants the use of a rich set of tools in the well-understood field of linear, time invariant systems theory. However, recent works in analysing speech provide strong evidence that these linear assumptions do not hold for most speech signals [74, 75, 76]. Non-linear signal analysis is still an experimental field and most algorithms estimate the true properties of underlying phenomena . Even determining the fundamental frequency from sustained vowel phonation is an inexact science as evident in Tsanas et al. (2014) [77].

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.

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 neurologically and physically cause dysphonia is possible.

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

Speech Signal Processing

Parkinson's disease diagnosis with speech exists as two distinct subfields: quantifying dysarthria in spoken sentences and quantifying dysphonia with sustained vowels (e.g, 'aaaaah...'). To obtain a clinical level diagnosis, both dysphonia and dysarthria related features will likely have to be considered.

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

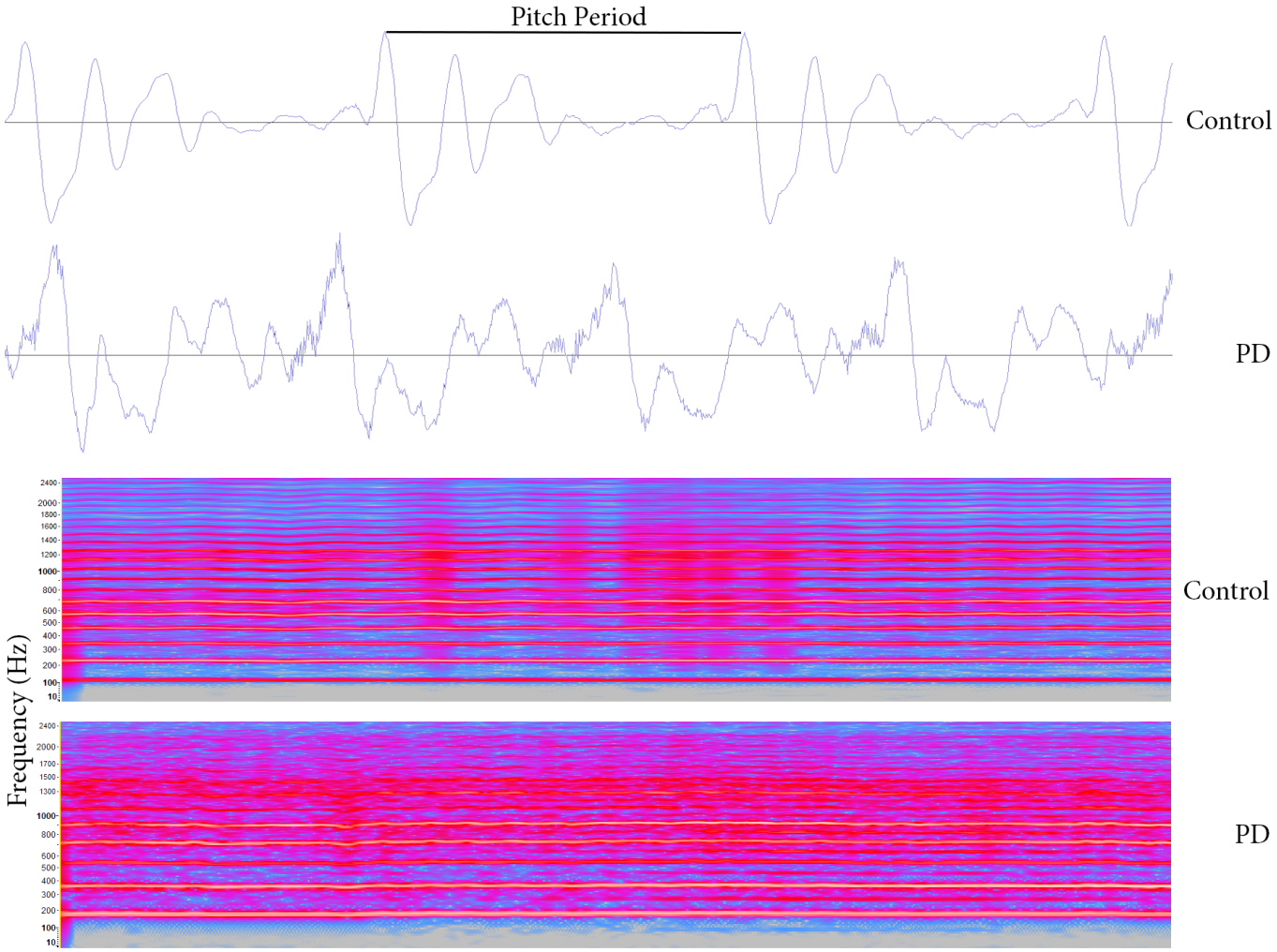


Fig. 1.1: A visualisation of prominent dypshonia in sustained vowel phonation on the time (above) and short-time frequency domain (below, Mel-scale [80]). Cases are generally not as extreme and the natural variation in voice makes differentiation a difficult task.

Although changes in speaking patterns (dysarthia) are very perceivable to human ears, features such as slurring or hesitation can only be roughly estimated 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. (2012) [7] investigates PD diagnosis on English and German sentences, however does not use any short-time features. Hazan et al. also observes that machine learning models trained on the English speakers do not generalize well to the German speakers and vice versa.

The Interspeech 2015 [53] competition featured a sub-challenge where the extent of PD dysarthria (as rated by the UPDRS) was to be 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 micro-

phone. The best performing papers in this sub-challenge only managed Pearson correlations of 0.4 to 0.64 against neurologist diagnosis [81, 82, 83].

However recent works point to evidence that speech can be a powerful predictor given better signal processing approaches. Vasqeuz et al. (2015) [84] is able to enhance noisy PD speech data using a technique proposed in Wang et al. (2007) [85] which decomposes speech into signal and noise subspaces. Orozco et al. (2015) [54] quantifies the transitions between voiced and unvoiced speech and presents significantly better results compared to using voiced speech as in prior works.

Sustained vowel phonations are the preferred method of quantifying dysphonia. Although features used in the general speech signal processing are applicable in dysphonia quantification, features developed specifically for dysphonia may be more robust as they are based on the non-linear model of speech production [22, 86]. Dysphonia specific features generally quantify the variation in each glottal cycle, relying on an an accurate fundamental frequency estimation algorithm [77].

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

Early dysphonia 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* [87, 88] the variation in amplitude (volume). The harmonics-to-noise ratio (*HNR*) [89] measures the amount of noise in a signal, which correlates with the 'hoarseness' or 'breathiness' from an incomplete closure of the glottis. The Glottal to Noise Excitation (*GNE*) ratio was introduced by Michaelis et.al (1997) [90] and is a more reliable measure of dysphonia than HNR [91].

More recently, methods used in non-linear dynamical systems[11] have been shown to be effective to dysphonia quantification. Detrended Fluctuation Analysis (*DFA*) was originally introduced by Peng et al. [92] as a measure of the autocorrelation of a signal. Little et al. (2007) [86] shows this correlates with the amount of turbulent airflow in speakers with dysphonia. Little et al. (2007) also proposes Recurrence Period Density Entropy (*RPDE*) which characterises the repetitiveness of a signal, which is generally lower for speakers with dysphonia due to jitter and shimmer. As the method does not rely on the

[11] 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. Dynamical systems are often viewed as a stochastic process for the purpose of analysis.

detection of the fundamental frequency it may be more robust for dysphonic speakers. Little et al. (2009) [22] builds upon RPDE to develop Pitch Period Entropy (*PPE*) which is a better measure of the impaired control of pitch experienced by PD patients.

Tsanas et al (2012) [93] extends GNE to develop Vocal Fold Excitation Ratios (*VFER*) and also introduces the Glottal Quotient (*GQ*). GQ measures the standard deviation of the duration when the glottis is opened and closed and is founded on the principles of the DYPSA [94] fundamental frequency estimation algorithm. We refer to Tsanas (2012) [95] for a more detailed summary of the signal processing involved.

Mel-Frequency Cepstral Coefficients (*MFCC*) have long been used for speech recognition [96], and have also shown promise in detecting dysphonia [97]. They are the most common and often the only feature used in speech recognition systems however lack interpretability and is very sensitive to noise [98]. There are also a variety of feature sets used in general speech classification, such as the 6,368 in the 2013 Intespeech ComParE set [99]. Although not all of these features may measure dysphonia, they are effective in fields such as speaker trait classification and may be useful in complex machine learning models. The incidence of PD varies based on age, gender and race [100, 101], and it is likely that dysphonia presents itself differently depending on speaker traits. We refer to Eyben (2015) [33] for a comprehensive description of these features as well as a summary of feature sets used in speech classification.

1.2.3 Movement

Despite a similar amount of literature existing in both movement and voice feature extraction, signal processing in the voice domain is more developed. Feature extraction for movement data diverges into a number of subfields, each developing different measurements for different sensors to quantify the extent of a movement disorder. Features are crafted specifically for dyskinesia[12] and akinesia[13] quantification. The signal processing techniques used in movement disorder quantification are basic compared to methods in voice and EEG.

People with PD exhibit increased tremor, particularly in the 3.5-7hz range [34] as well as distinct patterns of sway which can be quantified by recurrence analysis [36, 102]. These are best measured when the subject attempts to stand as still as possible. Both IMUs and

[12] *Dyskinesia* describes the presence of involuntary, often 'jerky' movements.
[13] *Akinesia* is the impairment of voluntary movement.

force plates are able to quantify this — IMUs have the advantage of being cheaper and more accessible however have lower resolution and may not be spatially accurate. There has not yet been a study comparing the information content of the two. The amount of tremor can be easily quantified with a Fourier transform, and recurrence can be quantified with general signal processing techniques such as DFA (see 1.2.2).

It is also possible to quantify gait with IMUs. Barth et al. (2011) [45] and Sijobert et al. (2015) [47] propose gait estimation algorithms for IMUs attached to the foot and shank respectively. It is also possible to estimate gait with handheld or in-pocket IMUs as done in Renaudin et al. (2012) [46] and Diaz and Gonzalez [103] respectively. However existing algorithms do not perform to the standards required to detect akinesia and are not very robust. Force Walkways and motion capture are more accurate alternatives for measuring gait however are more costly and only available in a clinical context.

Although expensive and difficult to setup, motion capture presents the possibility of completely quantifying all movement related components. However, feature extraction has not evolved to take advantage of the additional information and a significant amount of training data would likely be required to realise its full potential. Das et al. (2011) [50] uses motion capture on 4 PD and 2 control subjects, however does not explore any spatial features beyond what is provided by multiple accelerometers. Pose recognition in video is also an rapidly developing field which proposes similar capabilities to motion capture at a fraction of the cost. Current models are promising, however are not precise enough to be used in combination with akinesia detection.

