Australian National University

Applications of 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 9-34%. Hence, there is interest in investigating if machine learning can provide a more reliable and objective diagnosis.

Current machine learning literature uses simple models to differentiate between (often late stage) PD and control subjects. The setup of these experiments do not mirror real-life diagnosis as neurologists are faced with a number of diseases with similar symptoms - not just PD or control. These studies are also based on small datasets which suffer from a tendency to bias and overfitting due to limited data due to Freedman's paradox.

This thesis focuses on answering the question: "what can machine learning offer the field of PD diagnosis?". We approached this by investigating the ability for machine learning to differentiate PD and non-PD participants based on symptoms neurologists could not identify. This thesis also consolidates and replicates the body of work done on diagnosis with smartphone sensors on the much larger crowdsourced mPower dataset. We propose a number of techniques to handle the noise in the mPower data and show that the simpler machine learning models used in past works are insufficient to handle

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

Introduction

This thesis has been written for all audiences, however a background in machine learning is 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 re-iterate important information for those skim-reading and footnotes[1] provides contextual background information.

Highlight. Highlights re-iterate crucial information.

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.

Processing power was a significant limitation. All processing was done on a roasted potato.

[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 underlying causes of Parkinson's disease are still unknown.

Table 1.1: Symptoms of Parkinson's disease. Although commonly associated with tremor, only around 70% of patients experience resting tremor.

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. This test generally involves qualifying visible symptoms such as tremor and dysphonia, and assessing the

[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.

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 ranging from 9-34%. [2, 12, 13].

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 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 greatly 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 therefore prone to biases in the dataset, Freedman's paradox[4] [18] and overfitting on cross validation [18]. 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.

[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.

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

There has been some preliminary investigation into using machine learning to differentiate PD and other Parkinsonism disorders which show promising results [19, 20]. However a vast 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 visible symptoms - thus experts would likely be able to effortlessly classify all subjects correctly. The results of these papers are therefore 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 [21, 22]. A patient's progression of PD is monitored with a scale, the most common being the MDS-UPDRS [23] which quantifies the extent of 27 motor and non-motor symptoms on a scale of 0-4. 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 [24]. 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.

Highlight* 1.4 *(UPDRS). The MDS-UPDRS [23] 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. Often the set of features input into the

model is reduced using feature selection [25] or dimensionality reduction [26, 27] due to the curse of dimensionality[5] [28].

1.2 Feature Extraction and Signal Processing

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

Table 1.2 summarises some work in feature extraction relating to PD. As most datasets consist of data from a single sensor, literature generally focuses on quantifying the data for a single symptom of Parkinson's disease. Literature can be classified as diagnosing with movement or voice features and currently more research focuses on movement [29, 30].

Movement is the primary symptom considered by a neurologist in 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 changes in gait are very noticable. However, our ability to differentiate between different types of irregular gait is more limited [11]. Although most devices such as IMUs can only a capture a fraction of the information of human senses, it is possible that they are better at precisely quantifying the differences between the Parkinsonism disorders parkinsonismdifferential2.

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

Although speech is only a single component of the 44 component UPDRS [23] scale, it has received a great deal of attention in machine learning. There is also 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 [31]. Furthermore, microphones are able to

[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.

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: 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] [32, 33, 34] Smartphones [35, 36, 37]	Words and sentences [7, 51, 52]	Demographics
Postural Sway Force Plates [38] IMUs [39, 34]	Sustained vowel phonation [21, 53, 54]	UPDRS Patient Questionnaire [55, 56]
Gait Force Walkways [40, 41, 42] Video [41] Multiple IMUs [43, 44, 45]		Physical Changes Gene Expression [10, 57] MRIs [58, 59, 60] Olfactory [55] REM sleep [55, 56] Cerebrospinal Fluids [56] Gastrointestinal [61]
Handwriting [46, 47]		
Motion Capture [48]		
Tapping [49, 50]		

There is evidence that PD is heterogeneous and symptoms are present in distinct subsets [62, 63], however the underlying reasons not well understood. Studies have reported a large variation in the presence of speech dysfunction, ranging from 74-94% [64, 65, 66, 67]. To the best of the author's knowledge, no large scale movement study has been conducted. Studies report tremor in 70% of patients [68] and Akinesia in 80% [69]. However, there there is the possibility that some of these symptoms are imperceptible to a neurologist but detectable by a high resolution sensor.

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

Unless there is evidence that '*micro-symptoms*' are present in all people with PD, feature extraction in each of these areas are equally as important. We will investigate the existence of micro-symptoms in section X.X

[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.

1.2.1 General Signal Processing

There are a number of simple signal processing techniques which are informative in almost any application.

1.2.2 Voice

PD diagnosis with vocal features is a promising option for Parkinson's diagnosis as microphones are readily available and capture a comparable level of information to ears. Little et al. (2009) [21] 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 many feature extraction algorithms are currently very sensitive to minor changes in voice and speech.

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

Traditionally, the source-filter relationship of the vocal tract was assumed to be *linear*[8] and *time invariant*[9]. This assumption greatly simplifies the the analysis of speech and

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

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 many speech signals [70, 71, 72]. Non-linear signal analysis is less developed, and algorithms often involve estimation techniques. As evident in Tsanas et al. (2014) [73], extracting the fundamental frequency from sustained vowel phonation is an inexact science.

PD vocal symptoms can be broadly classified as dysphonia [74] - impairment in the production of sounds and dysarthria [75] - 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[10].

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, it is likely that both dysphonia and dysarthria related features must be considered.

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

[10] *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.

do not generalize well to the German speakers and vice versa.

The Interspeech 2015 [51] competition also featured a sub-challenge where the extent of PD dysarthria (as rated by the UPDRS) were 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 microphone. The best performing papers in this sub-challenge only managed Pearson correlations of 0.4 to 0.64 against neurologist diagnosis [76, 77, 78].

