University of the Witwatersrand

Masters Dissertation

Automated Discrimination of Tremor Severity in Movement Disorder Patients Based on Machine Learning of Hand-Drawn Spirals

Author:

Kelvin da Silva

A Dissertation submitted in fulfillment of the requirements

for the degree of Master of Science

in the

Biomedical Engineering Research Group
School of Electrical and Information Engineering

September 2020

Declaration

I declare that this dissertation is my own, unaided work, except where otherwise acknowledged. It is being submitted for the degree of Master of Science in Engineering to the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination to any other university.

Signed this 17 day of September 2020

Abstract

The standard scales used to assess the severity of tremor in Parkinson's Disease (PD) and Essential Tremor (ET) are subjective and prone to human errors. One of these assessments involves performing the Archimedean spiral drawing task. Most attempts to computationally quantify the severity of tremor based on hand-drawn spirals required digital tablets, which are inaccessible to many clinical settings and foremost to telemedicine. Pen and paper hand-drawn spirals are more appropriate for these settings, and can still be scanned into a computer for automated processing. This study investigated whether deep machine learning could automatically discriminate the different severities of tremor using pen and paper drawings. The discrimination was performed using a Convolutional Neural Network (CNN). The spiral drawing dataset used for the study was small and imbalanced in terms of tremor severity groups. Data augmentation techniques were therefore studied in order to determine an optimal solution for these limitations. Binary classifications of patients with zero, mild, severe tremor and control subjects were performed. Classifications between the two movement disorders, ET and PD, and controls were also performed. The classifications were validated through a stratified 5-fold cross validation.

The accuracy of the discrimination of mild from severe tremor for both movement disorders was higher than 70.2%. The accuracy of the discrimination of control subjects from both movement disorders was higher than 85.1%. The discrimination of ET from PD yielded an accuracy of 82.4%. The discrimination of controls from zero tremor ET patients yielded an accuracy of 98.8%. The results revealed that the optimal augmentation method for this data included random translations of up to 2-4% of the spiral height and width. Data augmentation was found to improve classification accuracy by 4.6% and reduce standard deviation by 9.3%. The CNN was effective at automatically assessing the severity of tremor in movement disorder patients using pen and paper spirals. The results enable quantitative clinical insight into the manifestation of tremor in movement disorders. The methods employed indicate a feasibility to employ the simple tool of pen and paper drawings for automated screening and monitoring of tremor in movement disorders.

To Maria Graça Vieira Barreto

Acknowledgements

I would like to thank Manny da Silva, Tina da Silva, Leandro da Silva and Robyn Heard for their support in all forms while carrying out this research. I would also like to thank Prof. Vered Aharonson and Prof. David M. Rubin for their constant guidance. Without the support of these individuals, this dissertation would not have been realised.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Parkinson's Disease (PD) and Essential Tremor (ET) are two of the most common movement disorders that are characterized by tremor [1]. The currently standardised rating scales utilized to assess tremor severity of these disorders are subjective and human error prone [2, 3]. Errors in tremor severity monitoring result in patients being administered incorrect types and dosages of treatment. There is also a major challenge in differentiating between the two disorders [4]. Misdiagnoses of ET and PD also result in patients receiving incorrect treatment.

There is a need for automated assessment of tremor severity for PD and ET patients. Methods were proposed to quantify tremor using handwriting tasks, where one of the most commonly utilized tasks is the Archimedean Spiral drawing [5, 6].

The majority of studies acquired the Archimedean spiral drawings using a digital tablet [7]. This approach is technologically advanced but does not satisfy the requirements of telemedicine [8]. A more appropriate method of spiral acquisition is by pen and paper drawing [8, 9]. The static images of scanned paper drawn spirals are not suitable for traditional signal processing techniques that were employed on digital drawings [10, 6, 11, 12, 13]. Convolutional Neural Networks (CNNs) are found to be best tool for acquiring the visual attributes of pen and paper spiral drawings [7, 14, 15, 16]. These studies have only investigated part of the problem: discrimination of movement disorder and control subjects. Discrimination of tremor severity has not been considered yet.

CNNs generally require large amounts of training data in order to function correctly [17]. However, previous studies have shown that smaller pen and paper spiral datasets are still appropriate when using CNNs [18, 19]. Data augmentation techniques can be implemented to increase the number of samples in these limited databases [20, 21].

No previous studies investigated the best type and degree of data augmentation for small pen and paper spiral datasets. This information is important for developments towards a method that can automatically assess the severity of tremor in movement disorder patients.

This study determines the extent to which a CNN automatically discriminates between tremor severities of movement disorder patients using pen and paper spiral drawings. It also investigates the extent to which the CNN discriminates between ET and PD as well as a control group. The most effective type and degree of data augmentation for a small spiral drawing dataset for CNN training is also determined.

Within this document is the literature review detailing and comparing the methods and findings of existing solutions in Chapter 2. Chapter 3 outlines the research specifications and acknowledges assumptions and constraints. Chapter 4 describes the methods used to perform the study which includes spiral drawing acquisition, pre-processing, CNN architectural and training specifications as well as all experimental procedures. The significant findings from the results are discussed in Chapter 6. Recommendations for future researchers are made and the study is concluded in Chapter 7.

Chapter 2

Literature Review

This chapter provides a review of literature in aspects that are important for understanding the study. Brief background information into Parkinson's Disease (PD) and Essential Tremor (ET) as well as their treatment is provided. Existing solutions to quantitatively analyse tremor using spiral drawings are referenced. The architecture, training and validation of Convolutional Neural Networks is also discussed. A discussion on techniques to mitigate implications caused by training data shortages is provided.

2.1 Diseases associated with tremor

There are several diseases and syndromes that have symptoms of tremor. These include PD and ET [22]. Since PD and ET are the disorders of interest in this study, a brief discussion on their nature is included in this section.

2.1.1 Parkinson's Disease

PD is a progressive neuro-degenerative disorder that may be characterized by motor and non-motor characteristics [23, 24]. Motor symptoms of the disease include resting tremor, rigidity and bradykinesia [23]. A major characteristic motor symptom of the disorder is hand tremor ranging in frequency from 4-6 Hz. ET is another disorder that is characterised by hand tremor with similar frequencies [1, 25].

2.1.2 Essential Tremor

ET is a clinical syndrome of action tremor in the upper limbs in up to 95% of patients [25]. The disorder is a degenerative disease made up of a family of neuro-degenerative diseases [26]. ET is characterized by tremor frequencies ranging from 4-12 Hz [25].

Both disorders can be treated through invasive and non-invasive procedures.

2.2 Disease treatment

Treatment for PD consists of pharmacotherapy and functional stereotaxic neurosurgery [27]. These include levodopa as well as deep brain stimulation (DBS) [28, 29]. Existing methods for treating PD are primarily treatments for PD symptoms.

There have been no breakthroughs in the pharmaceutical approaches to ET treatment [29, 30]. Primidone and Propranol remain the most popular pharmaceutical approach [30]. The most popular surgical treatment of ET is DBS [29].

While these approaches are the most frequently used, they may result in the patient experiencing adverse effects [31, 32, 30]. For these reasons, non-invasive approaches to treat these disorders were investigated [33]. Magnetic Resonance guided Focussed Ultrasound (MRgFUS) is a non-invasive method for treating ET and PD [24, 33, 34].

MRgFUS involves ablating portions of a patient's brain using focussed ultrasound [24, 33]. Ultrasound transducers are phased to target specific locations in the brain. This procedure takes place while the patient is in an MRI machine [35]. A full description of the MRgFUS treatment can be found in Appendix A. Monitoring the condition of the patient involves acquiring hand tremor data from Archimedean spiral drawings [36]. The procedure to acquire this tremor is detailed in Section 4.1 of Chapter 4. Spiral drawings are also performed while the patient is nearby a MRI machine [36]. Not all spiral acquisition methods are suitable for this environment. This is discussed further in Section 2.3.3.

Evaluations of tremor are performed subjectively by clinical physicians [36]. Evaluations are performed using standardised rating scales. A quantitative approach to determining the severity of tremor needs to be developed. Currently, there is no quantitative evaluation which can determine the severity of movement disorders in patients.

2.3 Disease Severity Monitoring

The decisions on treatment suitability and course necessitate accurate diagnosis and severity monitoring[1, 37]. These decisions are facilitated by tremor rating scales. Tremor rating scales cannot be used to conclusively distinguish between types of movement disorders. Rating scales have been created for the purpose of monitoring disease severity and progression using the qualitative characteristics that have been common in each disorder [37]. However, some studies have demonstrated that certain qualitative characteristics are common in both PD and ET [37]. These clinical misperceptions include:

1. Rest tremor = PD
2. Action tremor = ET
3. Bilateral and symmetric tremor = ET
4. Head tremor = ET
5. Jaw tremor = PD
6. Bradykinesia = PD

These misperceptions make the rating scales ineffective as diagnostic tools. The majority of existing approaches in movement disorder severity monitoring involve subjectivity. Two of the most popularly utilized rating scales are briefly discussed in this section: Unified Parkinson's Disease Rating Scale (UPDRS) and Clinical Rating Scale of Tremor (CRST).

2.3.1 UPDRS

The UPDRS is a standardised rating scale used to diagnose and measure the severity of PD. The scale is used to evaluate the severity of motor and non-motor symptoms in all currently existing forms of treatment [2]. Relevant tasks are performed and ratings are provided by a trained physician [2]. While the procedure is standardised, it is a subjective assessment based on the opinion of the physician. There is considerable variance between the ratings from different physicians [38].

The Motor-UPDRS can be divided into mild and severe cases using the ranges of 1-20 and 20+ respectively [39]. The maximum possible score that can be acquired from

Motor-UPDRS is 108 [39]. The mild-severe threshold is at 19.5% of the maximum score [39]. This ratio allows for the mild-severe threshold to be calculated on scores obtained using portions of the Motor-UPDRS. This was useful for dividing the PD dataset in this study into categories of severity.

The UPDRS is standardised for PD assessment only [2]. The CRST is used to assess general tremor [40]. It also offers more objective forms of tremor evaluation.

2.3.2 CRST

The CRST is a rating system used to evaluate ET as well as general tremor [40, 41]. Different tasks are performed and are rated by a physician. One of these tasks involves drawing an Archimedean spiral [40, 41]. This form of data acquisition is discussed further in Section 4.1.1. An Archimedean spiral is shown in Figure 2.1. The physician can perform physical measurements on the Archimedean spiral drawing, which is a more objective assessment [40, 41]. While this decreases subjectivity, there is still a risk of human error.

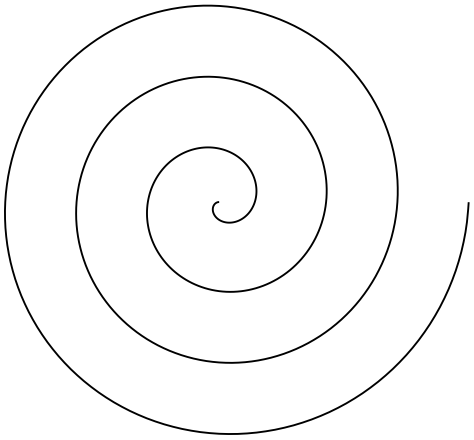


Figure 2.1: The Archimedean spiral.

A fully implemented CRST evaluation results in a maximum possible score of 144 [40]. Tremor severity is considered mild if the CRST score is within 24% of this maximum score (1-35) [40]. The ET dataset in this study was categorized according to severity using this threshold.

Bedside examinations of tremor severity entail subjectivity. No studies have determined how accurately trained physicians can discriminate between mild and severe

tremor (Binary classification). The standard error of measurement (SEM) in physician scores has been evaluated [42]. The value of SEM was never higher than 30% [42]. It can be inferred that any classification accuracy greater than 70% can be considered as a well performing classifier between different tremor severities [42]. For the purpose of routine and therapeutic trials, a clinical rating scale can produce reliable results [43]. This reliability is improved through the use of handwriting tasks to assess tremor severity [40, 43]. Archimedean spiral drawing is the most popular form of tremor data acquisition [5, 6]. It should be noted that while the degree of assessment subjectivity is decreased, it is still present. Spirals drawn by zero tremor and no disorder subjects are very similar to one another but have subtle differences. Detecting these differences can be a challenge for a trained physician [37]. These differences can present as changes in pen pressure when drawing spirals and small oscillations that are only visible when the drawing is analysed from a sample or pixel level [6, 9, 10, 11, 12, 13, 44].

There are different manners in which spirals drawings can be acquired. Certain characteristics are available for analysis depending on which spiral acquisition method is chosen.

2.3.3 Archimedean Spiral Acquisition

Archimedean spirals can be acquired by the patient drawing on a piece of paper or on a digitizing tablet [7]. The acquired drawing can then be analysed.

Based on the way the patient was instructed to draw a spiral, different drawing information could be acquired. There was a large amount of literature based on analysis of Archimedean spirals acquired from digitizing tablets. There was a shortage of literature regarding the analysis of spirals acquired from paper scans.

Telemedicine is the field of remote diagnosis and monitoring of patients by means of telecommunications technology [8]. The following reasons summarize the need for further investigation and analysis of spirals acquired with pen and paper:

1. All patients, doctors and hospitals are more likely to possess pen and paper compared to tablets.
2. Tablet acquired spirals have not been standardised.
3. A tablet cannot be used in or near an MRI machine while pen and paper can.

4. Paper acquired spirals meet the needs of Telemedicine more appropriately than tablets do.

Approaches analysing spirals that have been acquired from pen and paper was the focus of this literature review [45, 8]. However, approaches using tablet acquired spirals were acknowledged.

2.4 Current Tremor Quantification using Hand Writing

Approaches towards a quantitative measure of hand tremor severity using spiral drawings have been studied. Data required to achieve this has been acquired from both tablet and paper drawn spirals. Most approaches make use of tablet data.

Mean Relative Tremor was a quantitative feature to measure the amount of tremor in a spiral drawing [9]. It was defined as the mean and standard deviation of differences between the radius of a given sample and its left nearest neighbours [9]. This measurement was dependent on the number of manually selected sample points on the drawing. The displacement of the sample points were used to measure radius difference. The study did not detail the feature's correlation with existing tremor rating scales. All other approaches of tremor quantification using Archimedes spirals made use of tablet features.