Smartphones

Smartphones are becoming increasingly common, even in developing countries. As they contain a number of sensors such as accelerometers, microphones and cameras, they are a promising tool in *telemedicine*, where PD can be remotely diagnosed or monitored. The universal nature of smartphones makes large PD datasets possible, with the 8,000 patient mPower [104] dataset used in this paper crowdsourced from smartphone users.

Smartphone studies generally use features presented in speech and accelerometer research, along with additional tests such as memory or tapping tasks [51]. 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[14] [105, 106] is a similar research area, however techniques are less applicable as measures are often more coarse.

Little et al. (2009) [22] provides evidence that a high quality microphone is not required to classify dysphonia, obtaining good results on a dataset of 33. Brunato et al. (2013) [39], Boussios et al. (2013) [38] and Arora et al. (2014) [37] also manage to obtain good results with simple accelerometer based features. However all of these models have been tested on small datasets, which are prone to overfitting on cross validation [19], bias and uninformative predictors [18].

Zhan et al. (2016) [52] conducts a smartphone feasibility study on the largest dataset to date — 121 PD and 105 control. Participants were recruited into the study and asked to asked to conduct tasks such as walking, saying 'aaaah...' and alternated tapping [51]. However, Zhan et al. obtained results barely above the conditional baseline when predicting on features from all tasks (71% accuracy). This result is also 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 the previous works such as Arora et al. (2014) [37] which reported 98.0% accuracy on very similar accelerometer features. It is evident that reported results must be taken with a grain of salt. A possible cause is that Zhan et al. does not control the android smartphone used, hence the sensor data collected varies significantly between devices. Zhan et al. also uses very basic features to quantify speech, neglecting the state of the art speech signal processing features used in other works [33, 95].

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. Although many features have been crafted specifically for diagnosis of PD and Alzheimers with EEG [63, 107], this section will only cover those which may be applicable to speech and movement.

A variety of EEG signal processing techniques are inspired by non-linear dynamical systems theory. It is believed that EEG signals are generated by non-linear coupling interactions between neuronal populations [107]. Patients suffering from neurodegenerative

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

disorders often exhibit decreased complexity in EEG patterns, believed to be caused by the a decrease in non-linear cell dynamics [108]. Features developed with EEG signal processing aim to characterise the dynamic structure of this system. As these features are not inspired by human senses, these features are the very promising for the task of measuring the presence of symptoms undetectable by a neurologist.

***Highlight* 1.8.** The nature of features related to EEG make them very promising for the task of measuring the presence of symptoms undetectable by a neurologist.

The *Lyapunov Exponents* quantify the divergence of two systems with infinitesimally similar initial conditions. The Largest Lyapunov Exponent (L_1) characterises the chaos[15] or rate of divergence of a system and is commonly estimated with Rosenstein's algorithm [109] which reconstructs the system's dynamics using a time delay technique[16]. The L_1 has long been used in the EEG analysis of sleep and as a feature for machine learning [111, 112]. More recently, the L_1 has been applied to analyse the non-linearity of speech [113, 114], gait and balance [115, 116, 117].

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[17] [118]. A higher value implies 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 [119, 120]. The fractal properties exhibited in neuronal control are reflected in heartbeat and gait [121] with force plate data from elderly and Parkinsonism subjects showing a significant increase in fractal dimension compared to healthy young subjects [122, 123, 124]. Esteller et al. [125] compares algorithms estimating the fractal dimension of signals.

The *Hurst* exponent characterises the autocorrelation or long-range dependence of a signal [126]. 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 [127]. 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

[15] Chaos refers to the sensitivity of a dynamic system to its initial conditions.

[16] Additional Lyapunov exponents generally require known equations describing the system [110].

[17] 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.

fractal dimension, the Hurst exponent is a valuable tool in the analysis of gait and balance [128]. Detrended Fluctuation Analysis (DFA) is essentially a generalisation of the Hurst exponent for non-stationary[18] stochastic processes and has been applied in dysphonia diagnosis [86]. Although DFA is the more robust measure, the disparity between the measures may reveal information on the dynamics of the system.

Fisher Information is a measure relating to the uncertainty of measuring a variable (signal) about the unknown parameters modelling its distribution [129]. It is applicable in quantifying non-linear dynamics [130] and is often applied in the analysis of EEG [131]. General entropy will not differentiate two sequences where the frequency of each variable 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 which aim to quantify this unpredictability in a signal [132, 133]. The multi-scale sample entropy [134] is especially powerful tool in the analysis of biological signals [135, 136]. Although these are a prominent feature in EEG analysis, they are rarely used in voice and movement analysis.

1.2.5 Summary of Features

This section summarises all the features covered in previous sections which are used in the construction of models presented in this thesis, which are based on audio and accelerometer signals. The libraries and parameters used to implement are covered in detail in section 2.7. Features are organised by field of introduction or by the field they are most commonly applied in. All relevant general and EEG features are extracted for both the audio and accelerometer models used in this thesis. Additionally, RPDE and DFA are used as a feature for the accelerometer data.

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

General Signal Processing	
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

[18] Non-stationary systems are those with properties which evolve over time.

Fourier	Transforms the signal from time domain to frequency 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 [138]	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

Speech – Dysphonia

Power Cepstrum	From the inverse Fourier domain. Commonly taken in the Mel-log scale [80], resulting in the MFCC [96]. Minimal interpretability, however is the primary feature used in speech recognition [97].
Pitch [77]	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 [89, 139]	Measures the ratio of noise in a voiced signal (signal to noise)
Jitter [88]	Measures of the variation between the length of each glottal cycle.
Shimmer [87]	Measures of the variation of amplitude between each glottal cycle.
LPCC [140]	Coefficients of an *autoregressive* model which measures how well a signal can be modelled linearly by its previous values.
GNE [90]	An extension of HNR by Michaelis et al. [90] to improve reliability in dysphonia quantification
VFER [93]	An further extension of HNR, building upon the theory of GNE.
EMD-ER [141]	Another technique developed based on non-linear speech theory to quantify signal to noise
GQ [93]	Measures standard deviation of duration the glottis is opened vs closed.

DFA [86, 92]	Detrended Fluctuation Analysis. A generalisation of the Hurst exponent which measures the self-similarity of a time series.
RPDE [86]	Measures the repetitiveness of a signal, specifically designed with non-linear speech as the target.
PPE [22]	Measures the variation in successive glottal cycles.
Wavelet Measures [23]	A set of 180 measures for dysphonia based on wavelet transforms to the f_0 of speech introduced by Tsanas et al. (2011) [142].
GeMAPS [143]	A minimal acoustic feature set of 58 or 87 (eGeMAPS) parameters that performs well in general speech classification [33].
Interspeech ComParE [144]	An exhaustive 6,368 feature set for general speech classification [33]. 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 Parkinsonism tremor.
Jerk [145]	The change in acceleration. The jerk signal may be more effective when combined with certain signal processing methods.
Sway Area	Simple measures such as bounding ellipse can quantify the amount of sway. A 95% CI is often taken to remove outliers.
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 [103].

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

EEG

Hjorth Parameters [146]	Three simple statistical measurements of a signal which have been used as features in EEG and IMU models [147].
Lyapunov Exponents [109]	Characterises the divergence of systems with close initial conditions. The largest exponent (LLE) [115] is most commonly used.
Fractal Dimension [118]	A measure of how the detail in a signal changes with the scale at which it is measured. The Higuchi [148] and Petrosian [149] fractal dimensions are used in this thesis.
Hurst Exponent [126]	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 [129]	Quantifies the non-linear dynamics in the system generating a signal.
Ap/Samp Entropy [133]	Approximate and sample entropy quantify the unpredictability of a signal. Multiscale entropy increases information content [134].
SVD Entropy [150]	A measure of complexity. The entropy of the orthogonal vectors from singular value decomposition needed to describe the signal.

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 patient 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 or 1. 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.

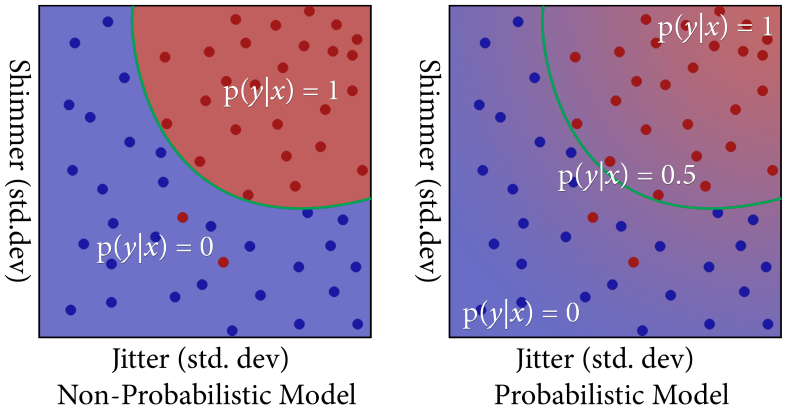


Fig. 1.2: A visualisation of binary classification with two features. Data is rarely as 'clean' as this artificial example.

Traditional machine learning models were built on statistical foundations. The mathematical backing these models are solid and the models well understood. However, the mathematics of these models were developed on assumptions that are rarely satisfied with real world data. Models such as deep neural networks have started to rise to popularity recently due to their modelling power. However the behaviour of deep neural networks are poorly understood and difficult to analyse.

Most models have strengths in different areas, and very rarely does a model strictly dominate another. The choice of model is often informed by the data. For example, models like deep neural networks may perform well when data is plentiful, however in small datasets the very simple decision tree may greatly outperform neural networks[19].

***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[20] or unrelated to accurately predict the data. An optimal model cannot achieve beyond this irreducible error. Bias describes a model 'fitting' the data poorly and is evident in a model with low accuracy. Variance describes how 'unstable' a model is — a model with high variance may score 100% accuracy but generalize poorly to new data. A model with high variance is essentially predicting results by 'memorisation.' Fitting a model with high variance is often

[19] These will be explained in section 1.3.1 and 1.3.2

[20] *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'.

known as *overfitting*. The bias-variance tradeoff [151] is a fundamental problem in machine learning, describing the difficulty in reducing bias without increasing variance and vice versa.

Models often have one or more adjustable parameters to balance bias and variance. These parameters are tuned with intuition combined with some form of search [152, 153].