There has recently been work which reveals the benefits of working with speech. Vasqeuz et al. (2015) [79] is able to enhance noisy PD speech data using a technique proposed in Wang et al. (2007) [80] which decomposes speech into signal and noise subspaces. Orozco et al. (2015) [52] detects 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 [31] 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 [21, 81]. Dysphonia specific features generally quantify the variation in each glottal cycle, relying on an an accurate fundamental frequency estimation algorithm [73].

***Highlight* 1.7.** As dysarthria is more difficult to quantify, dysphonia based signal processing currently shows 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* [82, 83] the variation in amplitude (volume). The harmonics-to-noise ratio (*HNR*) [84] 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) [85] and is a more reliable measure of dysphonia than HNR [86].

More recently, methods used in stochastic processes have been shown to be applicable to dysphonia quantification. Detrended Fluctuation Analysis (*DFA*) was originally introduced by Peng et al. (2007) [87] as a measure of the self-affinity of a signal. Lit-

tle et al. (2007) [81] 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 detection of the fundamental frequency it may be more robust for dysphonic speakers. Little et al. (2009) [21] build upon RPDE to develop the more robust Pitch Period Entropy (*PPE*).

Tsanas et al (2012) [88] 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 as is founded on the principles of the DYPSA [89] fundamental frequency estimation algorithm. We refer to Tsanas (2012) [90] for a more detailed summary of the signal processing involved.

Mel-Frequency Cepstral Coefficients (*MFCC*) have long been used for speech recognition [91], and have also shown promise in detecting dysphonia [92]. They are the most common and often the only feature used in speech recognition systems however lack interpretability and is very sensitive to noise [93]. There are also a variety of feature sets used in general speech classification, such as the 6,368 in the 2013 Intespeech ComParE set [94]. 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 [95, 96], and it is likely that dysphonia presents itself differently depending on speaker traits. We refer to Eyben (2015) [31] for a comprehensive description of these features as well as a summary of feature sets used in speech classification.

1.2.3 Movement

As the sensors used to measure movement are varied, and signal processing techniques are often not transitive between subfields. Furthermore, movement signal processing is almost exclusively applied in dyskinesia[11] and akinesia[12] processing as unlike dysarthria, it does not share similarities with a large research area like Automatic Speech Recognition. Smartphone step and motion mode recognition[13] [97, 98] is most similar major research

[11] *Dyskinesia* describes the presence of involuntary, often 'jerky' movements.
[12] *Akinesia* is the impairment of voluntary movement.
[13] *Motion mode recognition* involves classifying whether the user has their phone in their pocket, hand, bag

area, however techniques are less transferable as measures are often more coarse.

People with PD exhibit increased tremor, particularly in the 3.5-7hz range [32] as well as distinct patterns of sway which can be quantified by recurrence analysis [34, 99]. These are best measured when the subject attempts to stand as still as possible. Both IMUs and 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) [43] and Sijobert et al. (2015) [45] 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) [44] and Diaz and Gonzalez [100] 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) [48] 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.

It is evident that despite a similar amount of literature existing in both movement and voice diagnosis in feature extraction, signal processing in the speech domain is more developed due to the more focused approach.

1.2.4 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 resolution and accuracy of smartphone sensors varies significantly between models and generalizing a model trained on one smartphone to another would be difficult - if not impossible with current machine learning models. Smartphone sensors are often of lower quality than the well-beyond thousand dollar equipment used in medical contexts.

Little et al. (2009) [21] 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) [37], Boussios et al. (2013) [36] and Arora et al. (2014) [35] 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 [101] and uninformative predictors [18].

Zhan et al. (2016) [50] 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 [49]. 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 (and 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) [35] 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 [31, 90].

1.2.5 Dynamical Systems and Chaos Theory

There are a number of methods universal in both

1.2.6 Summary of Features

Table 1.3: Summary of most state of the art features used in the various subfields of Parkinson's Disease classification. Features in the speech section can be used in movement and vice versa, however are not commonly applied.

General Signal Processing

Moments	Statistical features - mean, variation, skewness, kurtosis, etc.
Crossing Rate	Rate at which the signal oscillates past a value - usually zero or mean.
Entropy	Information content of signal
Spectral Flux	Rate at which the power spectrum changes
Fourier	The frequency domain of a signal. Fourier transforms quantifies the *power* or *energy* of certain bands (e.g, tremor between 3.5 - 7hz).
Wavelet	A variation of the Fourier transform with a different bases, allowing it to quantify both time and frequency
DFA [87, 81]	Detrended Fluctuation Analysis. Measures the self-similarity of a signal which correlate with PD dysphonia and tremor patterns.
Energy Operators [102]	Quantifies the instantaneous amplitude and frequency of a signal. Common energy operators include Teager-Kaiser (TKEO) and Squared (SEO)
HNR [84, 103]	Measures the ratio of noise in a signal (signal to noise)
Lyapunov Exponents [104]	Characterise the divergence of nearby trajectories. The largest exponent (LLE) [105] is a measure of the predictability/chaos of a system. ||MOVE THIS and has been used as a gait feature [106, 107] as well as analysing the non-linearity of speech [108, 109]

Speech

Cepstrum	The Inverse Fourier domain. Commonly taken in the mel log scale [110], resulting in the MFCC [91]. Minimal interpretability, however is the primary feature used in speech recognition [92].
Pitch [73]	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.
Jitter [83]	Measures of the variation between the length of each glottal cycle.
Shimmer [82]	Measures of the variation of amplitude between each glottal cycle.