The Degree Of Severity equation correlated best with the UPDRS score of studied patients [10, 6]. Tablet spiral features of speed and pressure were used to develop a Composite Index of Speed and Pen-Pressure, which correlated strongly with the UPDRS scale [11]. The effectiveness of treatment was quantified by using mathematical parameters relating to spiral radius and oscillation [12]. Drawing acceleration and velocity were used to engineer fractional derivative features [13]. These engineered features were used in a supervised learning regression task to predict UPDRS scores.

There was less literature focussing on ET severity quantification from spiral drawings. Tremor metrics based on spectral features were calculated from tablet drawings [44].

There is a need for a model to automatically assess tremor severity using pen and paper spiral drawings. An investigation into the validation of the model is also necessary.

2.5 Quantitative Analysis of Archimedean Spirals

No studies have used pen an paper drawn spirals to distinguish between different severities of tremor. Existing literature involved using this form of data to discriminate between a movement disorder and a control group. Tremor severity prediction based on spiral drawings can be reduced into a classification problem. Therefore, investigating this literature was appropriate. More details of the specific problem reduction in this study is discussed in Section 3.4 of Chapter 3.

Figure 2.2 summarises the available features from spirals acquired from tablets and paper. Features available from paper acquired spirals include radial measures, error features and statistics [46, 47, 7, 48]. CNN-learned features are those attributes learned by a deep learning model. Deep learning is discussed further in Section 2.6.

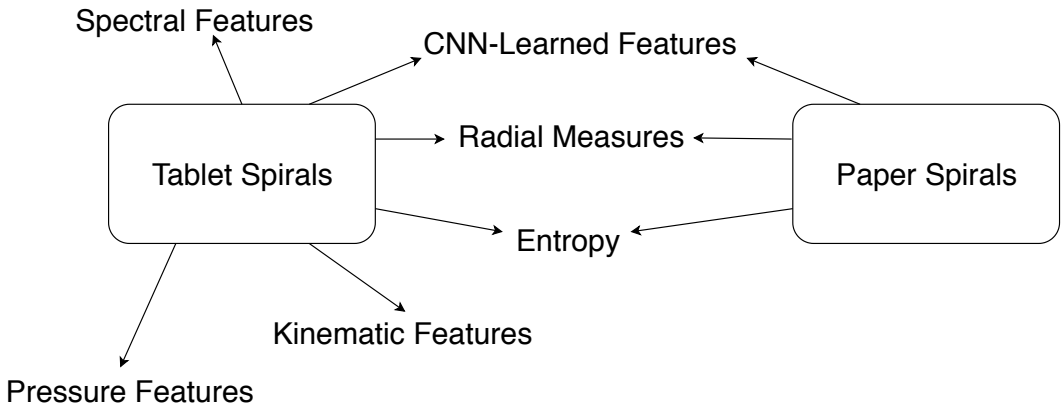


Figure 2.2: Summary of features types that can be extracted from tablet and paper acquired spirals.

Spiral drawings could be converted into vectors of radii and angles using a polar coordinate transformation [9, 49, 50]. Radial measures were used to discriminate between PD and control groups. Radial measures had high significance as a feature that distinguished between patients and controls [49]. A downfall was that there were no automated techniques to extract these features from pen and paper spiral drawings.

Improved classification accuracy was achieved by using a Structural Co-occurence Matrix (SCM) to distinguish between PD and control patients [50]. The resultant output of the SCM were image attributes. These attributes were used in a machine learning classifier as an approach toward diagnosing PD [50]. This technique was not automated. Its performance was also second to that of CNN-Learned features.

Spiral drawings could be used as input to a Convolutional Neural Network (CNN) [19]. This technique has been used to discriminate between PD and control subjects. Studies suggested that visual attributes of handwriting offered comparable classifier performance compared to dynamic features [7, 14, 15, 16, 19]. This made pen and paper features as powerful as tablet features. Using pen and paper to obtain spirals is more appropriate in terms of telemidicine requirements as mentioned in Section 2.3.3. Visual features are also acquired automatically using CNNs. Visual attributes offer comparable classification performance to dynamic features due to the architecture of CNNs. Architectural details of CNNs are discussed in detail in Section 2.6.

2.6 Deep Learning with CNNs

The use of CNNs for classification has two major benefits in practice, namely [51]:

- **Local Invariance**: When classifying whether an image contains an item, the positioning/location of the item is not important. A classification is only dependent on whether or not the item is present in the image.
- **Compositionality**: The composition of a local patch of lower level features into higher level representation. Image pixels within a neighbourhood can be represented as edges, shapes and higher dimensional objects. The higher level representations are used for performing the classification.

These benefits are achieved through the implementation of multiple layer types within a CNN .

2.6.1 Layers

CNNs utilize several layers that are only connected to a small region of the layer before it. This local connectivity enables large amounts of network parameters to be specified. The most common layers are summarized below:

1. **Convolutional** - Receives an input image having a particular width W_{in}, height H_{in} and depth D_{in}. The W_{in} and H_{in} refer to the physical dimensions. D_{in} refers to the number of channels in the activation map of the layer [51]. The number of $F \times F$ filters/kernels, K, must also be specified. The stride S, is

the number of pixels that the kernels are shifted by in every step of convolution. Zero-padding P is an optional parameter that can be specified [51]. Figure 2.3 highlights the description of this diagrammatically. The output dimensions of

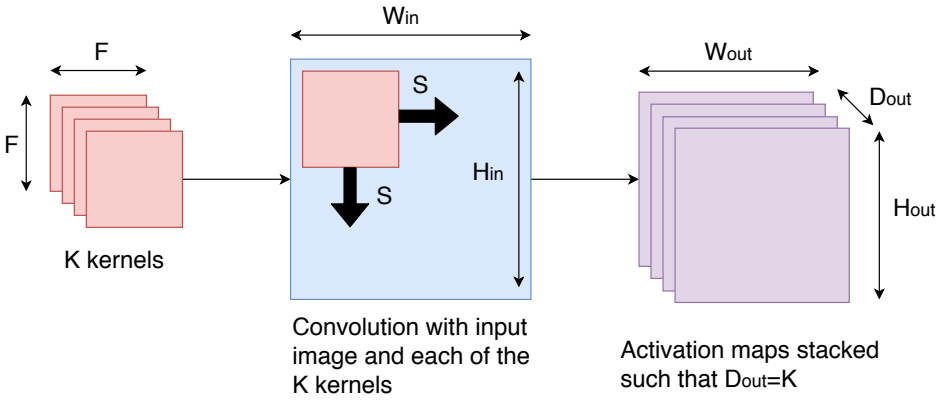


Figure 2.3: Process and high-level architecture of the convolutional layer of a CNN.

this layer are $W_{out} \times H_{out} \times D_{out}$, where:

- $Wout = ((W_{in} - F + 2P)/S) + 1$
- $Hout = ((H_{in} - F + 2P)/S) + 1$
- $D_{out} = K$

2. **Activation** - No parameters/weights are learned inside an activation layer. In this layer, a non-linear activation function such as a Rectified Linear Unit (ReLU) or sigmoid function is applied. The output of an activation layer is always the same dimension as that of its input.

3. **Pooling** - The main function of this layer is to reduce the spatial size of the input volume [52]. In doing so, the amount of parameters and computation required for training is reduced. This also aids in the control of over-fitting. The layer operates on each of the depth slices in an input volume independently by utilizing the max or average functions. The parameters to be specified in this layer are receptive field or pool size F and stride S. The W_{out}, Hin and D_{out} of this layer are related to these parameters as follows:

   - $W_{out} = ((W_{in} - F)/S) + 1$
   - $H_{out} = ((H_{in} - F)/S) + 1$
   - $D_{out} = D_{in}$

   The procedure and result of max pooling is depicted visually through Figure 2.4.

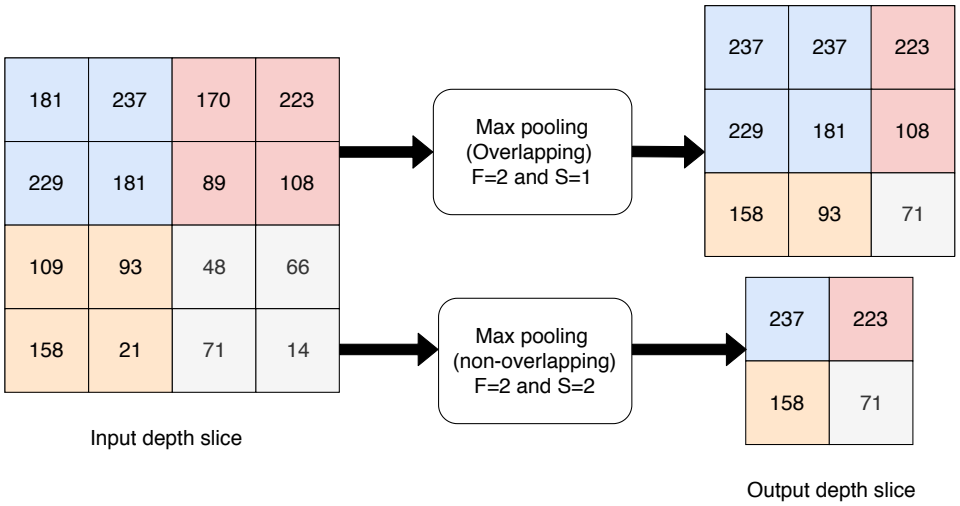


Figure 2.4: Overlapping and non-overlapping max pooling parameters and output (Adapted from [52]).

4. **Fully-Connected** - this layer is always placed at the end of a network and is fully connected to all activations of the previous layer [51]. It is common to use a fully connected layer before the application of a softmax classifier or a sigmoid activation function [51]. This layer uses the high level features extracted from previous layers to classify images into classes. These features are learned non-linearly.

   This layer contains a large number of parameters that require significant computational demands in training [53]. Computational demands can be reduced through applications of dropout techniques [53, 54].

5. **Batch normalization** - This layer is used to normalize activation of a given input volume before it is propagated to the next layer [51]. This layer increases classification accuracy and training stability. Computational demands during training are also reduced [51, 55]. A batch normalization layer should be implemented after activation layers in order to yield higher accuracy [56, 57].

6. **Dropout** - This layer randomly removes connections between consecutive fully connected layers based on a specified probability [58]. This causes the architecture of the network to be randomly manipulated at training time. This is a form of regularization [51, 58]. The layer aids in reducing over-fitting. It makes multiple layers responsible for activations due to similar patterns in input volumes [51].

Convolutional and fully connected layers are the only CNN components that have trainable weights/parameters. Training a CNN and the callbacks for optimizing the

process will be highlighted in the next section.

2.6.2 Training

Training a CNN involves minimizing the error between an expected output and the actual output based on a 2-dimensional input. This error is known as the loss function [51, 59]. After each iteration, trainable parameters are manipulated based on the size and sign of loss [51, 59]. The learning rate is the scalar factor used to control how much the trainable parameters change per iteration of training (epoch). For the loss to converge to a point where it does not improve further, an optimization algorithm must be specified. The desirable result from training is a final loss which converges to the lowest value possible [51, 59, 60].

It is common to constantly calculate the training loss and validation loss during training [51]. Training loss is the result when trying to relate a training dataset input image to its true class label. The validation loss is the result based on the prediction of the CNN [59, 60]. This loss shows how effectively the CNN can predict the true class label of an image that was not used to train it. The changes in training and validation loss per epoch should be closely related to one another and converge [59, 60]. The loss values will diverge for these reasons:

1. The CNN model architecture is not suitable to perform the desired classification
2. The model is over fitted

Over fitting can be caused by unoptimized CNN hyper-parameters and training for too many epochs [51]. CNN hyper-parameters are optimized by continuously changing parameters and recording corresponding performance metrics. Care should be taken to not over optimize, resulting in over-fitting [51]. There are callbacks used to ensure that training is automatically stopped once certain criteria are met. These callbacks aid in shortening training time and reduce over-fitting [51, 61]:

1. **Early Stopping** - CNN training is stopped once the validation loss or a selected performance metric does not improve for a specified number of epochs.
2. **Model Check-pointing** - During training, the model with the lowest validation loss or user specified performance metric is saved. Any subsequent models that have improvements can be saved.

These callbacks can be implemented easily through the use of Keras [61]. Keras also allows for other hyper-parameters to be tuned easily. Tuning these hyper-parameters has also been found to be effective in reducing cases of over-fitting [51]:

1. **Learning rate decay** - This increases the chance of locating the global minimum in terms of loss landscape by decreasing the learning rate after every training update. The learning rate (lr_{new}) is updated for every training iteration (i) based on Equation 2.1. The initial learning rate (lr_0) must be optimised. In practice, it is common to specify the *decay* rate as the quotient of lr_0 and the number of epochs needed to train the CNN [51, 52].

$$lr_{new} = lr_0 * (\frac{1}{1 + decay * i}) \tag{2.1}$$

2. **Momentum** - The momentum algorithm calculates an exponentially decaying moving average of current and previous update gradients and continues to update in the same direction. The lr_{new} is increased additively if previous updates are in the same direction. Updating power is decreased if previous gradients oppose the update direction. The momentum algorithm prevents training from becoming stagnant in loss landscape local minima and speeds up the location of the global minimum [51]. In practice, momentum is commonly specified as 0.9 [51, 52].

3. **Nesterov Acceleration** - This algorithm is closely related to momentum. The difference between the two algorithms is in the gradient computation phase [62]. Momentum calculates gradient using current and previous parameters whereas Nesterov acceleration computes interim parameters and uses them to calculate the gradient. The parameters are updated by multiplication of the calculated gradient and learning rate. This algorithm provides prevention against exploding gradients caused by momentum, which is common in the early stages of training [62].

These parameters and callbacks cannot function effectively across all forms of datasets. These attributes require tuning through performing several experiments. The useful information from these experiments are the changes in performance metrics. Useful performance metrics and evaluation procedures for small datasets are discussed in Section 2.6.3.

2.6.3 Evaluation and Performance Metrics

The performance of any CNN model can be validated by using different performance metrics. The metrics that were used to validate the results in this study are summarised in Table 2.1 [63, 59, 64].

Table 2.1: Equations of each performance metric for classification [64, 65]. The values tp, tn, fp, fn represent the number of true-positive, true-negative, false-positive and false negative classifications respectively.

Performance metric	Equation
Accuracy	$\frac{tp+tn}{tp+fn+fp+tn}$
Precision	$\frac{tp}{tp+fp}$
Recall	$\frac{tp}{tp+fn}$
F1-Score	$\frac{2(Precision \times Recall)}{Precision+Recall}$

Previous studies involving CNNs and spiral drawings only used accuracy as a classification performance metric [18, 19]. Accuracy is the ratio of correctly predicted labels to the total number of labels that were predicted. Accuracy can be a misleading metric in cases where classes are imbalanced. Since imbalance of classes in medical datasets are frequent, the additional performance metrics shown in Table 2.1 were included in the analysis of results to mitigate these effects [65].