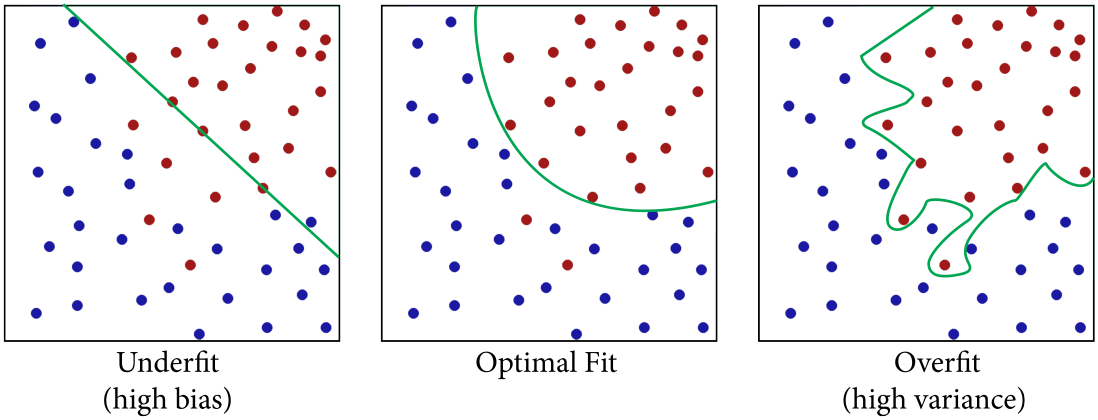


Fig. 1.3: Machine learning models and their parameters must be carefully chosen to ensure the optimal fit.

Overfitting is a major issue in machine learning as data is limited and models are often too complex to analyse. Visualising and detecting overfit may be simple when fitting a function in two dimensions, however it is significantly more difficult when the input has thousands of dimensions. A model that has overfit will appear to predict the data well, however fails to generalize to new data. Cross Validation is the gold standard in machine learning when it comes to model evaluation and recognising overfitting however it is not uncommon to find textbook examples which apply it incorrectly. Cross validation and other techniques used for model evaluation will be discussed in detail at section 1.3.4. Like any statistics based field, careful analysis of the results is required and unfortunately this of often neglected in machine learning literature.

The following sections will cover three common models, random forest classifiers, support vector machine and neural networks. The mathematical formulation of these models is abstracted in favour for intuition behind their behaviour. We refer to Bishop et al. (2005) [154] for a more formal description of these models.

1.3.1 Traditional

Traditional models are the approach favoured in current literature [30] due to the limited data and their interpretability. 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. However both are also non-probabilistic classifiers[21]. There exists models which are inherently probabilistic such as Gaussian Processes however they are less commonly used as they generally offer lower performance than decision boundary based classifiers.

Random Forest [156] classifiers are derived on the concept of Bootstrap Aggregation (*bagging*) [157] where the results of multiple models are aggregated 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.

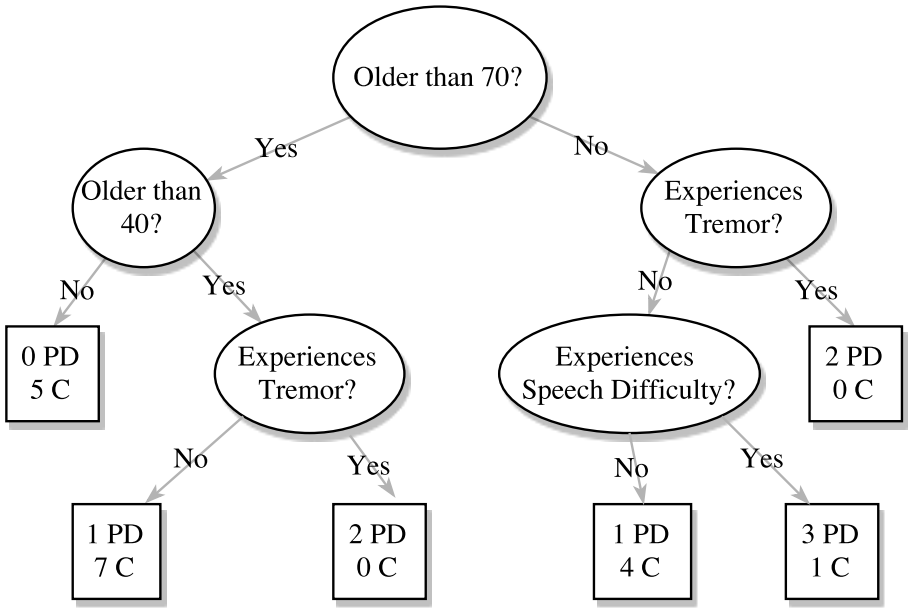


Fig. 1.4: A simple Decision Tree with cutoff depth 3. Data is split by rules until a leaf contains only one class exists or a cutoff criterion is satisfied.

Decision Trees are simple to interpret and are robust against high dimensional data. However, determining the optimal decision rules at each node as well as the optimal 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.

[21] In general. Methods of generating pseudo-probability with SVMs have been proposed [155]

Random Forests correct for the tendency of decision trees to overfit and provide 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 [158]. The square root of the number of features for classification is recommended by Breiman [156] and is commonly used in most applications. Hence it is rare to perform hyperparameter tuning on random forests.

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

Support Vector Machines [159] are built on the concept of creating the optimal decision boundary. The motivation is to create decision boundary which maximises the margin[22] between different classes. This can be computed by solving a Lagrangian, however, this is only mathematically possible with a linear decision boundary. As most problems are not linear, the *kernel trick* is used to transform the data into a linear space.

A kernel is a measure of similarity between two data points, and the kernel trick transforms the raw input into the feature space of the kernel[23]. Non-linear kernels enable a SVM to fit a non-linear function however the exact non-linearity in the data is rarely known. There are uncountably many kernels, and kernels such as the Radian Basis Function (RBF), Fisher and Polynomial are commonly used[24]. Kernels generally have adjustable parameters, such as the degree and constant coefficient for polynomial kernels.

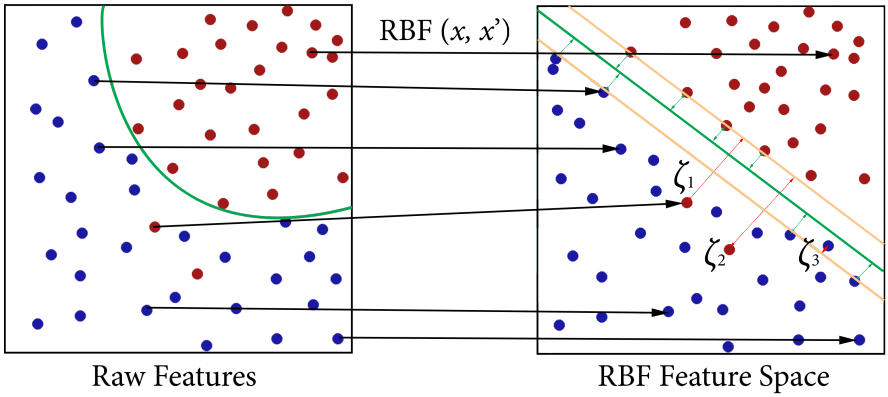


Fig. 1.5: A RBF kernel is used to transform the data into a more linearly separable space. ζ_i denote slack variables which lie beyond the margin (depicted by beige lines).

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

The original SVM algorithm was not able to handle cases where data was not separable. Cortes and Vapnik (1995) [160] introduced slack variables ζ_i, which define a penalty for data beyond the SVMs margins thus extending the use of SVMs to non-separable data. 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 minimize errors.

The combination of kernels and slack variables greatly improved the applicability of SVMs. SVMs became very popular in the machine learning community as they were simultaneously analysable and powerful. However, kernels and slack variables also introduce a number of hyperparameters, such as the scaling factor C and the type of kernel and its parameters. Although intuition and knowledge of the data can guide kernel and hyperparameter choice, techniques such as grid or random search [152] are generally used to fine-tune them. However, hyperparameter tuning increases the risk of overfitting, which will be discussed in detail in section 1.3.4.

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 pre-defined basis functions or the kernel trick to reduce non-linearity. Random forests and decision trees are powerful when data is readily available, however they do not model functions of data and are less suitable when predicting unseen outliers [161]. Neural networks are popular models used when the dataset is reasonably sized and there exists a difficult to describe structure in the input features. They are extremely powerful, however very difficult to interpret or debug and 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[25] [162]. In 1958, the simple perceptron learning algorithm was developed [163], which would become the building blocks of neural networks today. The fundamental concept of a neural network is connecting many perceptrons (acting as neurons) together to simulate the behaviour of a biological brain.

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

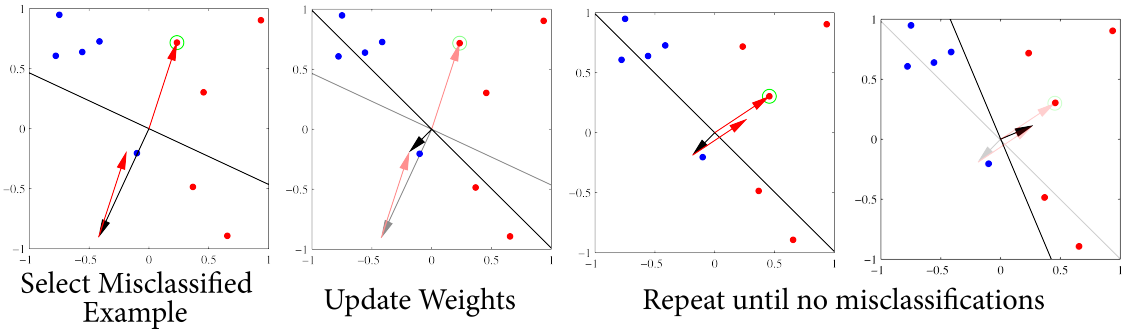


Fig. 1.6: The simple perceptron learning algorithm. The original incarnation could not handle inseparable data [163]. Images borrowed and modified from Bishop (2006) [154]

A perceptron by itself 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[26] *activation function*. At the time, computational power was limited and large networks impossible to train. Early works by Minsky and Papert (1969) [164] 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 layers of perceptrons could model the XOR function. In 1989, it was shown that a single layer with enough perceptons is able to approximate any non-linear continuous function [165].