LPCC [111]	Coefficients of an *autoregressive* model which measures how well a signal can be modelled linearly by its previous values.
GNE [85]	An extension of HNR by Michaelis et al. [85] to improve reliability in dysphonia quantification
VFER [88]	An further extension of HNR, building upon the theory of GNE.
EMD-ER [112]	Another technique developed based on non-linear speech theory to quantify signal to noise
GQ [88]	Measures the standard deviation of duration the glottis is opened vs closed.
RPDE [81]	Measures the repetitiveness of a signal, specifically designed with non-linear speech as the target.
PPE [21]	Measures the variation in successive glottal cycles
Wavelet Measures [22]	A set of 180 measures for dysphonia based on the wavelet transform to the f_0 of speech introduced by Tsanas et al. (2010)
Hammarberg Index [113]	The ratio of the strongest energy peak from 0-2kHz versus 2-5kHz. The *Alpha Ratio* is similar, measuring the largest peak 50Hz-1kHz versus 1kHz-5kHz.

Movement	
hi	pls

1.3 Machine Learning

***Highlight* 1.8.** Fundamentally, a machine learning model's goal 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).

This section abstracts the mathematical formulation for the models in favour of intuition behind their behaviour. We refer to Bishop et al. (2005) [114] for a more formal explanation of the models.

Fundamentally, supervised binary classification can be viewed as 'learning' a function which maps from a set of numerical input features to a class 0 or 1. Mathematically, a model $f : \mathbb{R}^d \mapsto \{0, 1\}$ where d is the number of features used in the model. The edge where the f transitions from zero to one can be viewed as a decision boundary (or 'hyperplane') which partitions the data into the two classes.

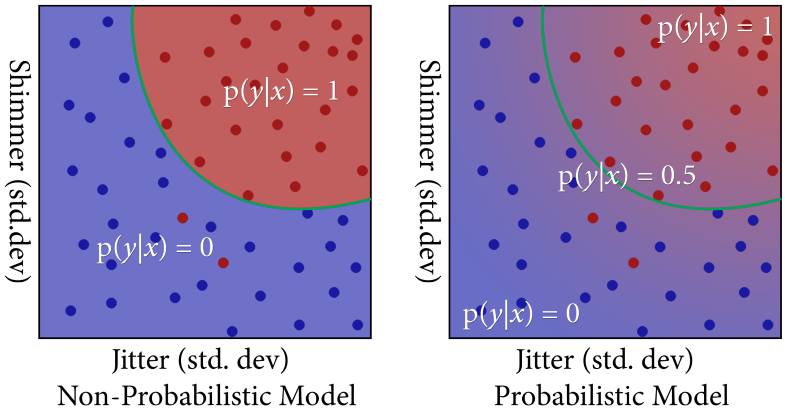

Fig. 1.1: A visualisation of binary classification with two features. Data is rarely as 'clean' as this artificial example.

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[14].

***Highlight* 1.9.** There is no 'best' model - the choice of model is informed by the data.

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

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[15] 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 known as *overfitting*.The bias-variance tradeoff [115] is a fundamental problem in machine learning, where it is very difficult to reduce bias without increasing variance and vice versa.

Models often have one or more adjustable parameters to adjust its bias and variance which are determined by intuition combined with gridsearch or large neighbourhood search [116, 117].

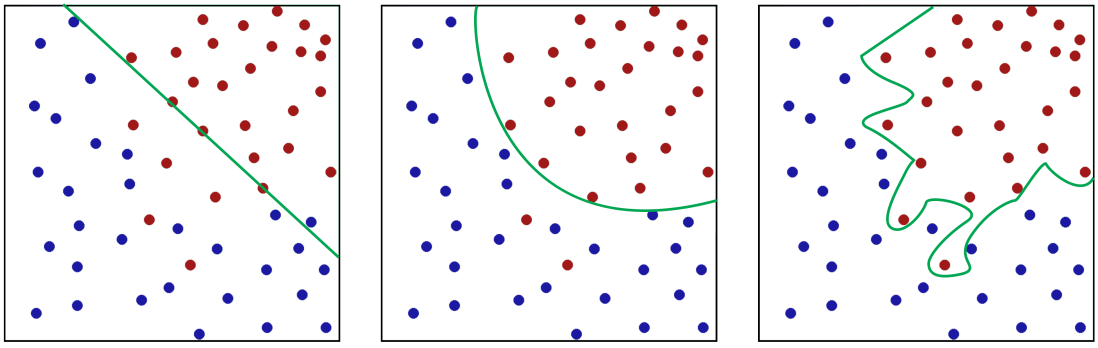

Fig. 1.2: Machine learning models and their parameters must be carefully chosen to ensure the optimal fit.

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 very simple polynomial 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 fit 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.

[15] *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'.

Like any statistics based field, careful analysis of the results is required, unfortunately this of often neglected in machine learning literature.

1.3.1 Traditional

Traditional models are the approach favoured in current literature [29] 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[16]. 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 in classification.

Random Forests [119] are derived on the concept of Bootstrap Aggregation (*bagging*) [120] 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.

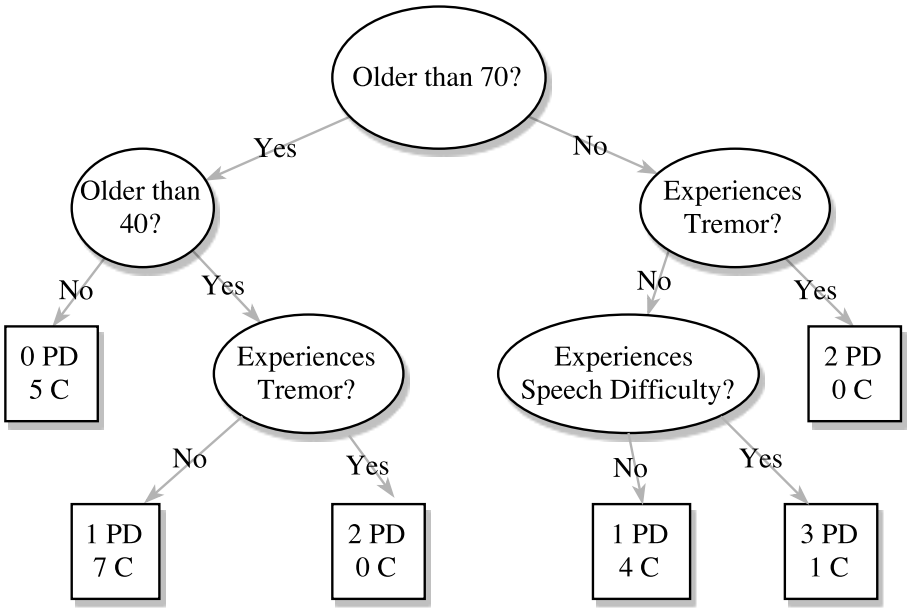

Fig. 1.3: A simple Decision Tree with cutoff depth 3. Data is split by rules until a leaf contains only one class exists or a cutoff criterion is satisfied.