Precision is a measure of how precise the model is. It determines how many of the labels that were predicted as true positive are actual positives [51, 65]. It is a favourable metric to use when the cost of false-positive predictions are high. In similar situations, it is also important to determine the recall metric. Recall is used to determine how many of the actual positives were captured by the model by labelling them as true positives [51, 65]. The metric is used to determine how frequently the model predicts false-negatives. F1-Score is a function of precision and recall which provides a balance between these metrics [51]. In cases of class imbalances, F1-Score is contributed largely by false-positives and false-negatives whereas these predictions remain hidden when using accuracy [65].

Performance metrics should be used to validate the performance of a CNN on different

training and testing datasets. This can be achieved by performing a stratified k-fold cross validation.

Stratified K-Fold Cross Validation

K-fold cross validation is used to evaluate the classification capabilities of machine learning models [66]. It accounts for the effect of over-fitting due to shortage of data [66]. The only parameter required for the technique is K. This refers to the number of groups or folds that the dataset will be split into. Figure 2.5 illustrates a 5-fold cross validation [67]. The final evaluation is provided as a mean and standard deviation of all the performance metrics across all folds.

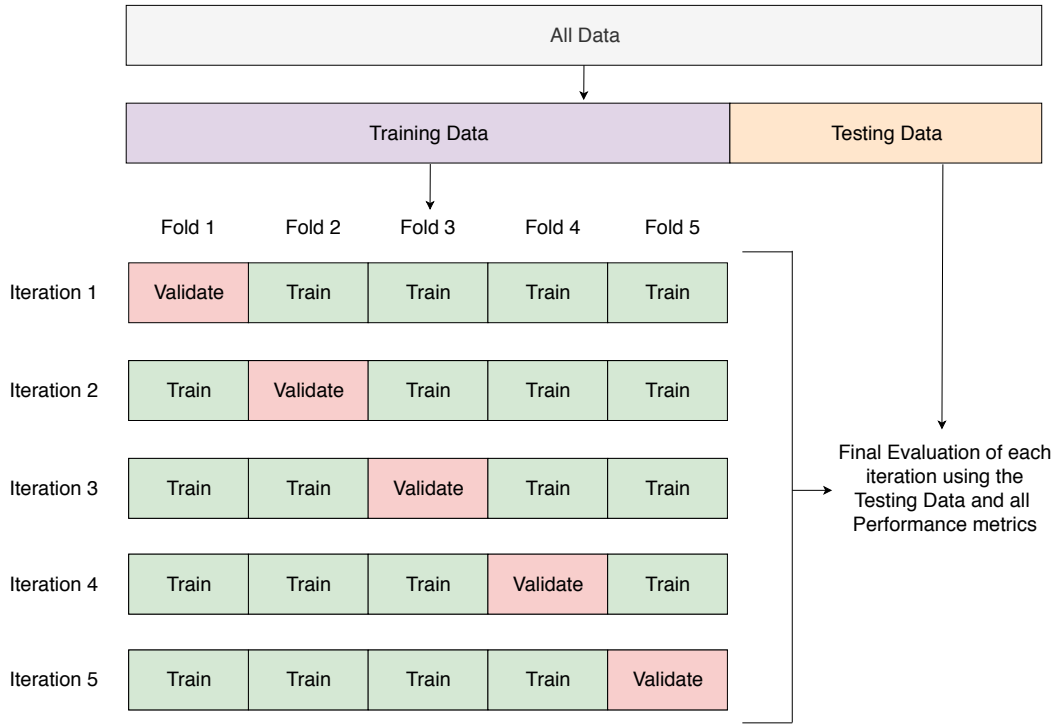


Figure 2.5: The procedure of a 5-fold cross validation. The method involves the first fold being used as a validation set while the remaining K-1 folds are used as the training dataset. This process is repeated K times such that each fold has made up the validation at some stage. Each of the folds are evaluated using an unseen testing dataset that was aside before commencing the K-fold cross validation.

The value of K is usually 5 or 10 [67, 68]. These are popular choices since they yield results with low degrees of bias and variance [67]. For lower values of K, computational requirements are less but the validation is more biased [69]. For larger

K values, computational requirements increase, validation becomes less biased but has more variability [69]. A trade-off exists when selecting a value of K.

The stratified variation of cross-validation ensures that each fold has the same ratio of all classes in the dataset [69]. This prevents class imbalances throughout a k-fold cross validation. It is best practice since it reduces validation bias and variability compared to regular cross-validation [69]. As will be discussed in Section 2.7.2, certain techniques can be applied if there are class imbalances in the entire dataset.

The quality of a model evaluation is not only dependent on the performance metrics and validation techniques. The total size of the training database and the proportion of its classes are also important. It is useful to know how to mitigate the effects of these undesirable database characteristics.

2.7 Small Datasets

A common problem with studies involving medical data are data shortages [17]. Studies have been performed in the deep learning domain using relatively small datasets. It is possible to use small datasets in deep learning. However, the implications on CNN performance must be acknowledged. This section discusses recommended mitigation techniques to minimize these effects.

2.7.1 Studies using relatively small datasets

Most deep learning studies for image classification used databases containing at least 1000 images per class [51, 52]. However, studies have achieved success having much less data. Database class sizes as small as 124 PD and 140 control subject spiral drawings have been used [18]. In the same study 30%, 20% and 50% of the dataset was used for training, validation and testing purposes [18]. This implies that 37 and 25 samples were used in each class for training and validation purposes respectively. The highest CNN classification accuracy was 90.38%. Another study used a total of 576 (36 PD and 36 control) drawings to discriminate between the two classes [19]. The highest accuracy obtained using spiral drawings was 76%. The ratios of data used for training and validation purposes were not specified. The study made use of a pre-trained CNN called AlexNet [19]. This was a method of mitigation against effects of data shortages and is known as transfer learning [19, 60].

There are other techniques used to address the implications of training a CNN with a small dataset. The mitigation techniques used to address specific implications are discussed in Section 2.7.2

2.7.2 Addressing Issues with Small Datasets

A requirement of using deep learning algorithms, such as CNNs, is a large volume of data [17]. However, there are means to address shortages of data where deep learning algorithms are required. Table 2.2 highlights the main implications as well as possible mitigations from having a shortage of data.

Table 2.2: The implications of having small datasets in deep learning and the techniques to mitigate them.

Implication	**Mitigation**
Lack of Generalization	Data Augmentation
	Data Generation
Data Imbalances	Class Weighting
	Up-sampling
	Down-sampling
Optimization Difficulties	Transfer Learning
	Problem Reduction

Lack of Generalization

A significant implication of having a small dataset is that the CNN would have poor abilities to generalize and result in an over fit model. Data augmentation is the most effective technique for improving model generalization. This technique is the process of applying different operations to image datasets in order to increase their size [20, 21]. The images used in this study did not consist of colour and hence, colour augmentation was less applicable compared to affine augmentation [21, 70, 71]. Affine augmentation techniques include rotations and translations and can be implemented with different degrees. A CNN's property of translational invariance would have different degrees of significance depending on the type of data that was used to train the model. The best augmentation method for increasing a CNN's receptive field cannot be known without conducting experiments. There have been no studies

detailing the optimal type and degree of affine augmentations for small spiral datasets. Performing research in this area would be a contribution to studies with these types of datasets. Generalization improves by creating variations of available data. Data can also be synthetically created using Graphic Adversarial Networks (GANs) [72]. GANs require more data compared to data augmentation in order to be effective [73].

Class Imbalance

In binary classification, having a dataset with imbalanced classes can impact the performance of the CNN. The CNN is insufficiently trained on the minority class, making it biased toward the majority class.

An effective technique for addressing this issue is class weighting [74, 71]. This makes the CNN account for a class imbalance by adding these weights to the loss function [74]. The minority class is weighted more compared to the majority class, making the representation of the two classes closer to equal. Scikit-Learn offers functionality for implementing class weighting easily [75].

Up-sampling involves copying samples within the minority class in order to balance its size compared to the majority class [76]. Down-sampling involves removing data from the majority class so that both classes are similar in size [74, 76].

Optimization Difficulties

Optimization difficulties refer to the inability of the CNN training procedure to locate the global minimum in the loss landscape [77]. A method of addressing this issue is transfer learning. Transfer learning is the approach of utilizing a pre-trained model, designed for a specific purpose, and adapting its learned parameters for a different purpose [60]. This technique is only effective if the data that was used to train the pre-trained network is similar to the current data of interest [78].

Problem reduction is another method of improving the optimization difficulties when training CNNs. This implies narrowing the scope of the initial machine learning task. The initial task may be too great based on the dataset size but more acceptable if the task size is minimized [77]. This may involve changing the format of the data. In other cases it involves reducing a regression task into a binary classification [77].

There is no general solution to these data shortage implications. Each mitigation technique requires experimental testing and tuning to be suitable for a particular domain. In the next chapter, the research details of this study will be discussed. This chapter will describe the problem statement and outline the research questions of the study. All significant assumptions and constraints will also be acknowledged.

Chapter 3

Research Specifications

This chapter outlines the problem addressed by the research that was performed. Contributions were made by addressing specified research questions that have been highlighted. It has been acknowledged that the study has been performed under certain assumptions and constraints.

3.1 Problem Statement

Tremor severity assessments, such as UPDRS and CRST, largely entail subjective and observational assessment. Severity monitoring is therefore subject to errors and bias. The existing approaches toward quantitative tremor severity assessment require digital tablets, and therefore have downfalls in the following areas:

1. **Accessibility**: A patient may not possess a tablet.
2. **Affordability**: Not all healthcare systems can afford to provide patients and doctors with tablets.
3. **Usability**: Many elderly people are unfamiliar with handling technology.
4. **Environment suitability**: Tablets cannot be operated in the immediate vicinity of MRI machine (MRgFUS treatment)

In order mitigate these issues with a deep learning approach, a large volume of medical data may be required. Data shortages are common in the medical domain. No studies have determined an optimal method of addressing data shortages.

Addressing the problems with tablet usage and shortages of data was seen as a contribution toward the development of a home monitoring system for tremor severity. The contributions were made by addressing specific research questions.

3.2 Research Questions

The primary research question of this study is as follows:

To what extent can a Convolutional Neural Network discriminate different tremor severities using pen and paper drawn spirals?

The specific aspects of this research question that were studied were:

1. To what extent can a CNN discriminate mild from severe tremor in ET and PD patients?
2. To what extent can a CNN discriminate ET from PD?
3. To what extent can a CNN discriminate movement disorder patients from controls?
4. To what extent can image data augmentation enhance discrimination performance in small data conditions?

Each of these aspects employed different subsets from the spirals dataset and entailed assumptions and constraints. These are discussed in Section 3.3 and Section 3.4 respectively.

3.3 Assumptions

The available data consisted of spiral drawings with corresponding tremor severity scores (UPDRS and CRST) and disorder diagnoses (ET and PD). These scores and diagnoses were performed by medical practitioners. This study assumed that this information was correct, both in terms of the tremor severity and of the diagnosis - PD or ET.

3.4 Constraints

The constraints of this study were:

1. Imbalance between classes of tremor scores and the two diseases.
2. Small dataset size for CNN training

The distribution of CRST and UPDRS data that were used as labels for the spiral drawings is shown in Figure 3.1. This distribution of data is not suitable for performing a regression using a CNN [79]. This motivated the reduction from regression or multiple-class discrimination to a binary classification between different tremor severity categories or different movement disorders. The small dataset sizes of these severity classes was another constraint of the study.

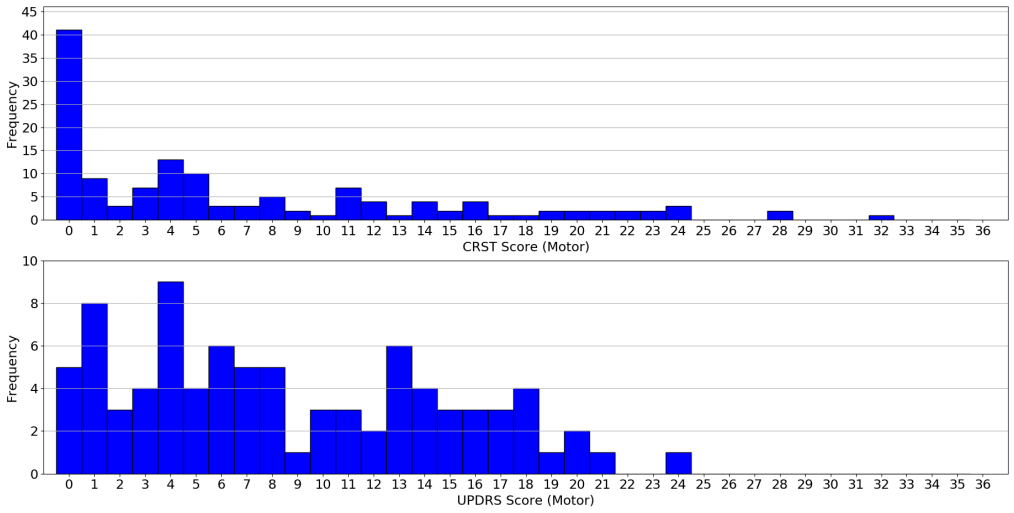


Figure 3.1: The distribution of CRST and UPDRS tremor scores for the available spiral dataset.

Chapter 4

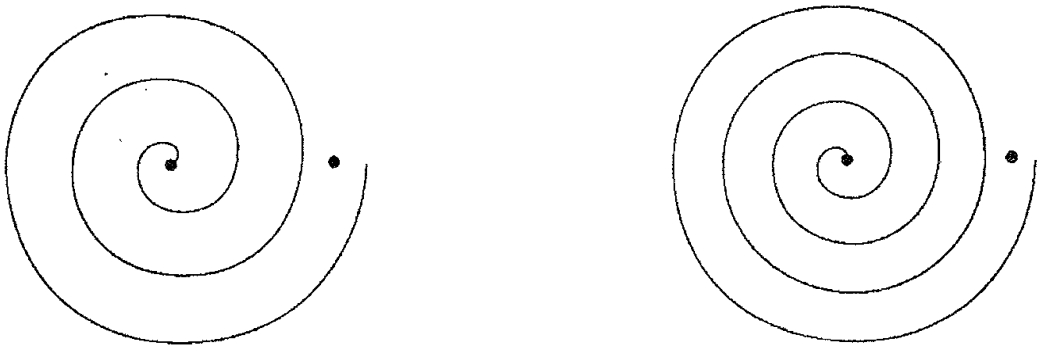
Methods

This chapter highlights the methods carried out in this study. It describes the nature of the utilized spiral drawing data, their acquisition and the techniques used to prepare them for training a CNN. The architecture and hyper-parameters of the utilized CNN are specified and the experiments that were performed are described.