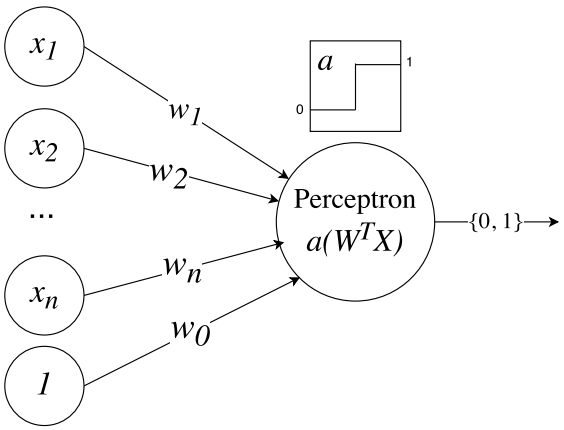


Fig. 1.7: A single perceptron node. Takes input X and learns the weight vector W to classify the output with the Heaviside activation function a.

Although multiple layers of perceptrons had always been the goal of neural network research, training them was not possible until backpropagation, a form of gradient descent[27]

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

[27] Surprisingly, Werbos' work on backpropagation [166] was lost and would be rediscovered a decade later in 1985 by Rumelhart et al. [167]

was introduced [166]. Backpropagation required the activation function to be differentiable, hence the sigmoid[28] replaced the Heaviside activation function. Neural networks were now able to reliably 'learn' complex non-linear functions, 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 a significant advantage over traditional models where this non-linearity must be pre-defined.

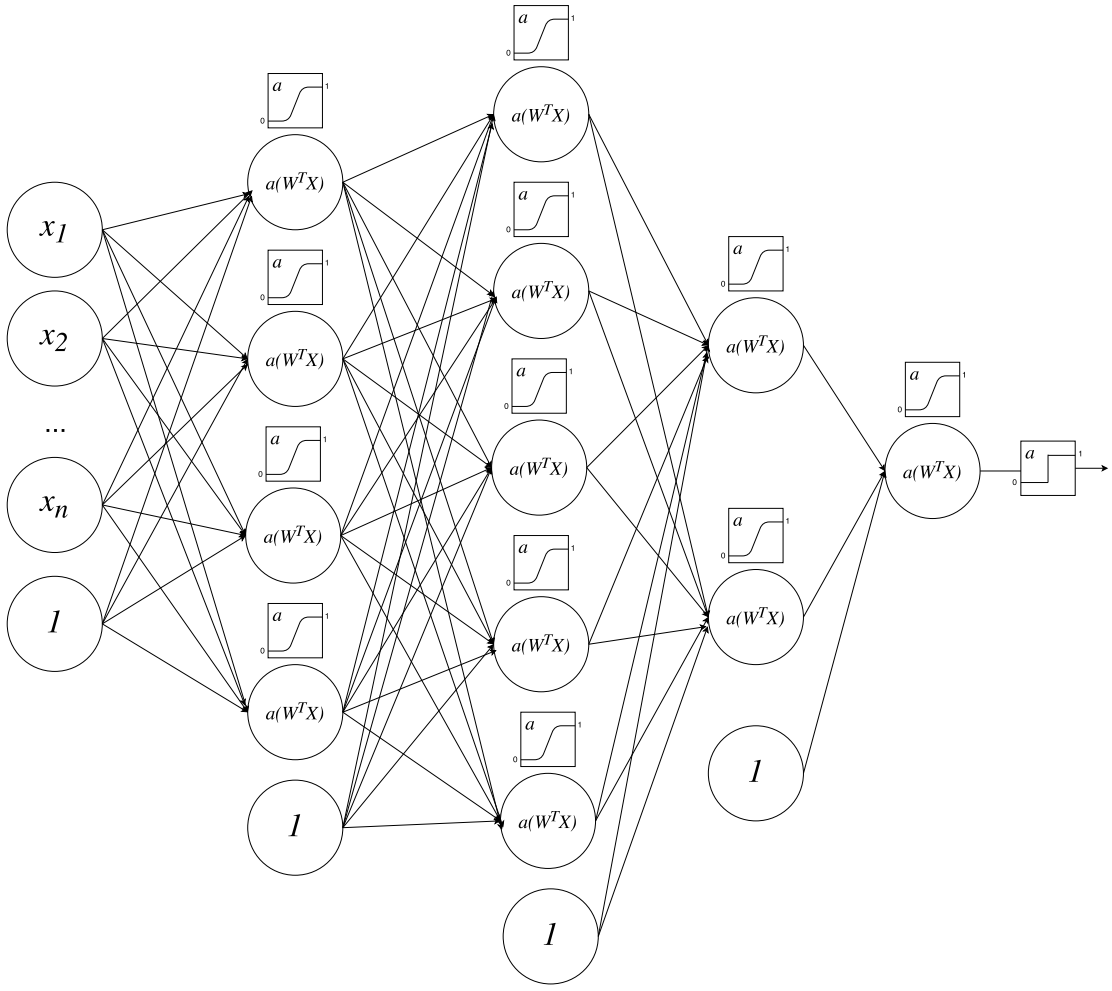


Fig. 1.8: A simple 3 hidden layer fully connected neural network with sigmoidal activations. By stacking non-linear activation functions, neural networks are able to learn any non-linear function of the input. The '1' nodes represent the bias at each layer.

Neural networks are computationally expensive models and unlike traditional models,

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

training requires optimising a non-convex function. This is computationally intractable and current neural networks are trained by finding a good local optima through gradient descent with backpropagation [168]. The vanishing gradient problem [169] limited the depth of neural networks until the recent development of batch normalisation [170]. Previously, careful management of gradient flow was required to train deep neural networks [171].

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 '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' and can negate the need for feature extraction for certain types of input data. This is especially evident in the task of image recognition, where CNNs have rapidly exceeded the performance of traditional models.

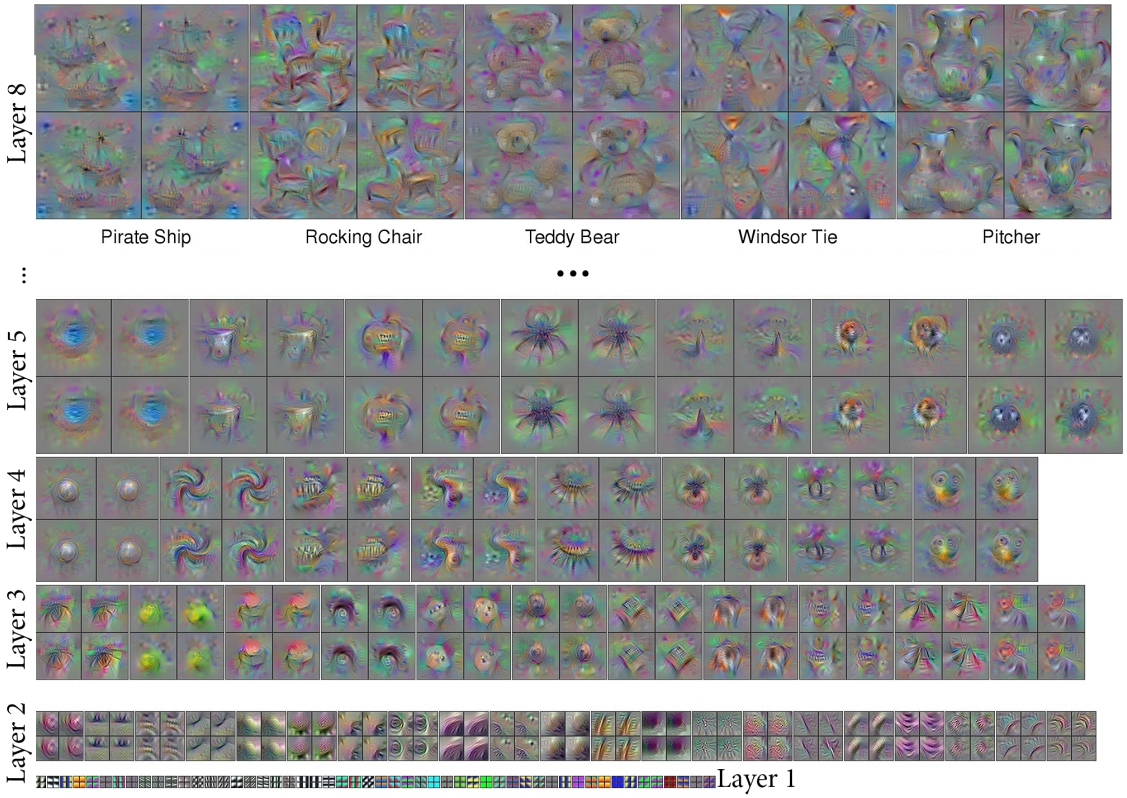


Fig. 1.9: A visualisation of CNN nodes from Yosinski et al. [172]. Layers capture increasingly complex relationships between pixels and act as features input into further layers.

Recurrent neural networks (*RNNs*) are neural networks with cyclic connections. This essentially creates an internal state which allows neural networks to better handle tempo-

ral data and provides flexibility for the type of input and output data. There are a number of RNN variants, however long short-term memory (LSTM) nodes are applied in practice as they are more robust to the vanishing gradient problem [173]. A recent model combining LSTM and convolution layers has been very effective at classifying visualised EEG data [174].

Neural Network Hyperparameters

Neural networks inherently have a large number of hyperparameters. 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 it's bias and is more suitable when the structure of data is complex or data is plentiful. It is thought that networks with many nodes per layer (*width*) are better at memorization whereas additional layers (*depth*) are better at generalisation of features [175]. Depth can also be exponentially more valuable than width for modelling the structure of complex non-linear data [176]. There is still no consensus on the balance between number of nodes and layers — these must be fine tuned for particular problems with intuition and search.

A large neural network has a tendency 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 is preferred. However in the context of 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 [177]. At first, this may appear unintuitive, however the idea is to promote redundant feature representations to improve its robustness. Dropout generally outperforms weight regularisation, and a combination of dropout and early stopping is commonly applied. There are also variations of dropout such as dropconnect [178] 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 [179, 180]. 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 [181] ($f(x) = \max(\alpha x, x)$ for $\alpha \leq 1$), Maxout [182], noisy ReLU [183] and exponential linear unit [184].

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. Glorot and Bengio [185] proposed 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. This is commonly termed Xavier initialisation and is effective for networks with sigmoidal or tanh activations however ReLUs rapidly tend to zero. He et al. [181] proposed a small modification to fix the dead ReLU issue by setting variance to $2/n_{\text{in}}$.

The method of gradient descent, referred to as the *optimizer* 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. Non-linear techniques developed in convex optimisation such as conjugate gradient descent and (Quasi-)Newton are powerful yet rarely applied in practice due to their computational complexity. The most popular optimizers for neural networks incorporate the concept of momentum, where previous gradients are considered in the descent. Adam [186] is one of the most recent optimizers and combines elements from two powerful optimizers before it, AdaGrad and RMSProp. Nesterov momentum [187] — which has favourable properties in convex optimisation — can also be incorporated into Adam, creating Nadam. [188].