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

[16] In general. Methods of generating pseudo-probability with SVMs have been proposed [118]

algorithms related to information criterion or search. A deep decision tree is prone to overfitting whereas a shallow one underfits.

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 [121]. The square root of the number of features for classification is recommended by Breiman [119] and is commonly used in most applications. Hence it is rare to perform hyperparameter tuning on random forests.

***Highlight* 1.10.** Random forests provide robust and consistent results without the need for hyperparamter tuning.

Support Vector Machines [122] are built on the concept of creating the optimal decision boundary. The motivation is to create decision boundary which maximises the margin[17] between different classes. Computationally, this is solvable and minimizing a Lagrangian will result in the optimal decision boundary. 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 datapoints, and the kernel trick transforms the raw input into the feature space of the kernel[18]. 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[19]. Kernels generally have adjustable parameters, such as the degree and constant coefficient for polynomial kernels.

The original SVM algorithm was not able to handle cases where data was not separable. Cortes and Vapnik (1995) [123] 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

[17] The *margin* is the smallest distance between the decision boundary and any of the samples
[18] Kernels perform the same role as basis functions in linear regression
[19] There is rich literature in kernel development however these innovative kernels are rarely used in practice

a large C is incentive to minimize errors.

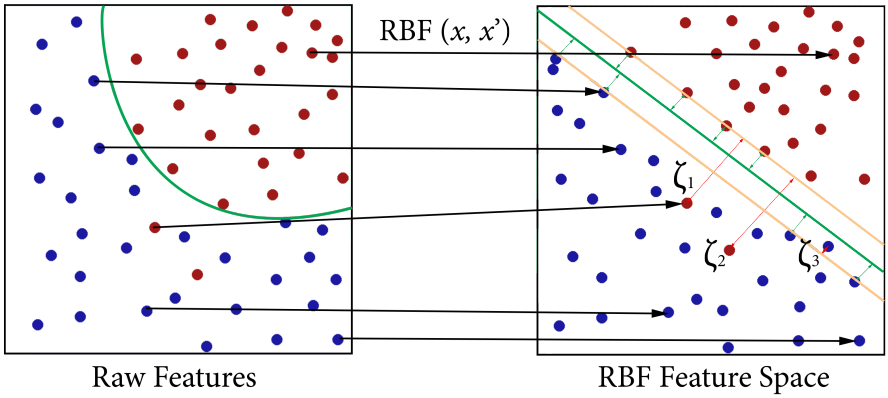

Fig. 1.4: A RBF kernel is used to transform the data into a more linearly separable space. ζ_i denote slack variables which lie beyond the margin (depicted by beige lines).

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 choice, techniques such as grid search [116] are generally used to tune these hyperparameters. 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

Although Artificial Neural Networks (ANNs) have only recently risen to the spotlight, their history begins when McCulloch and Pitts (1943) [124] introduced a model of biological neurons[20]. Rosenblatt (1958) [125] developed the perceptron, what would become a building blocks of ANNs today.

A perceptron by itself is a simple machine learning model, taking in a number of input features and outputting a value. As neurons were thought to have two states - either firing or not - a Heaviside[21] *activation function* was used.

A major breakthrough came when Werbos (1974) [126] introduced the concept of

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

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

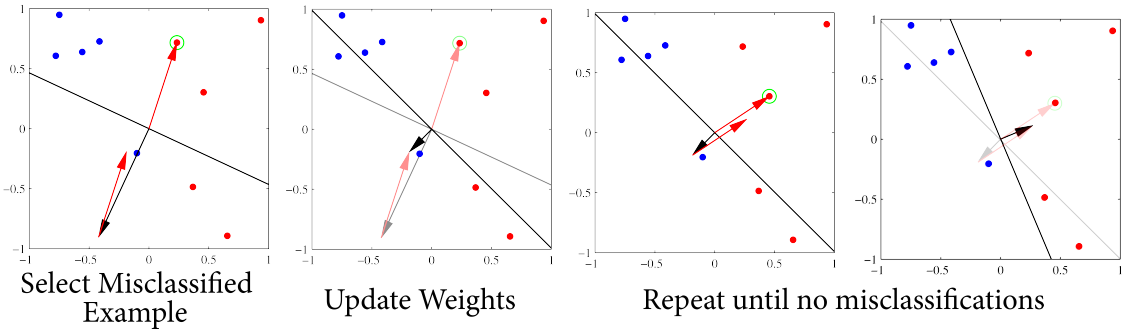

Fig. 1.5: The simple perceptron learning algorithm. The original incarnation could not handle inseparable data [125]. Images borrowed and modified from Bishop (2006) [114]

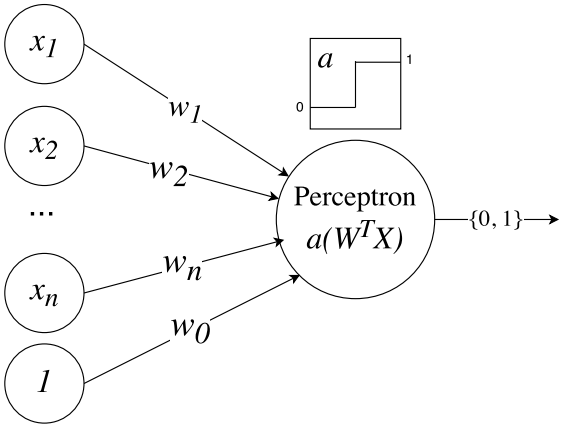

Fig. 1.6: A single perceptron node. Takes input X and learns the weight vector W to classify the output.

backpropagation, a form of gradient descent which enabled networks with multiple layers to be trained. For backpropagation to work, the activation function needed to be differentiable and the sigmoid replaced the Heaviside activation function. It was also shown that a single layer neural network which stacked enough nodes with sigmoidal activation functions was able to approximate any continuous function [127]. Unlike SVMs, a kernel does not have to be predefined - a neural network is able to learn a non-linear function of the data.