4.1 Data

The available dataset consisted of pen and paper Archimedean spiral drawings with corresponding clinical labels: CRST and UPDRS scores.

4.1.1 Spiral Drawings



(a) (b)

Figure 4.1: The two Archimedean spiral templates provided to patients on a single page: (a) Drawing A template (b) Drawing B.

The tremor data was taken from a previously-acquired database [24, 36]. The ethics clearance certificate for the secondary usage of this data is included in Appendix B. The assessment included two templates that the subjects needed to draw on, as shown in Figure 4.1. The differences between the two spirals was the number of revolutions: 3 and 4 revolutions for Drawing A and Drawing B respectively. A paper sheet with completed templates is shown in Figure 4.2. This figure illustrates the poor quality of scanned pen and paper tests in most clinical settings. In home-based telemedicine settings, the quality of these may be even worse. This problem necessitated extensive image pre-processing which is described in Section 4.2.

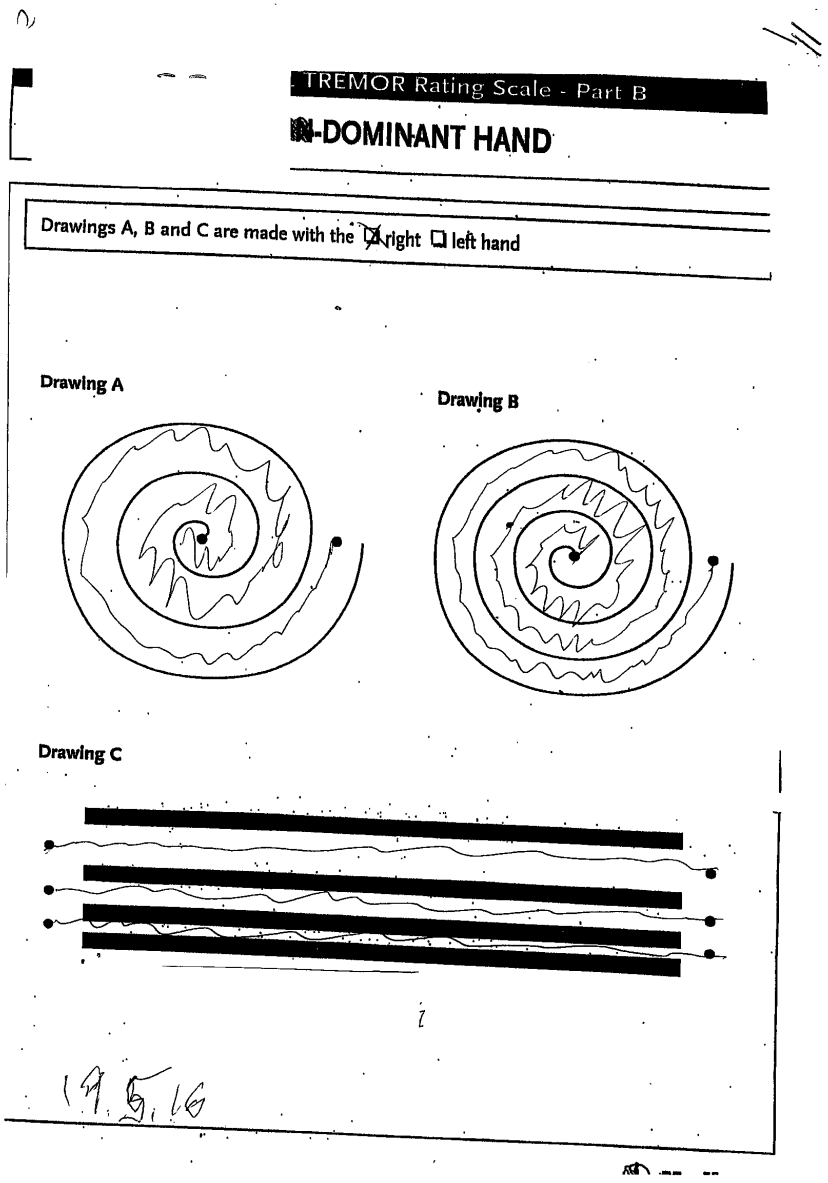


Figure 4.2: An example of a template sheet completed by a patient.

Spiral drawings were collected before and after MRgFUS treatment at the subjects discretion. This resulted in different amounts of data for each of the patients. The dataset had an imbalanced distribution of spiral drawings with different CRST and UPDRS scores.

4.1.2 Tremor Scores

Based on the MRgFUS treatment described in Appendix A, subjects had a choice of which hand (dominant or non-dominant) would receive treatment. Tremor scores were only available for the selected hand and were referred to as the treated hand score. Control subject spirals were evaluated by a medical professional using the CRST.

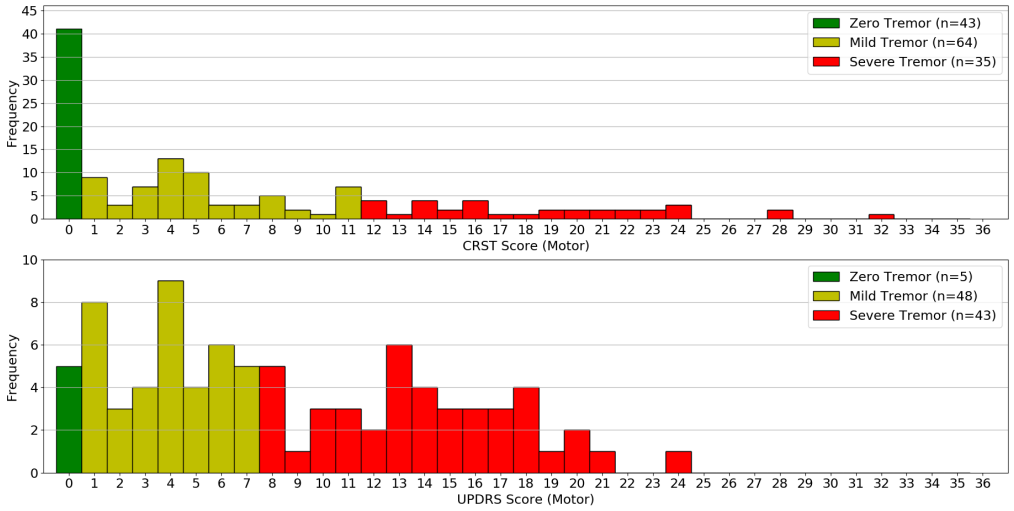


Figure 4.3: The distribution of CRST and UPDRS tremor scores for the spiral dataset.

The tremor score dataset consisted of the following subsets:

1. Treated hand CRST assessment of ET entries
2. Treated hand UPDRS assessment of PD entries
3. CRST assessment of control entries

Figure 4.3 shows the distribution of the CRST and UPDRS tremor scores for the ET and PD treated hand spiral drawings respectively. It illustrates the distribution

of zero, mild and severe tremor scores in the database. Only portions of the UPDRS and CRST were used to calculate the scores of spirals. The maximum possible scores for PD and ET spirals used in this study were 36 and 46 respectively. The threshold ratios in Section 2.3.1 and Section 2.3.2 of Chapter 2 were applied to these values.

The spiral drawings had to undergo certain pre-processing procedures. The automated cropping of spiral drawings and noise removal are discussed in the next section.

4.2 Data Pre-processing

Data pre-processing was divided into two stages: spiral cropping and noise removal.

4.2.1 Spiral Cropping

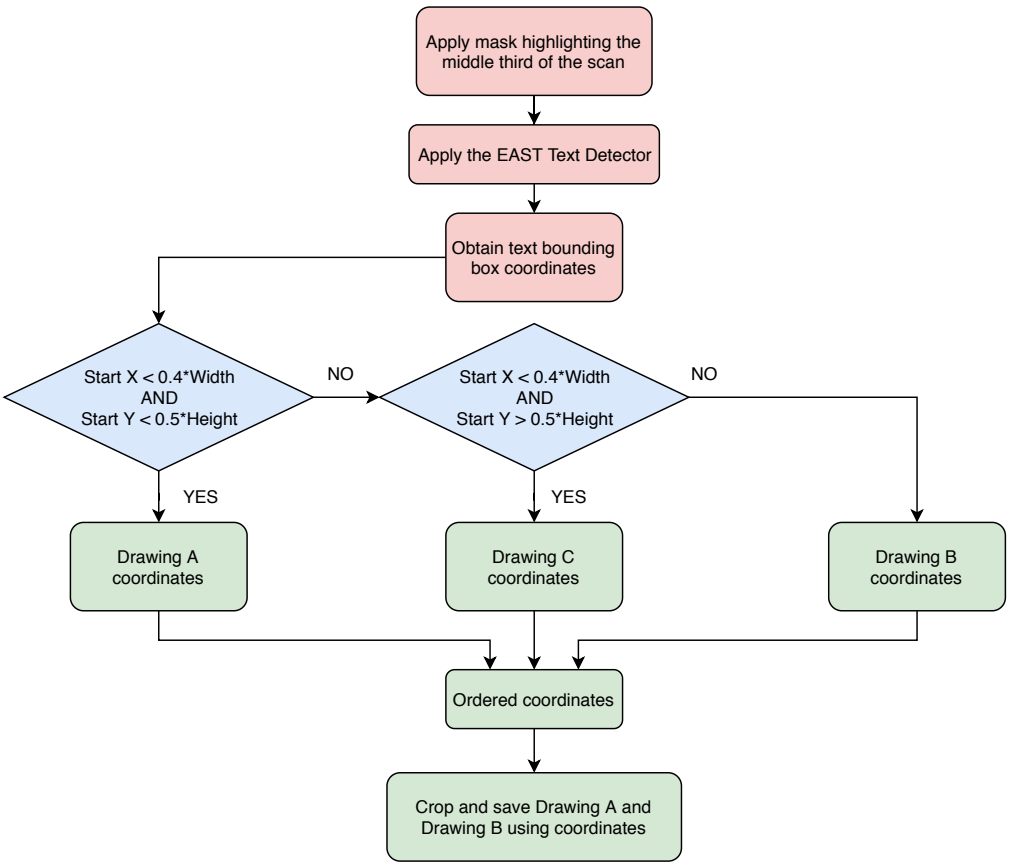
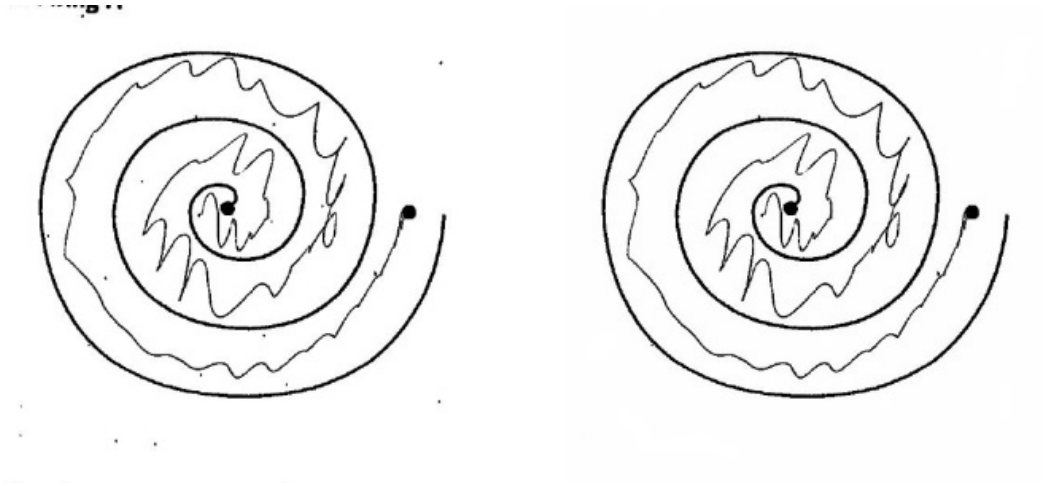


Figure 4.4: Algorithm for spirals cropping from the scanned data page. Start X and Start Y refer to the top left bounding box coordinates of a text scene. End X and End Y refer to the bottom right bounding box coordinates of a text scene.

The subjects' drawings and spiral templates were isolated from the rest of the scan, where Drawing A and Drawing B were separately stored. Figure 4.4 shows the algorithm that was used to crop the completed spirals of Drawing A and Drawing B from a completed template sheet. The Efficient and Accurate Scene Text (EAST) model was utilized for detection of text scenes [80]. Text scene coordinates in the middle third of Figure 4.2 were used to crop Drawing A and Drawing B. Cropped spiral images then underwent noise removal.

4.2.2 Noise Removal

Non-Local Means Denoising (NLMD) was an effective technique for removing random noise in spiral images [81]. The effect of the NLMD is shown by Figure 4.5. The random noise present in Figure 4.5a was removed effectively. The quality of the resulting Figure 4.5b was seen as acceptable to use in CNN experiments.



(a) (b)

Figure 4.5: The effect of Non-Local Means Denoising: (a) Before (b) After. (b) is the final result of pre-processing Drawing A from the template sheet shown in Figure 4.2.

The NLMD technique was applied to all of the cropped spiral images in the dataset. The fully pre-processed spirals were stored in a database structure shown in Figure 4.6. Table 4.1 lists the severity category abbreviations used in Figure 4.6. As will be discussed in Section 4.5, this database was used to perform binary classifications between different severity categories.

Examples of denoised spiral drawings from zero, mild and severe tremor levels from ET and PD patients and from control subjects are provided in the next section.