Each optimizer 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 (2016) have proposed a novel approach where the learning rate is fluctuated during training to create an ensemble in one training process [190].

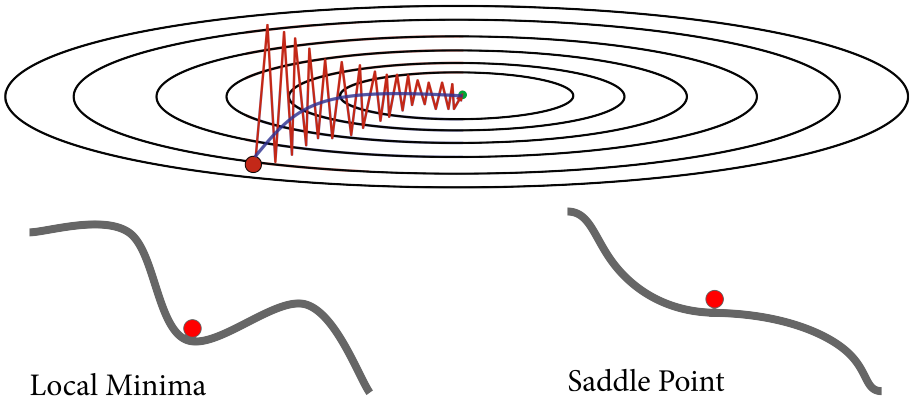


Fig. 1.10: Traditional gradient descent (red curve) performs poorly in 'long valleys'. Optimizers generally use momentum to simulate the behaviour of the optimal blue curve and avoid local minima. Diagram adapted from Stanford's CS231n [189].

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 determine which are most relevant[29]. Redundant or highly correlated features reduces the performance of most machine learning algorithms. Simple models like Naive 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 require more data or 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 [19] and speeds up training. Selecting an 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 exhaustive search would be required to determine the optimal subset however is computationally infeasible. 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 is generally fast and results are

[29] There are issues with this approach such as Freedman's paradox [18] however resolving this is the task of model evaluation (section 1.3.4).

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. This allows the selected features to be better tailored to each model. However wrappers can be computationally prohibitive when wrapping around models such as neural networks and also cater towards the model's tendency to overfit [191]. *Embedded* methods rely on machine learning models which inherently perform feature selection during their training, often from strong regularisation.

The performance of these approaches are highly problem-dependent. This thesis will employ a number of state of the art supervised feature selection algorithms as depicted in table 1.7. We refer to Li et al. [192] for a thorough description and comparison of these techniques.

Table 1.7: Feature selection methods used in this thesis. Neural network forward/backwards search was not performed as resources were limited.

Filter	Wrapper	Embedded
ReliefF [193]	SVM/Gaussian Process	RFS [200]
Fisher score [194]	Forward/Backwards	ls_l21 [201]
CIFE [195]	Search	
JMI [195]		
ICAP [196]		
MIFS [197]		
MRMR [198]		
CFS [199]		

Rather than eliminating features, *dimensionality reduction* aims to reduce the amount of information required to represent the set of features. This reduces the correlation between features and can significantly improve performance with simpler models. The two most common forms of dimensionality reduction are the unsupervised principal component analysis (PCA) and supervised linear discriminant analysis (LDA) and variations [154]. Neural networks can also be used to reduce the dimensionality of data by training a network to predict the input where the center hidden layer contains fewer nodes than the input. These are termed autoencoder networks and can out-perform PCA however are difficult to analyse [202].

***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 generalize well to new data. Accuracy over the entire dataset is evidently not a good metric, as an model which memorises the data (overfit) can appear to have perfect accuracy while failing to generalize to new data. Model selection and evaluation is the field in statistics which handles this. However the field is contentious — especially as model selection performance varies based on the type of data.

Cross validation (*CV*) has become the de-facto standard in machine learning. Conceptually, CV is very simple. The primary types used in machine learning are *leave one out* and *k-fold*. Lets assume there are 100 data points in a dataset. In leave one out CV (LOO), 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 fold 1 and testing on fold 2 then training on fold 2 and testing on fold 1. Common values of k are 2, 5 and 10.

In summary, we will be performing 10 fold cross-validation with random stratification[30] 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 [203] is used to test if the mean performance of one model is greater than the other. This decision will be justified in the next section with more background into model selection and hypothesis testing provided.

Model Selection and Hypothesis Testing

K-fold 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. Leave one out CV allows almost all the data to be used in training. When the data is clean (high signal to noise ratio) LOO performs nearly unbiased estimations [204]. However LOO has been criticized for preferring models with a high variance and is less computationally feasible for large data sets [205]. Kohavi (1995) [205] instead recommends 10 fold CV in the general case. CV variations such as exhaustive and

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

Monte-Carlo CV exist however they are not recommended by statistical literature [206, 204].

There are a number of catches when performing CV:

- 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 naively over the whole dataset can easily overfit.
- When performing hyperparameter optimisation the CV score is often used as a metric. The risk that the best model fits the validation sets well by pure chance increases as more parameters are explored [19].

Overfitting 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 (1997) [19] 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 models overfitting CV by chance. Bouckaert (2003) [207] 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[31] and may not be a good representation of a model's effectiveness in difficult tasks. Sensitivity, or true positive rate is a measure of the percentage of positive classes correctly identified and specificity or true negative rate measures the percentage of correctly identified negative examples. The F_1 *score* is the harmonic mean of sensitivity and specificity and is an 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 [208]) curve (*AUC*) 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

[31] 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% 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.

be a more robust and statistically consistent measure of model performance [209]. However recent empirical experiments have shown that AUC favours particular models [210] and it has been criticised for being incoherent [211, 212]. Modifications to AUC have been proposed [210, 212] however they are uncommon in practice. As a result, accuracy and F_1 score will be the primary performance measures utilised in this thesis as they are interpretable and model characteristic independent.

Hypothesis tests are used to determine if the results obtained in experiments are *statistically significant*. After obtaining the accuracies or F_1 scores for each fold of cross validation, a hypothesis test should be used to determine that the difference in results is not from chance. A paired t-test is the traditional approach to testing if one population mean is greater than another. There have been criticisms of frequentist hypothesis testing promoting the publication bias [213], citing the non-replicability crisis in psychology [214]. Recently, the American Statistical Association has officially endorsed Bayes factor [203] as their preferred method of hypothesis testing [215], and mass-replication studies have shown that almost half of previous psychological research do not meet the criteria for strong evidence when Bayes factor is applied [216]. The standard two tailed Cauchy distribution is used as a prior in this thesis [203].

In statistics, there is no agreed upon method for model selection and evaluation. Penalization based evaluation[32] criteria such as Akaline/Bayesian/General Information Criterion [217, 218] and Minimum Description Length [219] 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.

[32] *Penalization* based model criteria are inspired by Occam's razor, preferring simple model over a more complex one which obtains similar results as it is less likely to overfit

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. There have been two large scale literature reviews, Alhrics et al. (2013) [30] and Bind et al. (2015) [31]. In these reviews it is apparent that multiple sub-fields exist and research is often confined in its own sub-field. For example, the top papers in the Interspeech 2015 PD speech challenge [53] used methods independent of the dysphonia feature extraction 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 [220]. It is common to find a paper failing to cite prior work which performs the same experiments. A goal of this thesis is to consolidate and distil prior work into a easily digestible 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 good results, it is difficult to determine if these results are caused by biases in the dataset or overfitting. With any field based on empirical statistics, a publication bias exists [213] and there will exist results which are not replicable [214]. Section 1.3.4 details measures to avoid overfitting and evaluate models however their implementation is uncommon in applied machine learning literature. The variation of results on experiments with very similar setups shines doubt on the replicability of results for some of the best performing papers. Arora et al. (2014) [37] achieves 98.0% accuracy using smartphone IMU data from 20 participants. Zhan et al. (2016) [52] performs an experiment using all features in Arora et al. (2014) as well as additional speech and tapping measures however only manages 71% accuracy. Furthermore, the state of the in motion mode recognition rarely achieves such results despite the motion mode recognition likely being the 'easier' task [105].

The following sections detail the experiments performed as part of the project. We will apply a combination techniques used in state of the art on a larger dataset to assess true performance. The mPower dataset [104] has been chosen for this task and will be described in the following section.

- Section 2.1 discusses the dataset used (mPower) and how the data was filtered and pre-processed.
- Section ....

2.1 The mPower Dataset

To minimize the likelihood of bias or overfitting, a larger dataset was required. Currently, the only publicly available dataset that satisfies the size requirements is mPower [104].

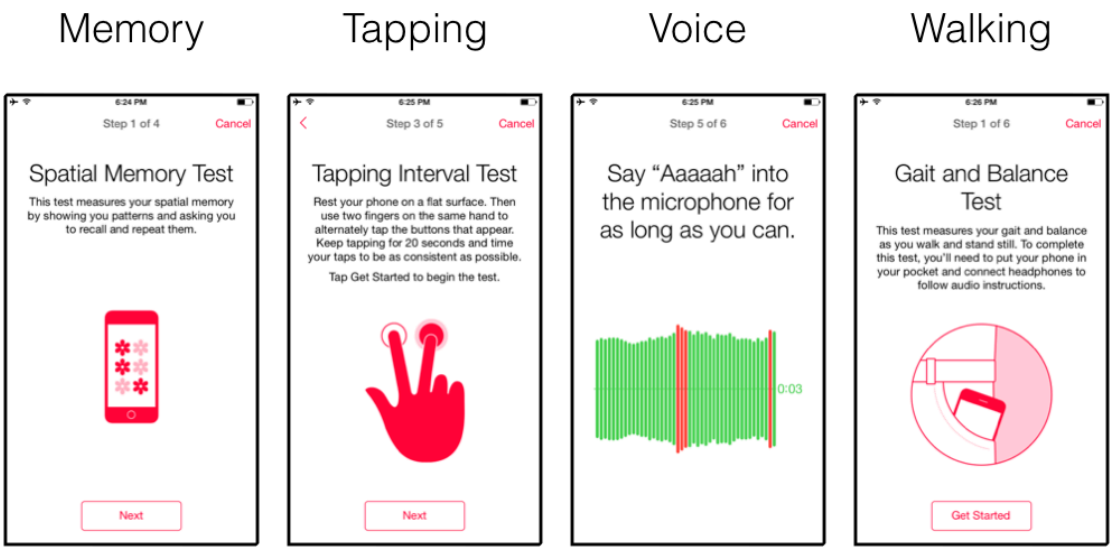


Fig. 2.1: The mPower app consists of several tasks to evaluate memory, bradykinesia, voice and gait.