***Highlight* 1.11.** A neural network is able learn non-linear functions of the data.

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 [128]. Depth can also be exponentially more valuable than width for modelling the structure of complex non-linear data [129]. *Deep Neural Networks* are a general term for neural networks with many (generally more than 3) layers. There is still no consensus on the balance be-

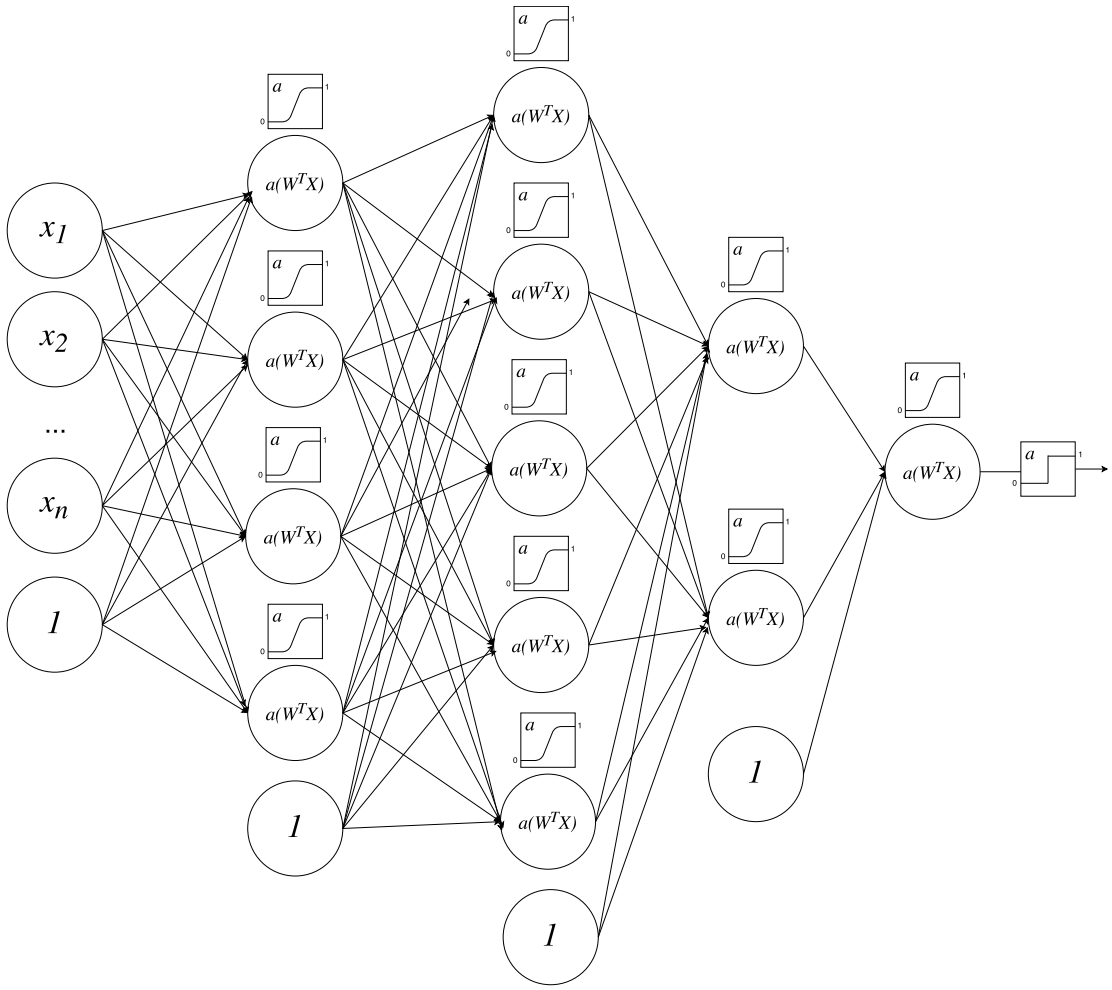

Fig. 1.7: A simple 3 hidden layer feedforward neural network with sigmoidal activations. By stacking non-linear activation functions, neural networks are able to learn any non-linear function of the input.

tween number of nodes and layers - these must be fine tuned for particular problems with intuition and search.

Neural networks are computationally expensive models and the vanishing gradient problem [130] limited the number of layers trainable with backpropagation. The recent increases in GPU power along with developments such as gradient descent with momentum [131, 132], weight initialization [133, 134] and the use of Rectified Linear Units (ReLU) activation functions [135] mitigated this weaknesses, enabling deeper and larger networks to be trained.

A neural network's ability to learn complex non-linear relationships provides a sig-

nificant advantage over traditional models where this non-linearity must be pre-defined. This is especially evident in the task of image recognition, where convolutional neural networks (*CNNs*) are able to model the relationships between pixels which constitute objects such as dogs and chairs. CNNs are able to extract features from some forms of raw data whereas these features needed to be pre-defined in traditional models. As a result, CNNs have rapidly exceeded the bounds of performance of traditional learning models in fields such as image recognition [136].