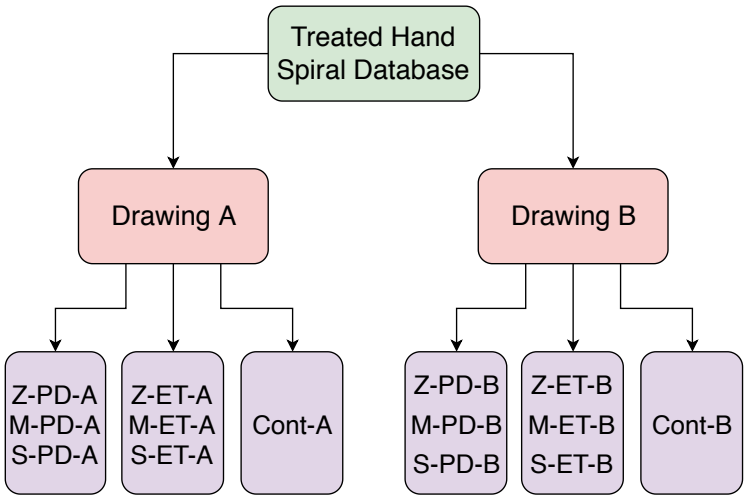


Figure 4.6: Structure of the treated hand cropped spiral database.

Table 4.1: Category abbreviation, sample size and description for the dataset disease severity levels.

Severity Abbreviation	**Sample Size**	**Description**
Z-PD-A/B	5	Zero tremor PD subject spiral Drawing A/B with a Motor-UPDRS score of x=0
M-PD-A/B	48	Mild tremor PD subject spiral Drawing A/B with a Motor-UPDRS score between $0<x\leq 7$
S-PD-A/B	43	Severe tremor PD subject spiral Drawing A/B with a Motor-UPDRS score of $x>7$
Z-ET-A/B	43	Zero tremor ET subject spiral Drawing A/B with a Motor-CRST score of x=0
M-ET-A/B	64	Mild tremor ET subject spiral Drawing A/B with a Motor-CRST score between $0<x\leq 11$
S-ET-A/B	35	Severe tremor ET subject spiral Drawing A/B with a Motor-CRST score of $x>11$
Cont-A/B	47	Spiral Drawing A/B from a subject with no known movement disorders and a Motor-CRST of x=0

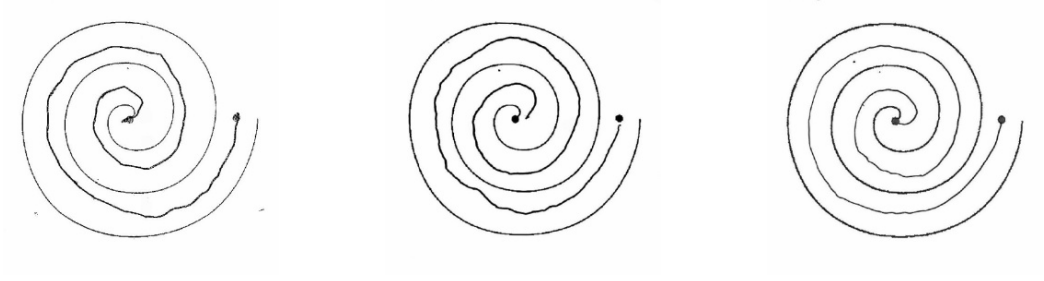
4.3 Spiral Drawing Examples

Figure 4.7 to Figure 4.9 show examples of spiral drawings of each movement disorder in each severity category. As the severity of tremor decreases, spiral drawings look more similar to control spiral drawings. It is a challenge to discriminate between a control subject and a mild movement disorder spiral drawing. It is even more challenging to discriminate between zero tremor disorder and control spiral drawings. The examples also show that it is not trivial to discriminate between each of the movement disorders. By visual inspection, it appeared that the difficulty of this task was constant at all severities.

These difficulties were addressed by using a CNN to perform binary classifications between the relevant classes as mentioned above.

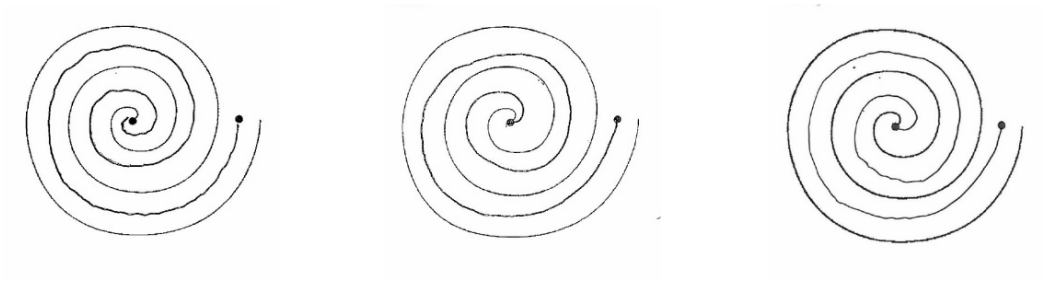
(a) (b) (c)

Figure 4.7: The Drawing A spirals from three different classes of subjects: (a) ET (b) PD and (c) Control. The ET and PD spiral drawings are the severest tremor cases.



(a) (b) (c)

Figure 4.8: The Drawing A spirals from three different classes of subjects: (a) ET (b) PD and (c) Control. The ET and PD spiral drawings are the mildest non-zero cases.



(a) (b) (c)

Figure 4.9: The Drawing A spirals from three different classes of subjects: (a) ET (b) PD and (c) Control. The ET and PD spiral drawings are zero tremor cases.

4.4 Convolutional Neural Network Design

The architectural details of the CNN used for all experiments in this study are described. These details included layer and training hyper-parameters as well as callbacks.

4.4.1 Architecture

As shown in Section 4.3, the data for this study was made up of images that were grey-scale. This is the reason why the input dimensions of the CNN were 64×64×1. As the database of cropped spiral images were saved with dimensions of 300×300, each image was resized to 64×64 prior to being propagated through the CNN. These input image dimensions were chosen as they were used in a study using spiral drawing data [19]. These image dimensions were the most the image could be shrunk without losing too many important features [19]. By reducing the input image size as much as possible, the time required for CNN training was significantly reduced [51]. The CNN architecture that the input images were propagated through is described in Table 4.2.

The CNN was a variant of the LeNet CNN architecture [82]. The LeNet architecture was chosen as a base architecture because it was the best performing architecture for classifying the MNIST handwritten digit dataset [51, 83]. MNIST dataset samples have similar features to the spiral drawing samples. Both datasets are made up of grey-scaled images and are mainly composed of lines. The original LeNet architecture was adapted through several experiments to obtain the architecture

summarised in Table 4.2. The experiments involved recording the performance of binary classifications of severe ET and PD spiral drawings when adding/removing layer types and changing the number of kernels and their dimensions. The original LeNet architecture was adapted by an additional convolutional and max pooling layer. Batch normalization layers were added after every activation layer and a single dropout layer was also introduced. These added layers improved classification performance resulting in an optimal layer type combination. This layer type combination was then optimized by changing each layer's parameters until the classification performance was maximized. It was assumed that the resulting CNN architecture shown in Table 4.2 was also optimized for the rest of the binary classifications that are discussed in Section 4.5.

The CNN architectural specifications were included for reproducing results in future or extended studies. Details of the CNN training hyper-parameters and callbacks are also necessary for reproducing results.

Table 4.2: The CNN Architecture that was used for all experiments. Each layer includes its specified parameters.

Layer #	**Layer Type**	**Number of Kernels**	**Kernel Dimensions**	**Stride**
1	Convolutional	40	5×5	1×1
2	Sigmoid Activation	-	-	-
3	Batch Normalization	-	-	-
4	Max Pooling	-	2×2	2×2
5	Convolutional	100	5×5	1×1
6	Sigmoid Activation	-	-	-
7	Batch Normalization	-	-	-
8	Max Pooling	-	2×2	2×2
9	Convolutional	200	3×3	1×1
10	Sigmoid Activation	-	-	-
11	Batch Normalization	-	-	-
12	Max Pooling	-	2×2	2×2
13	Dropout	-	-	-
14	Fully Connected	4000	-	-
15	Sigmoid Activation	-	-	-
16	Batch Normalization	-	-	-
17	Fully Connected	1	-	-
18	Sigmoid Activation	-	-	-

4.4.2 Hyper-Parameters and callbacks

Specific training parameters and callbacks were determined to ensure that the CNN was trained sufficiently while minimizing over fitting [51]. The hyper-parameters and callbacks utilized in this study have been summarized in Table 4.3. The significance of each of these parameters have been highlighted in Section 2.6 of Chapter 2.

Table 4.3: The hyper-parameter values and callbacks used for training CNN that was used for all experiments.

Hyper-Parameter	**Value or Type**
Learning Rate	10^{-5}
Loss Function	Binary Cross-entropy
Optimization Algorithm	Stochastic Gradient Descent
Epochs	4000
Decay rate	2.5×10^{-9}
Momentum	0.9
Nesterov Acceleration	True
Class Weighting	True
Cross Validation	Stratified 5-fold
Cross Validation Random State	Fixed
Callback	**Value or Type**
Early stopping monitoring	Minimum Validation Binary Cross-entropy
Early stopping patience	30 epochs
Model Checkpoint	Maximum Validation Accuracy

The learning-rate, training epochs and early stopping training epochs were optimised using the same method and datasets used to optimize the CNN architecture. These values had to be determined experimentally since no existing studies have investigated the loss landscape when optimising hyper-parameters in this domain. As mentioned in Section 2.6 of Chapter 2, the values selected for momentum and decay rate were based on the current best practices in the field of CNN-based image classification [51, 52]. The binary cross-entropy loss function was selected since this is the current best practice in binary image classification specifically [51, 52].

The hyper-parameters and callbacks were specified in order for training to be optimized for all binary classifications performed in this study. The next section details the experimental procedures that were carried out to address the research questions specified in Section 3.2.

4.5 Experiments

This section described the experimental procedures that were carried out to address the research questions. The experiments consisted of binary classifications using different tremor severity and disorder datasets. Abbreviations and descriptions of these datasets are shown in Table 4.4.

Table 4.4: Category abbreviation, sample size and description for the dataset disease severity levels.

Severity Abbreviation	**Sample Size**	**Description**
Z-PD-A/B	5	Zero tremor PD subject spiral Drawing A/B with a Motor-UPDRS score of x=0
M-PD-A/B	48	Mild tremor PD subject spiral Drawing A/B with a Motor-UPDRS score between $0<x\leq7$
S-PD-A/B	43	Severe tremor PD subject spiral Drawing A/B with a Motor-UPDRS score of $x>7$
Z-ET-A/B	43	Zero tremor ET subject spiral Drawing A/B with a Motor-CRST score of x=0
M-ET-A/B	64	Mild tremor ET subject spiral Drawing A/B with a Motor-CRST score between $0<x\leq11$
S-ET-A/B	35	Severe tremor ET subject spiral Drawing A/B with a Motor-CRST score of $x>11$
Cont-A/B	47	Spiral Drawing A/B from a subject with no known movement disorders and a Motor-CRST of x=0

Table 4.5 summarizes the eight experiments performed using the datasets listed in Table 4.4. The zero tremor PD dataset (Z-PD-A/B) was not used in any binary classification experiments due to its small size. The Python code used to train and validate the CNN is available at `https://github.com/kdasilva835842/tremor_classification`.

All the experiments were performed to determine:

1. How effective is the CNN at discriminating two different severities of the same movement disorder: either ET or PD.

2. How effective is the CNN at discriminating two movement disorders: ET and PD with the same severity level.

3. How effective is the CNN at discriminating controls from movement disorder subjects with mild tremor.

4. What are the differences in binary classification performance of each dataset, between Drawing A or Drawing B.

5. What are the effects of data augmentation on the performance of the CNN in binary classification.

Table 4.5: Summary of experiments that were performed.

	Description & Purpose	**Dataset**	**Performance Metrics**
1	Test CNN binary classification of mild and severe ET tremor entries	M-ET S-ET	Mean and standard deviation of Accuracy, Precision, Recall and F1 score
2	Test CNN binary classification of zero tremor rated ET and control tremor entries	Z-ET Cont	Mean and standard deviation of Accuracy, Precision, Recall and F1 score
3	Test CNN binary classification of mild and severe PD entries	M-PD S-PD	Mean and standard deviation of Accuracy, Precision, Recall and F1 score
4	Test CNN binary classification of control and mild PD entries	M-PD Cont	Mean and standard deviation of Accuracy, Precision, Recall and F1 score
5	Test CNN binary classification of severe ET and severe PD entries	S-ET S-PD	Mean and standard deviation of Accuracy, Precision, Recall and F1 score
6	Test CNN binary classification of mild ET and mild PD entries	M-ET M-PD	Mean and standard deviation of Accuracy, Precision, Recall and F1 score
7	Test CNN binary classification of mild ET and zero tremor ET	M-ET Z-ET	Mean and standard deviation of Accuracy, Precision, Recall and F1 score
8	Test the effect of data augmentation on CNN binary classification of mild ET and zero tremor ET	M-ET-A Z-ET-A	Mean and standard deviation of Accuracy, Precision, Recall and F1 score

4.5.1 Binary Classification

All the CNN binary classification experiments listed in Table 4.5 were performed through a stratified 5-fold cross validation. The relevant performance metrics for each fold were recorded. Accuracy was predominantly used to compare classification performance with previous studies. Precision, Recall and F1-Score were used to evaluate CNN classifications while accounting for class imbalances. The mean and standard deviation of each stratified 5-fold cross validation was also calculated. Figure 4.10 illustrates the procedure for performing the CNN binary classification experiments 1-7.

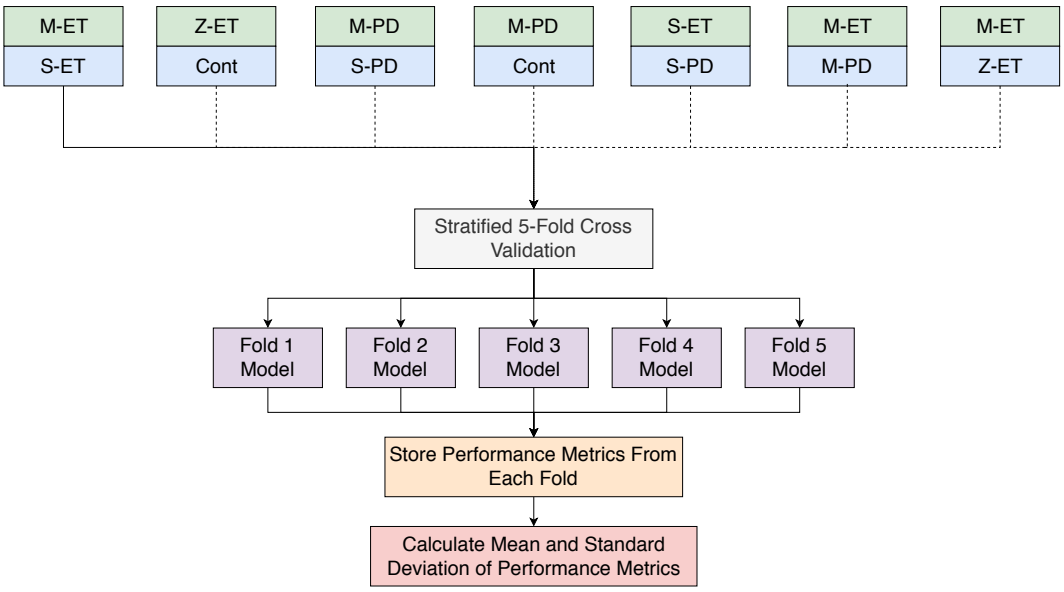


Figure 4.10: The experimental procedure for testing all of the CNN binary classifications. The relevant datasets that were used for experiments 1-7 are shown.