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 was presented with the tasks presented in figure 2.1 along with general demographics questions 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 a variety of backgrounds and may have other illnesses (however this was not recorded as part of the dataset).

The mPower dataset also contains a number of cases of young-onset Parkinson's disease[1] [221, 222] which has rarely been studied in a diagnosis context. Age is a bias in the dataset as a majority of the non-PD participants in the study were young adults. Using age alone, the prediction PD $\Leftrightarrow$ age $>$ 52 would result in 86.1% accuracy.

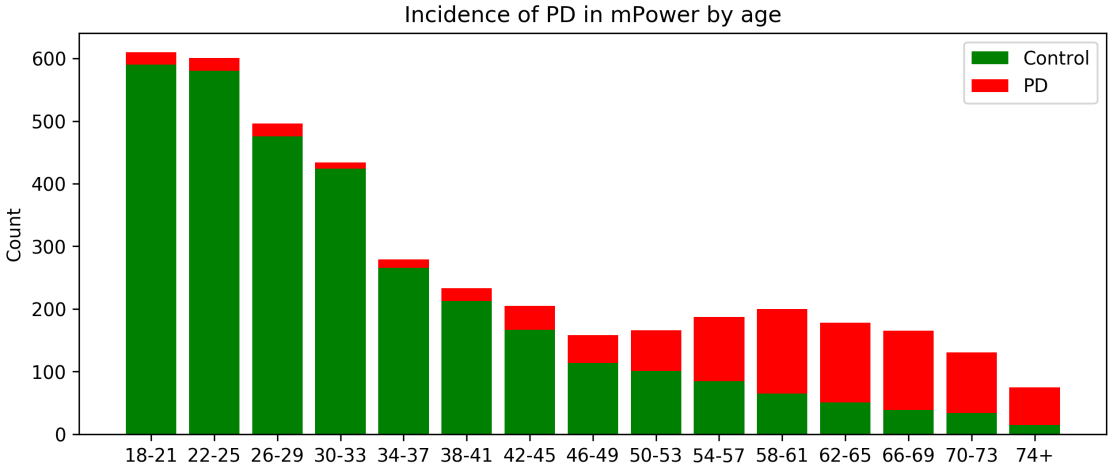


Fig. 2.2: Age is a bias in the mPower dataset as most non-PD participants are young. There are also some cases of rare young-onset PD[1].

Despite the dataset being released to the public in early 2016 and having multiple citations from machine learning and clinical papers, there has been no machine learning study published using the mPower data. The primary issue with the data is that it is quite 'noisy' — a major issue with any crowdsourcing project without significant precautions [223].

2.1.1 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. Although the mPower application prevented access to the voice task when background noise exceeded a certain threshold, this threshold was too lenient. A large number of participants also failed to complete the recording task properly — hesitation, interruptions and pronouncing vowels other than 'aaaaah...' were common. There was also a large variation in the distance to the phone during recording with some participants speaking directly into the microphone creating a large amount of 'wind noise' [224].

At the time of writing, there were 65,000 speech samples from 6,000 subjects in the

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

mPower dataset (a majority of these from a small number of users). We evaluated approximately 2,000 randomly selected samples for performing the task correctly and having acceptable levels of background noise, rejecting around 25%. 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 speech samples. After filtering, 4,100 users remained, 900 with PD. The highest ranked speech sample was selected for each of the users[2]. Machine learning could optimize this process, however it 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 for 30 seconds. During this task, accelerometer and gyroscope data is continually collected at 95±7 Hz. Although in-pocket IMU gait estimation exists [103], mPower does not record the parameters necessary (such as leg length) to estimate parameters other than cadence. The results of Esser et al. (2011) [20] suggests that although PD patients on average have a longer cadence, the separation is not clean.

The standing task is therefore more interesting in the context of machine learning. As the device is in the user's pocket or bag, data from the gyroscope would be minimally informative. Using gyroscope data, a rotation matrix was calculated to align the accelerometer's z axis to the direction of gravity.

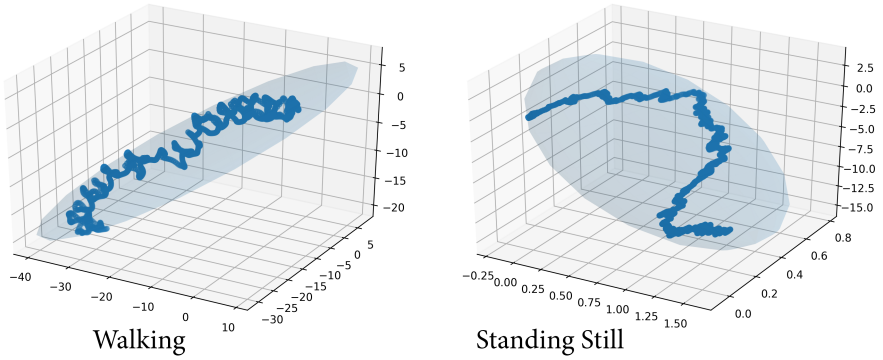


Fig. 2.3: A visualisation of device position after correcting for rotation with a bounding ellipsoid at 95% CI. Gravity (z) is not subtracted.

Unlike similar experiments carried out in force plates, the subject was not instructed to stand as still as possible. A majority of subjects show a significant amount of sway which could be consciously preventable. To map the accelerometer data more closely to force plate data, a 10^{th} order zero-phase 1hz Butterworth highpass filter was applied. The high-

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

pass filter removes preventable sway at the cost of removing valuable sway information below 1hz [225].

A 16 second extract of rest data between 4s and 20s and the first 10 seconds of the walking task were used for each subject for feature extraction. The choice of these values were solely informed by the nature of the dataset, with the first four seconds of rest data containing significant movement and most recordings of variable length. Features specified in section 1.2.5 were extracted using the tools and techniques specified in section **??**. Feature Extraction was done on both the original and filtered data for the resting task and only unfiltered for the walking task. The motion data was then filtered and ranked based on simple criterion such as average acceleration and the best selected for each subject.

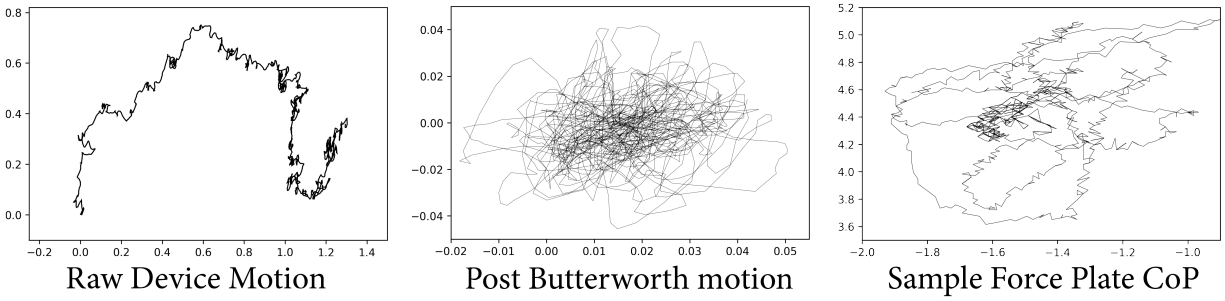


Fig. 2.4: The Butterworth filter results in a device path more similar to the centre of pressure, however low frequency sway information is lost. Note that the device motion recording is 30 seconds long while the force plate is 10 seconds.

2.2 Replicating Past Work

The two key results we will be replicating on the mPower dataset are the 98.6% accuracy from vowel phonation reported in Tsanas et al. (2011) [93] and the 98.0% accuracy with smartphone accelerometer data reported by Arora et al. (2014) [37].

2.2.1 Vowel Phonation

Tsanas et al. (2012) [142] uses the National Center for Voice and Speech (NCVS) dataset which consists of 33 people with PD and 10 healthy controls. 263 phonations in total were recorded in controlled circumstances using a professional grade microphone. HNR, GQ, RPDE, DFA, PPE, GNE, VFER, EMD-ER, MFCC and variants of shimmer and jitter were extracted, resulting in a set of 132 features (See 1.2.5).

Features were calculated on the 263 phonations and 10 fold, 100 repetition cross val-

idation used for evaluation of models. It is unclear whether Tsanas et al. has split the phonations on a per-subject scale. Failure to do so presents a high risk of overfitting as two phonations from the same subject may appear in both the training and validation set. Random Forests and SVMs were evaluated with hyperparameters selected by grid-search [152]. 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* 2.2.** It is unclear whether Tsanas et al. has split the phonations on a per-subject scale and failure to do so presents high risk of overfitting.

Fig. 2.5: Cross-validation accuracy of Tsanas et al. 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

We replicated Tsanas et al. on the 4,100 phonation samples selected after preprocessing mPower (see 2.1.1). Features were extracted from a 2 second window was of each audio sample which mirrors the phonation length used in fundamental frequency estimation datasets [226]. Gridsearch was performed to find (near) optimal SVM hyperparameters. The best performing feature subset of Tsanas et al., extracted with the ReliefF algorithm is initially evaluated.

Note that the NCVS data used in Tsanas et al. is at a ratio of 33PD:10C whereas the mPower data is at a ratio of approximately 9PD:32C. We stratify the data by random sampling to simulate NCVS split. On both the NCVS and mPower ratio, the SVM classifier exhibits the false positive paradox, where the most common class is predicted for all inputs. The results are summarised in table 2.2.

Table 2.1: Cross validation results of optimal SVM from random search using Tsanas' 10 feature ReliefF subset. Presented as mean ± stdev.

Equal Split (50P:50C)

	Pred PD	Pred C
True PD	30.1 ± 2.5%	20.0 ± 2.5%
True C	15.1 ± 2.5%	34.9 ± 2.5%
Accuracy		65.0 ± 3.3%
Sensitivity (TP)		60.1 ± 5.0%
Specificity (TN)		69.8 ± 5.0%

NCVS Split (33P:10C)

	Pred PD	Pred C
True PD	76.7 ± 0%	0 ± 0%
True C	23.3 ± 0%	0 ± 0%
Accuracy		76.7 ± 0%
Sensitivity (TP)		100 ± 0%
Specificity (TN)		0 ± 0%

The results using the mPower dataset are evidently poorer than the reported 98.6% accuracy. The ReliefF [193] feature subset consists primarily of MFCC coefficients. MFCC is a very powerful feature and is often the primary feature in speech recognition systems. The high and low MFCC coefficients are known to be rarely informative in speech recognition [227] and the ReliefF feature set contains both the 1^{st} and 11^{th} coefficients. The result suggest that these coefficients may be informative when used to detect abnormal speech and more MFCC coefficients should be extracted [96]. MFCC are known for being very sensitive to noise and frequency [98, 228]. Tsanas et al. used professional grade microphones whereas mPower audio data is recorded with a smartphone microphone;

Another possibility is overfitting. It is ambiguous if Tsanas et al. divided phonations of a per-subject level in cross validation. Naive 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 [228], phonations from different individuals are likely 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% vs 97.7%) as the hyperparameters of the RF classifier were not tuned by cross validation and RF is generally more robust against overfitting.