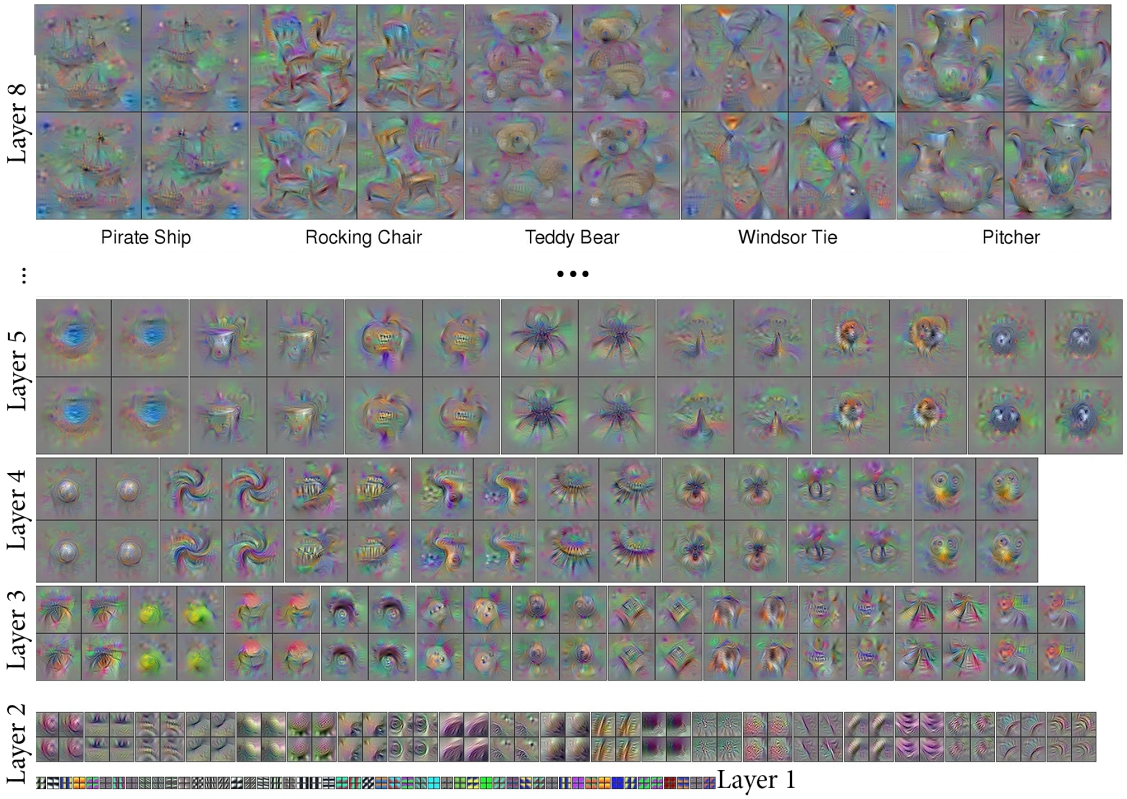

Fig. 1.8: A visualisation of a CNN from Yosinski et al. [137]. Layers capture increasingly complex relationships between pixels and act as features input into further layers.

However, neural networks also introduce a significant number of parameters which must be refined for each problem. The number of hyperparameters combined with the sheer memorisation power of neural networks increases the danger of overfitting, even on large datasets.

1.3.3 Feature Selection and Dimensionality Reduction

Te

1.3.4 Model Evaluation and Handling Overfitting

The primary goal of machine learning is to train a model which will generalize very well to new data. Accuracy over the entire dataset is evidently not a good metric, as an overfit (high variance) model 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 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. With 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 of the data-points 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. Commonly, 2, 5 and 10 are used as values of k.

In summary, we will be performing 10 fold cross-validation, randomly stratifying[22] each of the 10 sets and repeating this process 10 times. Taking the mean accuracy of each fold of cross validation, we will obtain a set of 100 accuracy values. Cross validation is performed with the same stratification sets, and Bayes Factor [138] is used to test if the mean performance of one model is greater than another when distributions are uncertain.

This decision will be justified in section 1.3.4 and more background into model selection and hypothesis testing provided.

Model Selection and Hypothesis Testing

Cross-validation is the de-facto standard in machine learning, and

Exhaustive (and LPO) and Monte-Carlo CV techniques also exist however they are not recommended by statistical literature [139, 140]. Cross-Validation is only valid if the data is independent of each other.

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

Leave one out CV provides a good estimate for a model's generalization error and allows almost all training data to be used. When the data is clean (high signal to noise ratio) LOO performs nearly unbiased estimations [140]. However LOO has been criticized for preferring models with a high variance and is less computationally feasible for large data sets [141]. Kohavi (1995) [141] instead recommends 10 fold CV in the general case.

In general, there is no agreed upon method for model selection and evaluation. Statistics is rich with penalization based evaluation[23] criteria such as Akaline/Bayesian/General Information Criterion [142, 143] and Minimum Description Length [144] however these are less suitable for machine learning as it is difficult to

However, later research by calibrated test reocmmends 10 repeats, random selected statified [145] Dieterrich recommends two fold, understimate variance [146] [147] As the size of the model increases, the

[23] *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 limit the comparability of different studies. There have been two large scale literature reviews, Alhrics et al. (2013) [29] and Bind et al. (2015) [30]. From these reviews, it is evident multiple sub-fields exist and research is often confined in their own sub-field. For example, the top papers in the Interspeech 2015 PD speech challenge [51] were 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 [148]. 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 likely exists [149] and there will exist results which are not replicable [150]. Section 1.3.4 details measures to avoid overfitting and evaluate models however their implementation is uncommon in PD 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) [35] achieves 98.0% accuracy using smartphone IMU data from 20 participants. Zhan et al. (2016) [50] performs an experiment using all features in Arora et al. (2014) as well as additional speech and tapping measures however manages only 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 [97].

We will apply a combination techniques used in state of the art on a larger dataset to assess true performance. The mPower dataset [151] has been chosen for this task and will be detailed in the following section.