Experiments 1-7 in Table 4.5 were performed using the same CNN architecture, hyperparameters and callbacks as discussed in Section 4.4. Each of these experiments involved evaluating the binary classification performance using Drawing A and Drawing B spirals. The obtained results indicated which spiral drawings were more suitable for performing the binary classifications. The results also indicated how effective the CNN was in performing a binary classification using the different spiral drawing datasets.

The procedure of Experiment 8, which was used to evaluate the effect of data augmentation, differed slightly to the rest.

4.5.2 Effects of Data Augmentation

Experiment 8 was conducted to observe the effects of different degrees of affine transformation data augmentation. Random rotation and translation data augmentation techniques were evaluated. Experiment 8 involved the binary classification between augmented Drawing A spiral data of mild and zero tremor ET patients only. These drawings were used as they resulted in average binary classification performance compared to the others described in Section 4.5.1. The effects of data augmentation were assumed to be the same for all other binary classifications. The procedure of Experiment 8 is illustrated in Figure 4.11. This process was repeated with different types and degrees of spiral drawing data augmentation. Table **??** details the different degrees of data augmentation that were applied with the repetition of the procedure in Figure 4.11.

The results of Experiments 1-8 are summarised in Chapter 5.

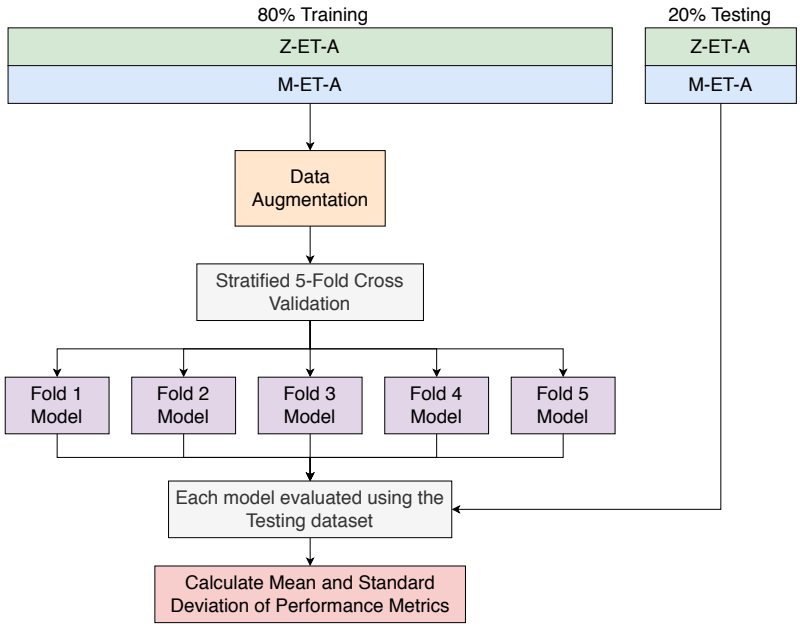


Figure 4.11: The procedure for obtaining results for one type and degree of data augmentation. It was repeated for all types and degrees of data augmentation. It shows the initial training and testing dataset splits of 80% and 20% respectively. The procedure was repeated with a 70% and 30% training and testing dataset split.

Table 4.6: Degrees of affine transformation data augmentation. 5 new spiral images were created for each case to make classes larger than in previous studies [18].

Augmentation Type	**Max degrees of rotation (°)**	**Max degrees of shift (%)**	**Description**
Rotation	0	0	The 5 generated images were rotated randomly in either clockwise or anti-clockwise directions by no more than the maximum rotation value.
	2	0	
	5	0	
	10	0	
	15	0	
	20	0	
	25	0	
	30	0	
Translation	0	0	The 5 generated images were randomly translated vertically and horizontally by no more than the maximum translation value. The maximum translation value was a percentage of image height and width (300x300).
	0	2	
	0	4	
	0	6	
	0	8	
	0	10	

Chapter 5

Results

Table 5.1 contains the results of Experiments 1-7, as detailed in Section 4.5. The results in Table 5.1 were sorted according to descending F1-Score. Since certain experiments involved imbalanced datasets, ordering results according to accuracy would not be appropriate. F1-Score is more invariant to data imbalances and is calculated using Precision and Recall. The results of Experiment 8 were presented in Figure 5.1 and Figure 5.2. Different degrees of data augmentation using different training and testing dataset sizes were presented. Rotation and translation techniques were compared separately.

Table 5.1: Summary of the results acquired from experiments 1-7.

	Description & Purpose	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
2	Test CNN binary classification of zero tremor rated ET and control tremor entries	Z-ET-A Cont-A	96.4 ±2.9	97.8 ±4.4	95.0 ±6.1	96.2 ±3.2
		Z-ET-B Cont-B	98.8 ±2.4	97.8 ±4.4	100.0 ±0.0	98.8 ±2.4
4	Test CNN binary classification of control and mild PD entries	M-PD-A Cont-A	85.1 ±7.7	89.1 ±9.5	81.4 ±12.5	84.3 ±8.3
		M-PD-B Cont-B	92.2 ±8.3	94.9 ±6.3	88.9 ±12.2	91.6 ±9.3
5	Test CNN binary classification of severe ET and severe PD entries	S-PD-A S-ET-A	82.4 ±16.7	94.0 ±8.0	72.8 ±30.1	78.1 ±23.4
		S-PD-B S-ET-B	71.2 ±9.0	79.6 ±7.8	64.7 ±16.6	70.1 ±12.8
6	Test CNN binary classification of mild ET and mild PD entries	M-PD-A M-ET-A	81.8 ±2.6	70.9 ±10.0	72.2 ±13.1	69.9 ±4.6
		M-PD-B M-ET-B	80.9 ±7.7	72.1 ±13.3	66.4 ±14.6	68.4 ±11.8
7	Test CNN binary classification of mild ET and zero tremor ET	M-ET-A Z-ET-A	72.4 ±10.0	62.4 ±10.2	73.9 ±19.3	67.1 ±13.0
		M-ET-A Z-ET-A	56.4 ±10.4	48.3 ±10.0	64.2 ±10.8	54.8 ±9.3
3	Test CNN binary classification of mild and severe PD entries	M-PD-A S-PD-A	70.2 ±11.8	72.0 ±12.1	65.3 ±20.9	67.0 ±13.9
		M-PD-B S-PD-B	68.7 ±9.3	71.2 ±16.3	65.6 ±22.5	64.7 ±13.3
1	Test CNN binary classification of mild and severe ET tremor entries	M-ET-A S-ET-A	75.9 ±5.2	51.1 ±10.8	69.1 ±17.2	57.6 ±9.3
		M-ET-B S-ET-B	67.3 ±9.7	42.4 ±10.9	58.6 ±7.0	48.1 ±6.9

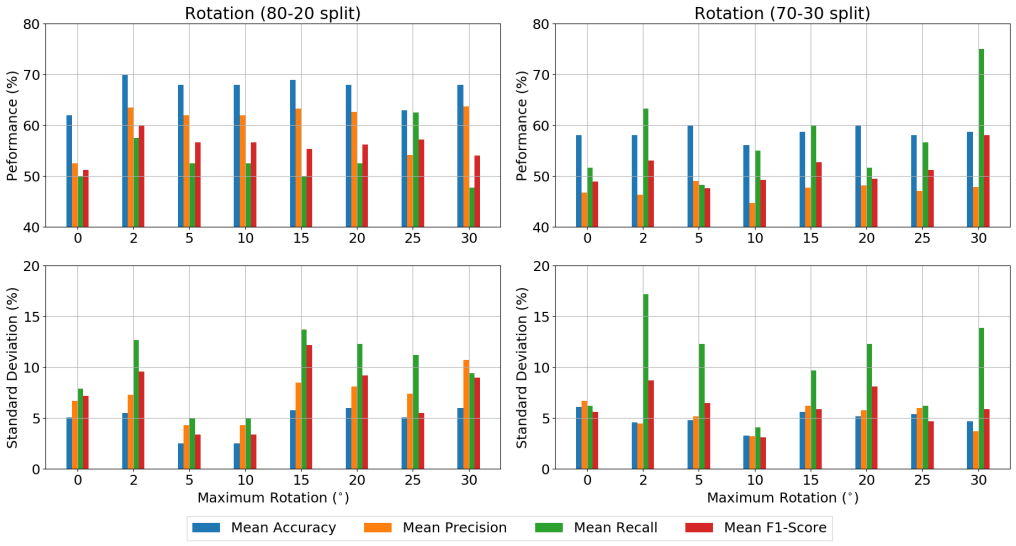


Figure 5.1: The data augmentation results (Experiment 8) showing the effect of different degrees of rotation on binary classification performance. These results were acquired using the method highlighted in Section 4.5.2. A comparison of the rotation data augmentation techniques using 80%-20% and 70%-30% train-test datasets is presented.

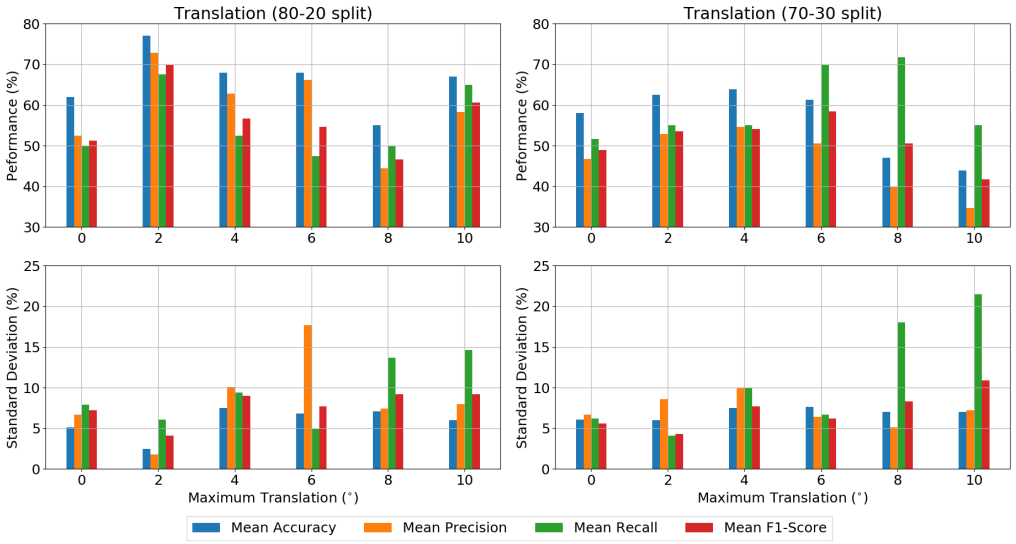


Figure 5.2: The data augmentation results (Experiment 8) showing the effect of different degrees of translation on binary classification performance. These results were acquired using the method highlighted in Section 4.5.2. A comparison of the translation data augmentation techniques using 80%-20% and 70%-30% train-test datasets is presented.

Chapter 6

Discussion

The results described in Chapter 5 reveal insights which pertain to ET and PD movement disorders, to the Archimedean spiral acquisition and to the methods used to analyse them: CNNs and data augmentation. These insights were discussed in the following sections.

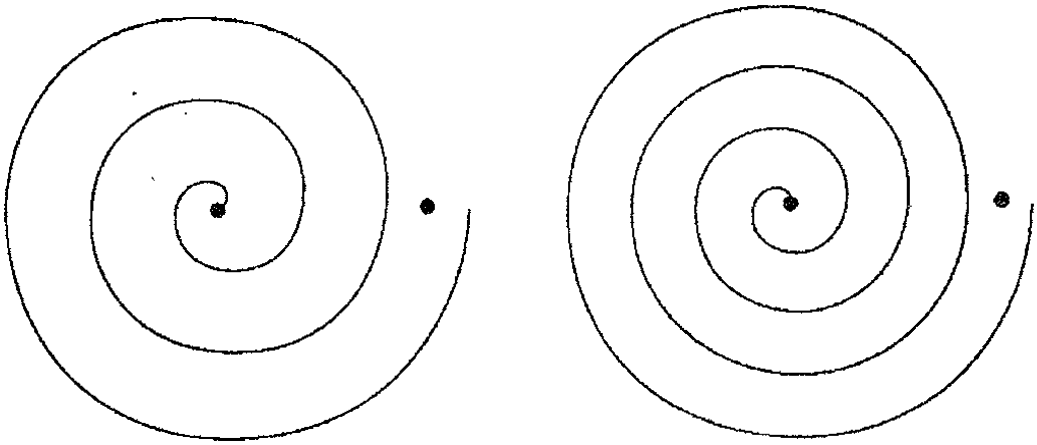
6.1 Template Effect on Classification Performance

Experiments 1-7 were performed on two types of drawings: A and B. All subjects performed both drawings such that these two classes (A and B) had the same size. This implied that differences in classification performance could not be assigned to data imbalance and stem from the differences between the two types of spiral template. The use of Drawing A datasets resulted in improved classification performance in all experiments except for that of Experiment 2 and Experiment 4.

As shown in Figure 6.1, Drawing A has one fewer revolution compared to Drawing B. Movement disorder subjects tire much quicker than controls. The severity of the tremor intensifies as the subject becomes more tired. This implies that use of Drawing B is more likely to tire the subjects compared to use of Drawing A, and fatigue effects may resemble higher tremor severity. This is corroborated in the results of experiments classifying between tremor severities of the same movement disorder (Experiments 1,3,7) and the discrimination between ET and PD spiral drawings (Experiments 5 and 6).

Drawing B offered improved classification performance in Experiment 2 and Experiment 4 where Control subject drawings were discriminated. This discrimination may intensify differences between able controls and severe movement disorders when

Drawing B is used. In all the other experiments (5 of the 7 experiments), Drawing A spirals yielded improved binary classification performance compared to Drawin B. This finding implies that acquiring Drawing B spirals in addition to Drawing A spirals may be redundant for the movement disorder patients' assessment. Future studies making use of pen and paper spirals may consider excluding Drawing B.