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: Mean Cross validation results of optimal SVM from random search using all features presented in Tsanas et al. (2012) [93]. Outperforms 2.1 with a Bayes factor of 10^{17}.

Equal Split (50P:50C)

	Pred PD	Pred C
True PD	32.4 ± 2.8%	17.6 ± 2.8%
True C	13.9 ± 2.4%	36.1 ± 2.4%
Accuracy		68.4±3.9%
Sensitivity (TP)		64.7±5.6%
Specificity (TN)		72.1±4.8%

mPower Split (9P:32C)

	Pred PD	Pred C
True PD	3.4 ± 0.8%	17.6 ± 0.8%
True C	1.7 ± 0.7%	77.3 ± 0.7%
Accuracy		80.7 ± 1.0%
Sensitivity (TP)		16.1 ± 3.7%
Specificity (TN)		97.8 ± 0.9%

2.2.2 Movement

Arora et al. [37] conducted a study with 10 control and 10 PD participants, obtaining 98.0% accuracy on 10 fold cross validation with 100 repeats on accelerometer data alone. Participants were provided with 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 however it can be assumed to be in the participant's pocket. No preprocessing was done to the data, and features extracted included simple statistical and entropy measures, DFA, mean TKEO and the dominant frequency. It should be noted that Zhan et al [52] extended the dataset and performed a similar experiment on 121 PD and 105 control using additional voice and tapping features, however only achieved 71.0% accuracy.

This thesis coincides with the Parkinson's Disease Digital Biomarker DREAM Challenge which involves using accelerometer data to classify PD or predict the UPDRS motor score [229]. Sage Bionetworks, sponsor of the mPower dataset and a organiser for the challenge released a baseline feature set [230] which included all features in Arora et al [37] and additional jerk based measures and the peak of the Lomb-Scargle periodogram [231]. The work completed as part of this thesis was submitted to the challenge, earning Xth place. The challenge evaluated the trained models on an unreleased portion of the mPower data, validating that our models have not overfit.

After filtering out samples which are too short (<10 seconds), have too much variability in recording rate or are missing measurements, XXXX people remain with a valid walk recording. The most recent valid walk was selected for each patient and all relevant features specified in section 1.2.5 extracted.

Before

2.3 Dynamical Systems Features and Data Boosting

We decided to investigate the potential of traditional

2.4 Visualising The Features

The models replicated in the previous sections clearly perform 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 also likely that introducing a greater diversity of subjects increases the problem difficulty as there are natural variances in speech and gait. No individual feature was able to achieve greater than 60% classification accuracy on an equal split of the data.

2.4.1 Speech

A large number of dysphonia related measures rely on precise measurements of the length of each glottal cycle. The SWIPE [232] fundamental frequency estimation algorithm was primarily 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 gathered in real-world conditions. Issues with f_0 extraction would invalidate the measurements from f_0 based signal processing. A simple investigation shows that the standard deviation of f_0 exceeds 10Hz for 347 subjects, a value that is almost certainly indicates a failure of the algorithm or a poorly executed recording.

Little et al. [86, 22] introduces three measures to distinguish dysphonia — DFA, RPDE and PPE. DFA and RPDE are measured of the autocorrelation of a signal. As evident in figure 2.7, people with PD exhibit 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, the natural variance in speech production makes distinguishing dysphonic speech difficult.

Benba et al. [233] distinguishes PD with a 82% success rate primarily with MFCC. The MFCC are also present in most of the feature subsets derived by Tsanas et al. [93]. MFCC

Fig. 2.6: Standard Deviation of the fundamental frequency during vowel phonation. Females and older individuals exhibit notably higher variations in f_0. People with PD seem to exhibit an increased variation in f_0 compared to age matched control subjects, however the distinction is less clear.

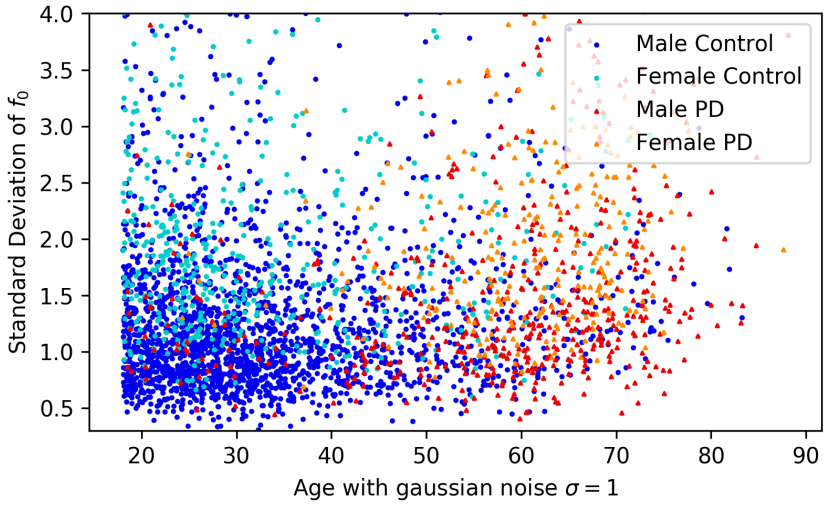
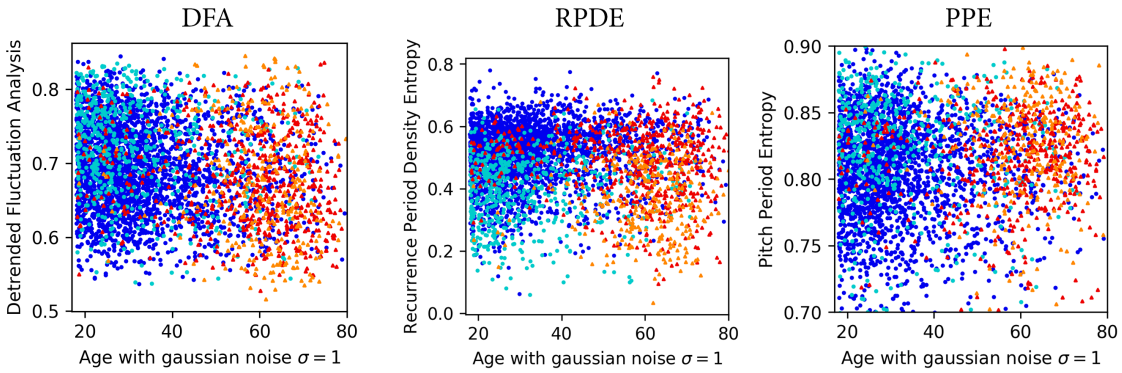


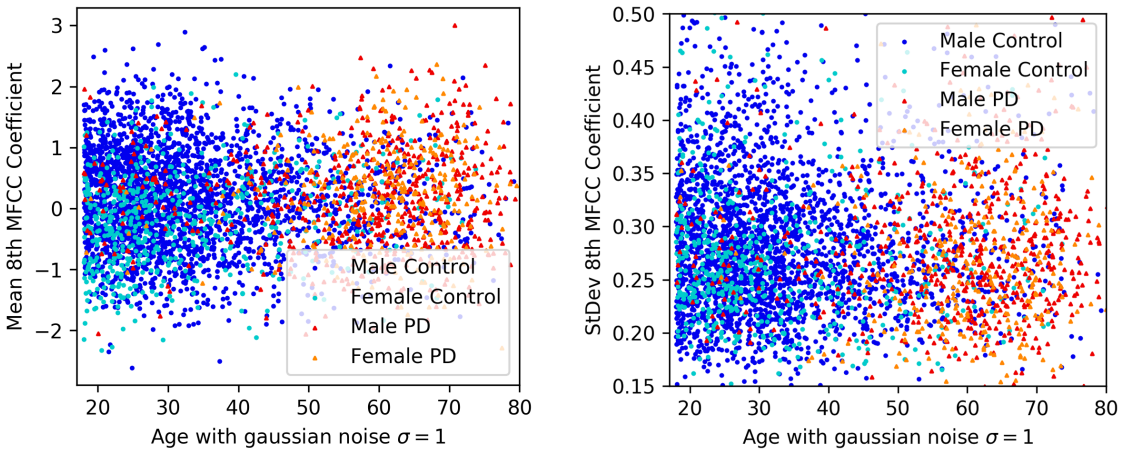
Fig. 2.7: PD subjects exhibit a lower DFA and RPDE, and a higher PPE. However these features are hardly separable.



are the primary feature in most speech recognition systems so it is not surprising they are strong features in detecting dysphonia. However, figure 2.8 suggests that there is a very minimal correlation between mean and standard deviation measures of the MFCC and the occurrence of 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. This will be explored in section 2.5.

YYYY was one of the stronger features in distinguishing

Fig. 2.8: The 8th MFCC coefficient is not a particularly notable feature for distinguishing dysphonia. The visualisation of other coefficients appear roughly the same.



2.4.2 Movement

2.4.3 Conclusions and Recommendations

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

Many features are very sensitive to minor fluctuations in the signal and their value can change dramatically depending on the segment used, as shown in section 2.3. The large fluctuations in value reduces their effectiveness in simple predictors, however these may be valuable when the feature is extracted over short-time intervals. Models such as CNNs or LSTMs (ref 1.3.2) can utilise the additional temporal information and 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.

As no individual feature can achieve good classification accuracy, the machine learning model must combine multiple features to reach a reasonable level of performance. Traditional machine learning models are best suited to modelling features which are independent and often perform worse when two features are correlated despite there being additional information. Feature selection [192] can be used in traditional machine learning models to reduce the impact of this however information is definitely lost. It is possible

that non-linear models such as neural network may be able to utilise more of this information, however a lot of training data will be required.

2.5 Automatic Feature Extraction: Neural Networks

A notable weakness of traditional machine learning to feature extraction is that information is often lost in the process as it is difficult for features to perfectly describe a signal. As evident in computer vision and EEG signal processing [174], CNN and LSTM based neural network architectures present a viable solution, automatically 'learning' the best features from a more raw representation of the data.