The following sections will detail the experiments performed as part of the project. Although each section may be read independently, reading sequentially is recommended as later sections may reference conclusions of prior ones.

- 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 [151].

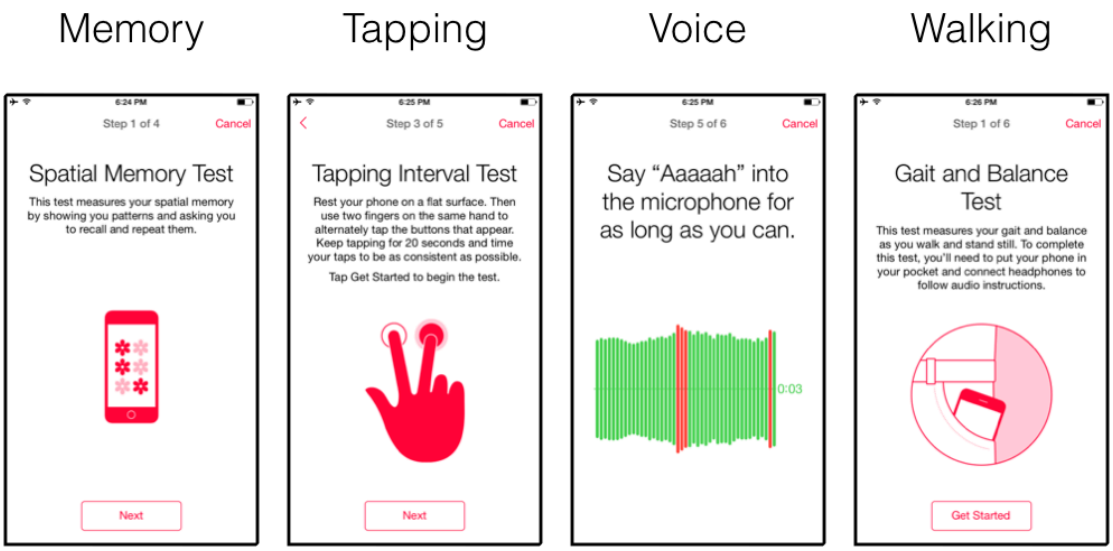

Fig. 2.1: The mPower app consists of several tasks to evaluate memory, bradykinesia, voice and gait.

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] [152, 153] 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.

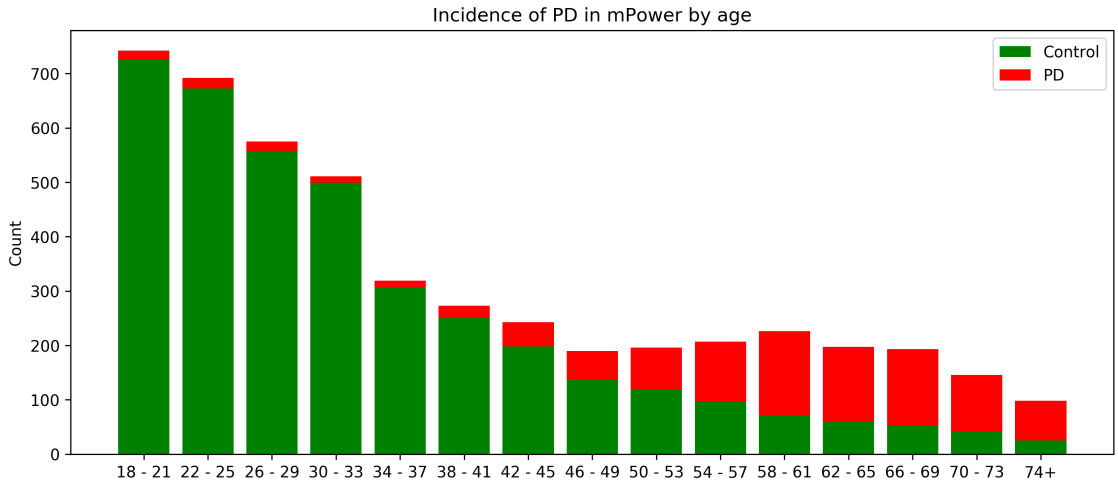

Fig. 2.2: Age is a bias in the mPower dataset as most non-PD participants are young. There are also some cases of rare young-onset PD[1].

The mPower data study was an experiment in data-collection and was not created with machine learning as a focus. 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 proper precautions [154].

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...'

[1] Assuming the honesty of the participants

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' [155].

At the time of writing, there were 65,000 speech samples from 6,000 subjects in the 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 further 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 100 ± 5 Hz. Although in-pocket IMU gait estimation exists [100], mPower does not record the parameters necessary (such as leg length) to estimate parameters other than cadence. The results of Esser et al. (2011) [19] 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.

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 Butterworth 1hz-45hz bandpass filter was applied. The high-pass filter reduces the noise from the iPhone accelerometer recordings and the low-pass removes most of the preventable sway. However the low pass filter also removes valuable sway information below 1hz [156].

A 16 second extract of motion data between 4s and 20s was used for each subject

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

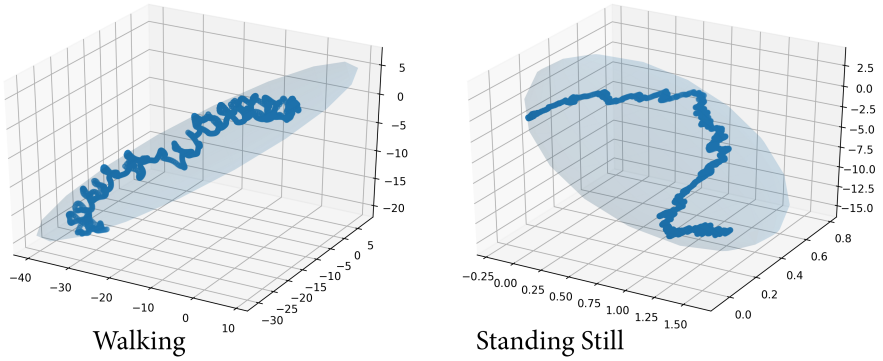

Fig. 2.3: A visualisation of device position after correcting for rotation. The gravity vector was not subtracted for a better visualization.

for feature extraction. Features specified in section 1.3 were extracted using the tools and techniques specified in section **??**. Feature Extraction was done on both the original and filtered data. The motion data was then filtered and ranked based on simple criterion such as average acceleration and the best selected for each subject.