(a) (b)

Figure 6.1: The two Archimedean spiral templates provided to patients on the assessment page: (a) Drawing A template (b) Drawing B template.

6.2 Tremor Severity Discrimination

The performance of the binary classification between mild and severe cases in ET and PD (Experiments 1 and 3) are discussed below. Figure 6.2 illustrates the distribution of the CRST and UPDRS tremor scores in the dataset and was used to aid the discussions of tremor severity classifications in ET and PD.

The CNN was able to discriminate between zero tremor ET and control subjects (Experiment 2) with an accuracy of 98.8%. Since both of these classes consisted of spiral drawings with CRST ratings of zero, the only difference between them was the presence of a movement disorder or not. This implies that the CNN may be used to detect whether person has ET even before the symptoms of tremor are present. This may prove to be useful as an early warning system for movement disorders.

The small size of the zero tremor PD patients data and the large imbalance of this class and the controls class could not enable reliable CNN performance. The analysis of zero tremor in PD will need to be addressed in future studies with a larger dataset.

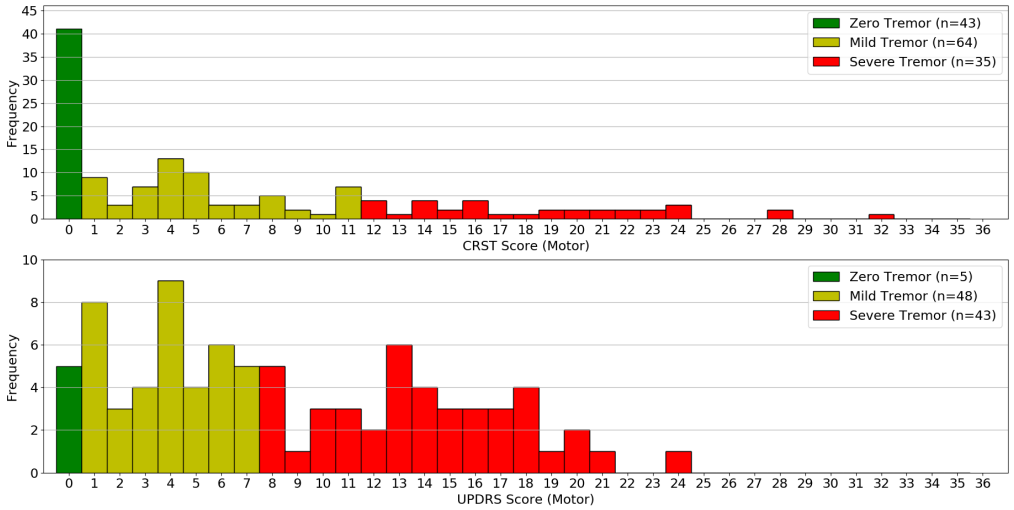


Figure 6.2: The distribution of CRST and UPDRS tremor scores for the spiral dataset. The number of control subjects was n = 47.

6.2.1 ET Severity Discrimination

No previous studies performed automated discrimination of severity classes in ET. These results will serve as a benchmark for future studies aiming to determine ET tremor severity using CNNs.

Experiment 7 (discriminating mild from zero tremor cases) outperformed Experiment 1 (discriminating mild from severe tremor cases). The results may be attributed to a larger class imbalance between mild and severe ET tremor compared to zero-rated and mild ET tremor. There was a non-uniform distribution of data within the severe tremor ET class. The performance of Experiment 1 may improve if more spiral drawings with higher tremor severities within the severe tremor ET class are made available in future studies.

6.2.2 PD Severity Discrimination

No previous studies discriminated mild from severe PD tremor cases. The results of Experiment 3 could serve as a benchmark for future studies on PD tremor severity. The CNN performance was better at discriminating between mild and severe tremor when using PD data compared to when using ET data. This implies that differences between the mild and severe tremor classes may be more evident in PD spirals drawings compared to ET spiral drawings.

The discrimination between mild and severe cases of ET yielded similar performance to that of PD. The classification performances of Experiments 1, 2, 3, 4 and 7 were all greater than 70%. This suggests that the CNN could discriminate between two different tremor severities more effectively than a trained physician [42]. This also indicated that the CNN was suitable for this discrimination in both movement disorders where tremor was labelled using different tremor rating scales (CRST and UPDRS).

6.3 Type of Movement Disorder Discrimination

Experiments 5 and 6 involved discriminating between ET and PD using mild and severe tremor severity cases. The results showed a superior classification performance between the two movement disorders when severe tremor cases were used, compared to mild tremor cases. The thresholds for severity classification are comparable in both disorders and comparing the severities of each disorder was valid. However, it is possible that the underlying subjectivity in physician ratings could have added uncertainty to the severity labels. This was particularly the case for drawings with scores close to the thresholds. At these scores, there was a possibility that the drawing should have a different severity classification. This could have affected the results of Experiment 5 and 6 as the data labels would be different. Determining whether or not classification performances would improve could not be certain and requires further investigation.

It was a relatively simple task to discriminate between controls and any movement disorder with very severe tremor. A greater challenge was to discriminate between controls and movement disorders with mild tremor because samples from each class look very similar to one another. The results indicated that the CNN was able to perform these challenging discriminations (zero tremor ET from control cases in Experiment 2 and mild tremor PD from control cases in Experiment 4) with high accuracy. Since previous studies performing the same binary classification as Experiment 4 also used accuracy as a performance metric, a valid comparison could be made. The performance in Experiment 4 (92.2% accuracy) was much improved compared to previous studies that used CNNs to discriminate PD from control spiral drawings. The best accuracy obtained in the previous studies was 90.38% [18].

No previous studies using CNNs to discriminate between ET and control spiral drawings have been performed.

6.4 The Effects of Data Augmentation

Based on the results shown in Figure 5.1 and Figure 5.2, translation was the superior choice of data augmentation compared to rotation. The best classification performance in terms of all metrics was achieved when using translation as an augmentation technique. The optimal degree of data augmentation was found to be random translations of no more than 2% of image height and width when using the 80% and 20% training and testing data. The optimal degree of data augmentation was found to be random translations of no more than 4% of image height and width when using the 70% and 30% training and testing data split. It is possible that translation augmentation led to improved classification performance because it increased the receptive field of the CNN more than rotation. The CNN was possibly more likely to distinguish factors between different types and degrees of tremor when the data was translated rather than rotated.

The changes in performance metrics as the degrees of translation increased were parabolic in nature. There was a definite peak in performance. This parabolic nature was more clearly seen when using the 70% and 30% training and testing dataset. This behaviour could also be seen using the 80% and 20% dataset split but there appeared to be an outlier in performance when performing translations of up to 10% of image height and width. With exception of this outlier, a consistent trend in classification performance using different degrees of translation was seen.

The effects of augmentation were analysed using spiral data of mild and zero tremor ET patients. Therefore, binary classification performance in Experiment 7 and Experiment 8 could be compared. Figure 5.2 showed that an average accuracy of 77% was achieved using augmented spiral data. This was 4.6% higher than the 70.2% average accuracy obtained using spiral data that was not augmented in Experiment 7. The standard deviation of accuracy using augmented data decreased by 9.3% (from 11.8% to 2.5%).

The effects of augmentation were analysed using spiral data of mild and zero tremor ET patients only. These drawings were used as they resulted in average binary classification performance compared to the others mentioned in Section 6.2 and Section 6.3. The effects of data augmentation were assumed to be the similar for all the other spiral drawings and binary classifications. There have been no prior studies indicating the effect of different augmentation techniques on this type of data, making these results useful for studies involving small spiral drawing datasets.

6.5 Overall Findings

The insights made in this study contributed to the field of digital health. Current tremor rating scales are subjective. The assessment determining these scales are labour intensive and unsuitable for telemedicine. Scanned pen and paper spiral drawings can be used for monitoring of movement disorders from home. The pre-processing employed in this study was able to amend even badly scanned spirals for tremor severity discrimination. The classification performance implied that one of the two spiral templates traditionally used in clinical tests for movement disorders is redundant and that only one should be used, which can further facilitate home-use and decrease clinical assessment time. The CNN was able to discriminate between mild and severe tremor and between different movement disorders from spiral drawings, which may serve as a benchmark for future studies.

The study is the first to convey results of discriminations between ET and control subjects from spirals drawings. It also demonstrated accurate discriminations between zero tremor ET and control subjects, highlighting the possibility of being used as an early detection mechanism of movement disorders. The performance of the discriminations between PD and control spirals were similar or higher than previous studies. There were no correlations between sample sizes and classification performances. The best type and degree of data augmentation for the spirals was determined. Data augmentation demonstrated the ability to improve classification performance and reduce variability. All these findings suggest that a CNN can be used to automatically discriminate between tremor severities using pen and paper spiral drawings.

These findings will serve as building blocks for developing a home tremor severity monitoring system. Scanned template pages with spiral drawings allow accessibility and ease of use. This study has contributed the fundamentals for such a system. Recommendations for future researchers are provided in Chapter 7.

Chapter 7

Conclusions

7.1 Future Recommendations

This study provided methods for automated tremor analysis based on hand-drawn spirals on paper. These insights formed a foundation for further development. The results and insights drawn can be extended with the following recommendations.

7.1.1 Transfer Learning

As mentioned in Section 2.7.2 of Chapter 2, transfer learning has shown successes in addressing optimization difficulties in CNN training. Research groups aiming to extend this study should investigate the effect of a deep transfer learning on this study's results. It will be useful to determine which pre-existing image dataset would be useful to train a CNN for this purpose.

7.1.2 Spiral Segmentation

The preprocessing procedure that took place on the cropped spiral drawings was noise removal. The spiral template that patients used to trace their drawings was not removed. It would be useful to determine how spiral segmentation effects the performance of CNN predictions. This investigation may require the next research group to develop a tool to perform this image segmentation.

7.1.3 CNN Regression

Binary discrimination between mild and severe cases is something that should be developed further in terms of severity resolution. The efficacy of using CNN to predict CRST and/or UPDRS scores from spiral drawings should be investigated. The obstacle preventing this investigation from being performed in this study was the relatively small available dataset. Depending on the results of the recommended transfer learning investigation, this challenge may be mitigated. Additional spiral drawing data may also be collected.

7.2 Study Conclusion

PD and ET are two movement disorders that are characterized by tremor. While being standardised procedures, the clinical scales for tremor severity today are subjective, labour intensive, and may be biased. The existing approaches that quantify tremor severity require the usage of digital tablets which are not appropriate in many clinical settings. There was a need to automatically assess the severity of tremor of ET and PD patients. This study determined how effectively a CNN could automatically discriminate between different tremor severities using pen and paper spirals. A second challenge was to determine an efficient method to overcome the problem of small dataset size, a common issue in medical data acquisition. This was done by restricting the classification to binary classifications of pairs of classes from the zero, mild and severe tremor levels, from ET, PD and controls. The results showed accuracy between 70.2% and 98.8%. An interesting result was that patients who had zero tremor were still discriminated from controls. Another solution to the small data size was the data augmentation. The results indicated an optimal type and degree of data augmentation that could improve the performance in all datasets of spiral drawings. The findings of this study are preliminary, and may serve as building blocks for the development of a home tremor monitoring system and early movement disorder detection mechanism.

Bibliography

[1] K.-Y. Kwon et al. "Comparison of motor and non-motor features between essential tremor and tremor dominant Parkinson's disease." *Journal of the Neurological Sciences*, vol. 361, pp. 34–38, 2016.

[2] C. G. Goetz and W. Poewe. "The Unified Parkinson's Disease Rating Scale (UPDRS): Status and Recommendations." *Movement Disorders*, vol. 18, pp. 738–750, 2003.

[3] R. Elble et al. "Reliability of a new scale for essential tremor." *Movement Disorders*, vol. 22, pp. 1567–1569, 2014.

[4] M. Algarni and A. Fasano. "The overlap between Essential tremor and Parkinson disease." *Parkinsonism and Related Disorders*, vol. 46, pp. 101–104, 2018.

[5] J. Alty, J. Cosgrove, D. Thorpe, and P. Kempster. "How to use pen and paper tasks to aid tremor diagnosis in the clinic." *Practical Neurology*, vol. 17, no. 6, pp. 456–463, 2017.

[6] R. Saunders-Pullman, C. Derby, K. Stanley, A. Floyd, S. Bressman, R. B. Lipton, A. Deligtisch, L. Severt, Q. Yu, M. Kurtis, and S. L. Pullman. "Validity of spiral analysis in early Parkinson's disease." *Movement Disorders*, vol. 23, no. 4, pp. 531–537, 2008.

[7] P. Drotar, J. Mekyska, I. Rektorova, L. Masarova, Z. Smekal, and M. Faundez-Zanuy. "Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease." *Artificial Intelligence in Medicine*, vol. 67, pp. 39 – 46, 2016.

[8] M. C. Barraso, G. P. Esteves, T. P. Nunes, L. M. G. Silva, A. C. D. Faria, and P. L. Melo. "A telemedicine instrument for remote evaluation of tremor: design and initial applications in fatigue and patients with Parkinson's Disease." *BioMedical Engineering OnLine*, vol. 10, pp. 1–17, 2011.

[9] C. R. Pereira et al. "A Step Towards the Automated Diagnosis of Parkinson's Disease: Analyzing Handwriting Movements." *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, 2015.

[10] S. L. Pullman. "Spiral Analysis: A New Technique for Measuring Tremor With a Digitizing Tablet." *Movement Disorders*, vol. 13, pp. 85–89, 1998.

[11] P. Zham et al. "Distinguishing Different Different Stages of Parkinson's Disease Using Composite Index of Speed and Pen-Pressure of Sketching a Spiral." *Frontiers in Neurology*, vol. 8, pp. 1–7, 2017.

[12] M. Wang, Bei, et al. "Quantitative Evaluation of Hand Movement in Spiral drawing for Patients with Parkinson's Disease Based on Modeling in Polar Coordinate System with Varied Origin." *International Conference on Complex Medical Engineering*, 2011.

[13] J. Mucha et al. "Identification and Monitoring of Parkinson's Disease Dysgraphia Based on Fractional-Order Derivatives of Online Handwriting." *Applied Sciences*, vol. 8, 2018.

[14] 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.