The biggest tradeoff in using automatic feature engineering is that understanding the features extracted is a bigger investment than developing the model. A great deal of 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 an similarly difficult to interpret model must be used to fit these uncertain features.

A....

WALK:: USE SUPPL trained model TALK:: USE CROSS VAL.

features unclear.

add noise and understand what happens – what do the numbers represent, if anything?

Computational power may have been too limited Due to the sensitivity of some features, it would be very interesting to extract all features on a short time scale and use a LSTM to observe however computational resources were too limited.

2.6 The Power of Machine Learning

There are a non-linearities

medication on off — difference diagnosis. Would be useful in real world diagnosis.

Most sensors can only measure a small

to the quality of the machine learning model.

In this thesis we setup experiments to provide evidence of machine learning's ability to classify PD and control patients. Experiments involve:

2.7 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 C, R (rpy2) and Matlab (Matlab Engine). The code for this project is made available .... under the CRAPL license???

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 extracted using the official openSMILE [234] program, which uses the sub-harmonic summation method of f_0 estimation [235]. Most dysphonia-specific features were extracted using Tsanas' toolbox [95] with the SWIPE [232, 77] f_0 estimation algorithm. Following Tsanas (2012) [95], 120hz and 190hz were used as the mean healthy f_0 for males and females respectively. The Toolbox features code for DFA and RPDE which were proposed and implemented by Little (2007) [86] which were also used in processing the walking code.

EEG and non-linear signal analysis was performed using the PyREM library [236], which builds upon PyEEG [237], correcting some implementation flaws. The false nearest neighbour implementation in the pypsr [238] library was used to calculate the embedding dimension for the relevant non-linear signal processing algorithms. The embedding dimension chosen for accelerometer data was 6, and speech 4.

Features were Jerk was considered [145]

Most traditional machine learning algorithms were based on the standard scikit-learn [239] implementation and Gaussian processes implemented with GPy [240]. Neural networks were implemented in Lasagne [241] and Keras [242]. Hyperopt [243] was used to aid in finding the optimal hyperparameters for some models. Scikit-feature [192] was used to

implement the filter and embedded feature selection methods.

3 Conclusion

3.1 Recommendations for the Future

Machine learning in PD is clearly a young field and there remains a significant amount of work that must be completed before it can feasibly be applied in a medical context. Our recommendations for the field in general are:

1. Tools such as spectrograms and signal processing should definitely be introduced in the diagnosis process as they can identify markers missed by human senses.
2. Although machine learning is not suitable as a definitive diagnosis tool, it can help in the diagnosis process under the discretion of a trained neurologist.
3. Telemonitoring is one of the most feasible applications of current machine learning and will provide valuable feedback to
4. Focus on real world impact of.

Many ideas were inspired throughout the development of this thesis, however only a small subset were explored. Some of the more interesting ones are below:

1. Diagnosis based on before and after medicine.
2. Active learning to figure out who to monitor.
3. Improve the current set of signal processing features and develop more robust models based on short-time features, as discussed in section 2.4.3.

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] 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 disease," *Neurology*, vol. 66, no. 8, pp. 1200–1206, 2006.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] K. R. Chaudhuri and Y. Naidu, "Early Parkinson's disease and non-motor issues," *Journal of neurology*, vol. 255, pp. 33–38, 2008.

[10] 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.

[11] N. Quinn, "Parkinsonism–recognition and differential diagnosis.," *BMJ: 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 disease," *Archives of neurology*, vol. 57, no. 3, pp. 369–372, 2000.

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

[14] 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.

[15] 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.

[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] 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.

[19] A. Y. Ng, "Preventing" overfitting" of cross-validation data," in *ICML*, vol. 97, pp. 245–253, 1997.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[27] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[28] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, vol. 46. John Wiley & Sons, 2004.

[29] 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.

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

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

[32] 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.

[33] F. Eyben, *Real-time speech and music classification by large audio feature space extraction*. Springer, 2015.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[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*, 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] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

[47] 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," *ElectronicHealthcare*, pp. 704–714, 2015.

[48] 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.

[49] 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.

[50] 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.

[51] R. Nakamura, H. Nagasaki, and H. Narabayashi, "Disturbances of rhythm formation in patients with Parkinson's disease: part i. characteristics of tapping response to the periodic signals," *Perceptual and motor skills*, vol. 46, no. 1, pp. 63–75, 1978.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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 speech dataset with multiple types of sound recordings," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 4, pp. 828–834, 2013.

[57] R. Armañanzas, C. Bielza, K. R. Chaudhuri, P. Martinez-Martin, and P. Larrañaga, "Unveiling relevant non-motor Parkinson's disease severity symptoms using a ma-

chine learning approach," *Artificial intelligence in medicine*, vol. 58, no. 3, pp. 195–202, 2013.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] A. W. Przybyszewski, "Applying data mining and machine learning algorithms to predict symptom development in Parkinson's disease," in *Annales Academiae Medicae Silesiensis*, vol. 68, pp. 332–349, 2014.

[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, pp. 1332–1338, 2013.

[66] F. Wang, J. Liang, and C. Xiao, "Subtyping Parkinson's disease with deep learning models," *The Michael J. Fox Foundation*, 2016.

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

[68] 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.

[69] 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.

[70] 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.

[71] 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.

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

[73] L. V. Kalia and A. E. Lang, "Parkinson's diagnosis," *Lancelet*, p. 896–912, 2015.

[74] 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.

[75] M. Little, "Biomechanically informed, nonlinear speech signal processing.," 2007.

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

[77] 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.

[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. 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] Y. Horii, "Jitter and shimmer differences among sustained vowel phonations," *J Speech Hear Res*, vol. 25, no. 1, pp. 12–4, 1982.

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

[89] 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.

[90] 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.

[91] 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.

[92] 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.

[93] 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.

[94] 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.

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

[96] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern recognition and artificial intelligence*, vol. 116, pp. 374–388, 1976.

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

[98] 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.

[99] 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.

[100] 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.

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] 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.

[106] 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.

[107] J. Jeong, "Eeg dynamics in patients with alzheimer's disease," *Clinical neurophysiology*, vol. 115, no. 7, pp. 1490–1505, 2004.

[108] 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.

[109] 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.

[110] I. Shimada and T. Nagashima, "A numerical approach to ergodic problem of dissipative dynamical systems," *Progress of Theoretical Physics*, vol. 61, no. 6, pp. 1605–1616, 1979.

[111] 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.

[112] 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.

[113] 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.

[114] 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.

[115] 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.

[116] 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.

[117] 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.

[118] B. B. Mandelbrot, "How long is the coast of britain," *science*, vol. 156, no. 3775, pp. 636–638, 1967.

[119] 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.

[120] 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.

[121] 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," *Nonlinear Dynamics, Self-organization and Biomedicine*, pp. 66–96, 2000.

[122] 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.

[123] 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.

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

[125] 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.

[126] H. E. Hurst, R. Black, and Y. Simaika, "Long term storage, an experimental study," 1965.

[127] 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.

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

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

[130] 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.

[131] M. Martin, F. Pennini, and A. Plastino, "Fisher's information and the analysis of complex signals," *Physics Letters A*, vol. 256, no. 2, pp. 173–180, 1999.

[132] 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.

[133] 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.

[134] 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.

[135] 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.

[136] 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.

[137] 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.

[138] 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.

[139] 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.

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

[141] 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.

[142] 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.

[143] 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.

[144] 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.

[145] 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.

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

[147] 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.

[148] 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.

[149] 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.

[150] 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.

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

[152] 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.

[153] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems*, pp. 2546–2554, 2011.

[154] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[155] 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.

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

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

[158] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?," in *MLDM*, pp. 154–168, Springer, 2012.

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

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

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

[162] 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.

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

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

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

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

[167] 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.

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

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

[170] 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.

[171] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

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

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

[174] 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.

[175] 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.

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

[177] 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.

[178] 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.

[179] 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.

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

[181] 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.

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

[183] 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.

[184] 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.

[185] 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.

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

[187] 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.

[188] T. Dozat, "Incorporating nesterov momentum into adam," *ICLR 2016 workshop submission*, 2016.

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

[190] I. Loshchilov and F. Hutter, "Sgdr: stochastic gradient descent with restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[191] 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.

[192] 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.

[193] 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.

[194] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Advances in neural information processing systems*, pp. 507–514, 2006.

[195] 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.

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

[197] 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.

[198] 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.

[199] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu, "Advancing feature selection research," *ASU feature selection repository*, pp. 1–28, 2010.

[200] 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.

[201] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient l 2, 1-norm minimization," in *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pp. 339–348, AUAI Press, 2009.

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

[203] 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.

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

[205] 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.

[206] L. Breiman and P. Spector, "Submodel selection and evaluation in regression. the x-random case," *International statistical review/revue internationale de Statistique*, pp. 291–319, 1992.

[207] 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.

[208] 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.

[209] 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.

[210] 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.

[211] 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.

[212] 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.

[213] 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.

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

[215] R. L. Wasserstein and N. A. Lazar, "The asa's statement on p-values: context, process, and purpose," 2016.

[216] 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.

[217] 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.

[218] 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.

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

[220] "Michael J. Fox Foundation launches $10,000 Parkinson's data challenge." `https://www.michaeljfox.org/foundation/publication-detail.html?id=325`, 2013.

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

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

[223] 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.

[224] 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.

[225] 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.

[226] 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.

[227] 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.

[228] 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.

[229] "Parkinson's disease digital biomarker DREAM challenge." `http://dreamchallenges.org/project/parkinsons-disease-digital-biomarker-dream-challenge/`, 2017.

[230] "Feature extraction toolkit for mPower modules." `https://github.com/Sage-Bionetworks/mpowertools/blob/master/FeatureDefinitions.md`, 2017.

[231] 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.

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

[233] 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.

[234] 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.

[235] 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.

[236] Q. Geissmann, "PyREM: package for sleep staging from EEG data." `https://github.com/gilestrolab/pyrem`, 2017.

[237] 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.

[238] H. Harrison, "Phase space reconstruction: pypsr." `https://github.com/hsharrison/pypsr`, 2017.

[239] 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.

[240] GPy, "GPy: A Gaussian process framework in python." `http://github.com/SheffieldML/GPy`, since 2012.

[241] S. Dieleman, J. Schlüter, C. Raffel, E. Olson, S. K. Sønderby, D. Nouri, *et al.*, "Lasagne: First release.," Aug. 2015.

[242] F. Chollet *et al.*, "Keras." `https://github.com/fchollet/keras`, 2015.

[243] 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.