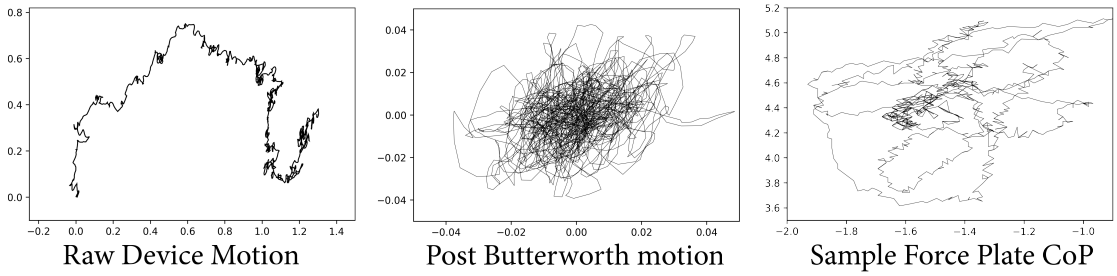

Fig. 2.4: The Butterworth filter results in a device path more similar to the centre of pressure, however low frequency sway information is lost. Note that the device motion recording is 30 seconds long while the force plate is 10 seconds.

2.2 Replicating Past Work: Traditional Models

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) [88] and the 98.0% accuracy with smartphone accelerometer data reported by Arora et al. (2014) [35].

2.2.1 Vowel Phonation

Tsanas et al. (2012) [157] 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 1.3).

Features were calculated on the 263 phonations and 10 fold, 100 repetition cross validation 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 [116]. 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 [158]. 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 almost all inputs. The results are summarised in table 2.2.

Table 2.1: Cross validation results of optimal SVM from grid 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 significantly poorer than the reported 98.6% accuracy.

The ReliefF [159] feature subset consists primarily of MFCC coefficients. MFCC is a very powerful feature and is often the primary feature in speech recognition systems. However MFCC are known for being very sensitive to noise and frequency [93, 160]. Tsanas et al. used professional grade microphones whereas mPower audio data is recorded with a smartphone microphone.

However another likely hypothesis is overfitting. The high and low MFCC coefficients are known to be rarely informative in speech recognition [161]. As the ReliefF feature set contains both the 1^{st} and 11^{th} coefficients, this implies the possibility of overfitting. If cross-validation did not divide the phonations of a per-subject level, phonations from same individuals may appear in both the training and validation sets. As MFCCs are sensitive to minor changes in frequency [160], 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 hyperpa-

rameters 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 grid search using Tsanas' 10 feature ReliefF subset.

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

2.2.3 Limits of Traditional Machine Learning

We decided to investigate the potential of traditional

2.3 Improving Results: Deep Neural Networks

There are a lo of non-linearities We aim to

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:

2.4 Implementation

?? We would like to extend our thanks to all open-source machine learning and signal processing libraries. Without these libraries, development would have been a significantly

slower process.

Machine Learning

Machine learning code was scripted in Python. Wherever possible, standard library code was used

Feature Extraction

A summary of the list of features extracted can be found in section 1.3.

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 [162] program, which uses the sub-harmonic summation method of f_0 estimation [163]. Most dysphonia-specific features were extracted using Tsanas' toolbox [90] with the SWIPE [164, 73] f_0 estimation algorithm. Following Tsanas (2012) [90], 120hz and 190hz were used as the mean healthy f_0 for males and females respectively.

The PyREM library builds upon PyEEG [165], correcting a number of implementation flaws. The

PYREM PYEEG

3 | Summary

WE SHOULD DO ACTIVE LEARNING!

3.0.1 Machine Learning

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] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

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

[28] 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.

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

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

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

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

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

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

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

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

[64] 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.

[65] 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.

[66] 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.

[67] 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.

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

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

[70] 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.

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

[72] 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.

[73] 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.

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

[75] 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.

[76] 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.

[77] 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," *INTERSPEECH*, 2015.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

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

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

[84] 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.

[85] 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.

[86] 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.

[87] 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.

[88] 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.

[89] 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.

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

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

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

[93] 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.

[94] 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," 2013.

[95] 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.

[96] 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.

[97] 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.

[98] 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.

[99] 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.

[100] 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.

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

[102] 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.

[103] 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.

[104] 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.

[105] 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.

[106] 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.

[107] 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.

[108] 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.

[109] 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.

[110] 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.

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

[112] 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.

[113] B. Hammarberg, B. Fritzell, J. Gaufin, J. Sundberg, and L. Wedin, "Perceptual and acoustic correlates of abnormal voice qualities," *Acta oto-laryngologica*, vol. 90, no. 1-6, pp. 441–451, 1980.

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

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

[116] 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.

[117] 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.

[118] 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.

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

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

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

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

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

[124] 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.

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

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

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

[128] 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.

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

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

[131] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

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

[133] 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.

[134] 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.

[135] 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.

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

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

[138] 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.

[139] 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.

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

[141] 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.

[142] 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.

[143] 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.

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

[145] 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.

[146] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[147] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," *Journal of machine learning research*, vol. 5, no. Sep, pp. 1089–1105, 2004.

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

[149] 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.

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

[151] 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.

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

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

[154] 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.

[155] 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.

[156] 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.

[157] 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.

[158] 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.

[159] 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.

[160] 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.

[161] 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.

[162] 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.

[163] 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.

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

[165] 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.