[15] C. R. Pereira, S. A. T. Weber, C. Hook, G. H. Rosa, and J. P. Papa. "Deep Learning-Aided Parkinson's Disease Diagnosis from Handwritten Dynamics." In *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 340–346. Sao Paulo, Brazil: IEEE, 2016.

[16] C. R. Pereira et al. "Handwritten dynamics assessment through convolutional neuralnetworks: An application to Parkinson's disease identification." *Artificial Intelligence in Medicine*, vol. 87, pp. 67–77, 2018.

[17] L. Bote-Curiel et al. "Deep Learning and Big Data in Healthcare: A Double Review for Critical Beginners." *Applied Science*, vol. 9, 2019.

[18] C. R. Pereira et al. "Convolutional Neural Networks Applied for Parkinson's Disease Identification." *Machine Learning for Health Informatics*, pp. 377–390, 2017.

[19] M. Moetesum, I. Siddiqi, N. Vincent, and F. Cloppet. "Assessing visual attributes of handwriting for prediction of neurological disorders - A case study on Parkinson's disease." *Pattern Recognition Letters*, vol. 121, pp. 19–27, 2019.

[20] A. Namazov and Y. I. Cho. "An Improvement for Medical Image Analysis Using Data Enhancement Techniques in Deep Learning." *Ministry of Science and ICT, Korea*, pp. 19–27, 2018.

[21] T. Hayashi, K. Gyohten, H. Ohki, and T. Takami. "A Study of Data Augmentation for Handwritten Character Recognition Using Deep Learning." In *2018 16th International Conference on Frontiers in Handwriting Recognition*, pp. 552–557. IEEE, 2018.

[22] A. Anouti and W. C. Koller. "Tremor Disorders - Diagnosis and Management." *West J Med*, vol. 162, pp. 510–513, 1995.

[23] G. D. Maagd and A. Philip. "Parkinson's Disease and Its Management, Part 1 - Disease Entity, Risk Factors, Pathophysiology, Clinical Presentation, and Diagnosis." *Pharmacy and Therapeutics*, vol. 40, p. 504, 2015.

[24] I. Schlesinger, A. Sinai, and M. Zaaroor. "MRI-Guided Focused Ultrasound in Parkinson's Disease: A Review." *Journal of Parkinson's Disease*, 2017.

[25] R. J. Elble. "What is essential tremor?" *Curr Neurol Neurosci Rep*, vol. 13, p. 353, 2013.

[26] L. N. Clark and E. D. Louis. "Essential tremor." *Handb Clin Neurol*, vol. 147, pp. 229–239, 2018.

[27] W. Oertel and J. B. Schulz. "Current and experimental treatments of Parkinson's disease: A guide for neuroscientists." *Journal of Neurochemistry*, vol. 139, pp. 325–337, 2016.

[28] J. Jankovic and L. G. Aguilar. "Current approaches to the treatment of Parkinson's disease." *Neuropsychiatric Disease and Treatment*, vol. 4, pp. 743–757, 2008.

[29] M. S. Okun. "Deep-Brain Stimulation for Parkinson's Disease." *The New England Journal of Medicine*, pp. 1529–1538, 2012.

[30] A. H. Rajput and A. Rajput. "Medical Treatment of Essential Tremor." *Journal of Central Nervous System Disease*, vol. 6, pp. 29–39, 2014.

[31] D. Salat and E. Tolosa. "Levodopa in the Treatment of Parkinson's Disease: Current Status and New Developments." *Journal of Parkinson's Disease*, vol. 3, pp. 255–269, 2013.

[32] A. Fytagoridis and P. Blomstedt. "Complications and Side Effects of Deep Brain Stimulation in the Posterior Subthalamic Area." *Stereotactic and Functional Neurosurgery*, pp. 88–93, 2010.

[33] I. Schlesinger et al. "MRI Guided Focused Ultrasound Thalamotomy for Moderate-to-Severe Tremor in Parkinson's Disease." *Journal of Parkinson's disease*, 2015.

[34] M. Zaaroor et al. "MR guided focused ultrasound as a treatment tool for essential and Parkinsonian tremor." *Journal of Therapeutic Ultrasound*, vol. 3, 2015.

[35] V. Rieke et al. "MR Thermometry." *Journal of Magnetic Resonance Imaging.*, vol. 27, pp. 376–390, 2008.

[36] I. Schlesinger, Department of Neurology at Rambam Healthcare Campus, 2019.

[37] M. A. Thenganatt and E. D. Louis. "Distinguishing essential tremor from Parkinson's disease: bedside tests and laboratory evaluations." *Expert Rev Neurother*, vol. 12, pp. 687–696, 2012.

[38] M. Vallejo et al. "Exploring Diagnostic Models of Parkinson's Disease with Multi-Objective Regression." *IEEE*, pp. 1–8, 2016.

[39] S. Grover et al. "Predicting Severity Of Parkinson's Disease Using Deep Learning." *Procedia Computer Science*, vol. 132, pp. 1788–1794, 2018.

[40] S. Fahn, E. Tolosa, and C. Marin. "Clinical Rating Scale for Tremor." *Parkinson's Disease and Movement Disorders*, pp. 225–234, 1988.

[41] N. Mohammed. "A meta-analysis of outcomes and complications of magnetic resonance-guided focused ultrasound in the treatment of essential tremor." *Neurosurgical Focus*, vol. 44, pp. 1–9, 2018.

[42] P. Y. Chan, Z. M. Ripin, et al. "Biomedical System Versus Observational Rating Scale for Parkinson's Disease Tremor Assessment." *Nature*, vol. 9, pp. 1–11, 2019.

[43] P. G. Bain et al. "Assessing Tremor Severity." *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 56, pp. 868–873, 1993.

[44] F. Tam, Y. Huang, M. L. Schwartz, T. A. Schweizer, K. Hynynen, and S. J. Graham. "A computerized tablet system for evaluating treatment of essential tremor by magnetic resonance guided focused ultrasound." *BMC Neurology*, vol. 17, no. 1, 2017.

[45] A. Sadikov, J. Zabkar, M. Mozina, V. Groznik, D. Georgiev, and I. Bratko. "ParkinsonCheck A Decision Support System for Tremor Detection." 2015.

[46] M. Memedi, A. Sadikov, V. Groznik, J. Žabkar, M. Možina, F. Bergquist, A. Johansson, D. Haubenberger, and D. Nyholm. "Automatic Spiral Analysis for Objective Assessment of Motor Symptoms in Parkinson's Disease." *Sensors*, vol. 15, no. 9, pp. 23727–23744, 2015.

[47] K. Lopez de Ipina, M. Iturrate, P. Calvo, B. Beitia, J. Garcia-Melero, A. Bergareche, P. De la Riva, J. Marti-Masso, M. Faundez-Zanuy, E. Sesa-Nogueras, J. Roure, and J. Sole-Casals. "Selection of entropy based features for the analysis of the Archimedes' spiral applied to essential tremor." In *2015 4th International Work Conference on Bioinspired Intelligence (IWOBI)*, pp. 157–162. San Sebastian, Spain: IEEE, 2015.

[48] P. Zham, S. P. Arjunan, S. Raghav, and D. K. Kumar. "Efficacy of Guided Spiral Drawing in the Classification of Parkinson's Disease." *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1648–1652, 2018.

[49] C. R. Pereira et al. "A new computer vision-based approach to aid the diagnosis of Parkinson's disease." *Computer methods and programs in Biomedicine*, vol. 136, pp. 79–88, 2016.

[50] J. W. M. de Souza et al. "A New Approach to Diagnose Parkinson's Disease Using a Structural Cooccurrence Matrix for a Similarity Analysis." *Computational Intelligence and Neuroscience*, 2018.

[51] A. Rosebrock. *Deep Learning For Computer Vision With Python - Starter Bundle*. PyImageSearch.com, 2016.

[52] A. Rosebrock. *Deep Learning For Computer Vision With Python - Practitioner Bundle*. PyImageSearch.com, 2017.

[53] S. Albawi, T. A. Mohammed, and S. Al-Zawi. "Understanding of a Convolutional Neural Network." *ICET2017, Antalya, Turkey*, 2017.

[54] A. Krizhevsky et al. "ImageNet Classification with Deep Convolutional Neural Networks." *Communications of the ACM*, vol. 60, pp. 84–90, 2017.

[55] S. Ioffe and C. Szedegy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." *ICML'15 Proceedings of the 32nd International Conference on International Conference on Machine Learning*, vol. 37, pp. 448–456, 2015.

[56] F. Chollet. "BN Questions." URL `"https://github.com/fchollet/keras/issues/1802#issuecomment-187966878"`.

[57] D. Mishkin. "CaffeNet-Benchmark – Batch Norm." URL "https://github.com/duchaaiki/caffenet-benchmark/blob/master/batchnorm.md".

[58] R. Maalej et al. "Online Arabic Handwriting Recognition with Dropout applied in Deep Recurrent Neural Networks." *12th IAPR Workshop on Document Analysis Systems*, pp. 417–421, 2016.

[59] A. Rosebrock. *Hands-On Machine Learning with Scikit-Learn & tensorFlow - Concepts, Tools, And Techniques To Build Intelligent Systems*. O'Reilly, 2017.

[60] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[61] F. Chollet et al. "Keras." https://keras.io, 2015.

[62] Y. Bengio, N. Boulanger-Lewandowski, and R. Pascanu. "Advances in Optimizing Recurrent Networks." 2012. URL http://arxiv.org/abs/1212.0901.

[63] A. Zheng. *Evaluating Machine Learning Models*. O'Reilly, 2015.

[64] M. Sokolova and G. Lapalme. "A systematic analysis of performance measures for classification tasks." *Information Processing and Management*, vol. 45, pp. 427–437, 2009.

[65] M. Hossin and M. N. Sulaiman. "A Review on Evaluation Metrics For Data Classification Evaluations." *International Journal of Data Mining and Knowledge Management Process*, vol. 5, 2015.

[66] S. Yadav and S. Shukla. "Analysis of K-fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification." *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, 2016.

[67] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 1st ed., 2013.

[68] M. Kuhn and K. Johnson. *Applied Predictive Modelling*. Springer, 1st ed., 2013.

[69] R. Kohavi. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." *International Joint Conference on Artificial Intelligence*, 1995.

[70] K. Dutta, P. Krishman, M. Mathew, and C. V. Jawahar. "Improving CNN-RNN Hybrid Networks for Handwriting Recognition." In *2018 16th International Conference on Frontiers in Handwriting Recognition*, pp. 80–85. IEEE, 2018.

[71] A. Mikolajczyk and M. Grochowski. "Data augmentation for improving deep learning in image classification problem." In *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, pp. 117–122. IEEE, 2018.

[72] J. Patel, S. Pandya, and V. Shah. "Review on Generative Adversarial Networks." *International Journal of Technical Innovation in Modern Engineering and Science*, 2018.

[73] A. Odena. "Semi-Supervised Learning with Generative Adversarial Networks." *Data Efficient Machine Learning workshop at ICML*, 2016.

[74] A. Somasundaram and U. S. Reddy. "Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data." *Proceedings of 1st International Conference on Research in Engineering, Computers and Technology*, 2016.

[75] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[76] A. Anand, G. Pugalenthi, G. B. Fogel, and P. N. Suganthan. "An Approach for Classification of Highly Imbalanced Data using Weighting and Undersampling." *Amino Acids*, vol. 39, pp. 1385–1391, 2010.

[77] J. P. Maheswari. "Breaking the curse of small data sets in Machine Learning: Part 2." URL "https://towardsdatascience.com/breaking-the-curse-of-small-data-sets-in-machine-learning-part-2-894aa45277f4".

[78] A. Naseer et al. "Refining Parkinson's neurological disorder identification through deep transfer learning." *Neural Computing and Applications*, 2019.

[79] S. Lathuiliere, P. Mesejo, X. Alameda-Pineda, and R. Horaud. "A Comprehensive Analysis of Deep Regression." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 1–17, 2019.

[80] X. Zhou et al. "EAST: An Efficient and Accurate Scene Text Detector." *Megvii Technology Inc., Beijing, China*, 2017.

[81] A. Buades, B. Coll, and J.-M. Morel. "Non-Local Means Denoising." *Image Processing On Line*, vol. 1, pp. 208–212, 2011.

[82] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-Based Learning Applied to Document Recognition." *Proceedings of the IEEE*, pp. 1–46, 1998.

[83] C. Deotte. "Digit Recognizer: Learn computer vision fundamentals with the famous MNIST data." URL "https://www.kaggle.com/c/digit-recognizer/discussion/61480".

Appendix A

Magnetic Resonance guided Focused Ultrasound (MRgFUS)

MRgFUS has been found to be effective in treating tremor due to PD and ET [34]. MRgFUS treats tremors by ablating a portion of the thalamus, the part of the brain that serves as the relay station for transferring electrical signals, through a non-invasive approach [24]. When the treatment takes place the patient is situated within an MRI machine. The MRI machine allows the physician to locate the target area of the brain to be ablated as well as to monitor the temperature of the site of ablation using MR thermometry [24, 33, 35]. The patient is able to specify which hand they would prefer to receive treatment: the dominant or non-dominant hand [36].

After the patient's head is completely shaven, a stereotactic frame lined with an elastic silicone diaphragm is fixed to the skull [24, 33, 34]. In conjunction with the frame, the ultrasound transducers helmet is also fitted onto the patient's head and pre-planned sonications are applied to desired portions of the brain while regularly monitoring the patient's condition and adjusting the treatment accordingly [24, 33, 34]. Monitoring the condition of the patient involves acquiring hand tremor data. The procedure undertaken to acquire the data to monitor changes in tremor is detailed in Section 4.1.1 of Chapter 4. Spiral drawings are also performed while the patient is nearby a MRI machine [36]. Not all spiral acquistion methods are suitable for this environment. This was discussed further in Section 2.3.3.

Appendix B

Ethics Clearance Certificate

This appendix contains the ethics clearance for the secondary use of data in Figure B.1

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Mr Kelvin da Silva

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M190405

NAME: (Principal Investigator)	Mr Kelvin da Silva
DEPARTMENT:	School of Electrical and Information Engineering Faculty of Engineering and Built Environment
PROJECT TITLE:	Quantitative analysis of treatment effectiveness based on hand-drawn spirals by movement disorder patients
DATE CONSIDERED:	26/04/2019
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Prof Vered Aharonson and Prof David Rubin
APPROVED BY:	Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	29/04/2019

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the Third Floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in **April** and will therefore be due in the month of **April** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

Date 04/06/2019

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Figure B.1: Ethics clearance certificate for the secondary use of data.