TALLINN UNIVERSITY OF TECHNOLOGY

School of Information Technologies

Tanel Kossas 164119IAPB

OPTIMAL LENGTH OF FINE MOTOR TESTS FOR DIAGNOSING PARKINSON'S DISEASE

Bachelor's thesis

Supervisor: Sven Nõmm

PhD

Tallinn 2019

TALLINNA TEHNIKAÜLIKOOL

Infotehnoloogia teaduskond

Tanel Kossas 164119IAPB

PEENMOTOORTESTIDE OPTIMAALNE PIKKUS PARKINSONI TÕVE DIAGNOOSIMISEKS

Bakalaureusetöö

Juhendaja: Sven Nõmm
PhD

Tallinn 2019

Author's declaration of originality

I hereby certify that I am the sole author of this thesis and this thesis has not been presented for examination or submitted for defence anywhere else. All used materials, references to the literature and work of others have been cited.

Author: Tanel Kossas

27.05.2019

Abstract

Among the most employed diagnostic tests for Parkinson's disease (PD) are the pen-and-paper versions of Luria's alternating series tests, in which the patient is asked to draw a series of two alternating shapes, repeating for 5-7 times. In recent times, the same test has also been used in digitised form, using a digital pen and a smart device to measure precise information during the drawing process. When graduating the use of tools analysing digitised alternating series tests from prototypes to use in medical diagnosis, however, it is important for the lengths of the tests to be as short as possible – yet also still provide a valuable reference for diagnosis – in order to maximise their usability in the diagnostic tool. Despite this, to the best knowledge of the author, no research on the optimal length of the aforementioned digitised tests is available.

The main aim of the thesis was to determine the optimal length of six variations of Luria's alternating series tests, i.e. determine the shortest length of a test that at the same time does not decrease the classification power between PD patients and healthy controls by a significant amount. The thesis found that the length of a test could be reduced by up to 70% of the initial test length for half of the test types chosen, and by at least 30% in all the cases. This means that depending on the test type, only 1-4 repetitions are required for maintaining the same classification power, in comparison to the 5.5 repetitions that were used in most of the tests in the the test data. In addition, for facilitating diagnostic use and further related research, an application was developed for displaying detailed results of a given test drawn by the patient and data on the classification power of the test type.

This thesis is written in English and is 68 pages long, including 7 chapters, 31 figures and 6 tables.

Annotatsioon

Peenmotoortestide optimaalne pikkus Parkinsoni tõve diagnoosimiseks

Üks kasutatumaid meetodeid Parkinsoni tõve diagnoosimiseks on paberil teostatav Luria vahelduvate seeriate test, milles patsient joonistab üksteise järele kaks vahelduvat kujundit ning kordab seda 5-7 korda. Viimastel aastatel on sama testi kasutatud ka digitaalsel kujul: digitaalset pliiatsit ja nutiseadet kasutades mõõdetakse joonistamisprotsessi kohta suure täpsusega informatsiooni. Kui aga viia digitaalset vahelduvate seeriate testi analüüsivad vahendid prototüübist edasi kasutusse diagnostilise vahendina meditsiinis, on tähtis, et testid oleks võimalikult lühikesed, et suurendada testide kasutatavust, ent mis seejuures aitaksid endiselt võimalikult palju patsiendi õigele diagnoosile kaasa. Autorile teadaolevalt ei ole aga mainitud testide optimaalse pikkuse kohta uuringuid teostatud.

Käesoleva töö peamiseks eesmärgiks oli leida optimaalne pikkus kuue erineva Luria vahelduvate seeriate testi jaoks, ehk lühim testi pikkus, mille analüüs omab endiselt peaaegu sama suurt eristusvõimet Parkinsoni tõvega patsientide ja tervislike inimeste vahel kui tervet testi analüüsides. Käesolevas töös leiti, et testide pikkust võis poolte testide korral vähendada kuni 70% algse testi pikkusega võrreldes, ja igas testis vähemalt 30%. Seega on digitaalses vahelduvate seeriate testis vajalik olenevalt testist vaid 1-4 kordust, kui andmestikus oli enamike testide puhul kasutusel viis ja pool kordust.

Lisaks optimaalse pikkuse leidmisele arendati käesoleva töö käigus välja ka rakendus, eesmärgiga aidata kaasa tulevastele uuringutele antud valdkonnas ning ka tulevaseks diagnostiliseks kasutuseks. Rakenduses kuvatakse kasutajale patsiendi joonistatud testi detailsed tulemused ja antud testitüübi klassifitseerimise kvaliteeti mõõtvad näidud.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 68 leheküljel, 7 peatükki, 31 joonist ja 6 tabelit.

List of abbreviations and terms

PD	Parkinson’s disease
HC	Healthy control
LAST	Luria’s Alternating Series Test
dLAST	digitised Luria’s Alternating Series Test
SVC	Support Vector Classification
CART	Classification and Regression Tree
KNN	K-Nearest Neighbours
LR	Logistic Regression
UI	User Interface

Table of contents

List of figures

List of tables

1 Introduction

Recent advances in the digitisation of fine motor tests have allowed to collect more information compared to the classical pen and paper setting. Modern techniques in this area are based on the analysis of an already completed test. This means that one of the most important advantages of digitised tests, i.e. the ability to record and analyse precise information, remains unused.

Many recent studies of using digitised fine motor tests pertain to diagnosing Parkinson's disease (PD) [1, 2, 3]. PD has symptoms that greatly affect fine motor skills, and also a reliable and reasonably quick test is not yet available for diagnosing it [4]. These make PD a good candidate for developing a method of its diagnosis using digitised tests.

However, having fine motor tests designed to diagnose PD serves no use if the tests are too long to carry them out on patients, especially so if the patient already has problems with fine motor skills. Moreover, if the time taken by doctors and specialists to correctly diagnose PD could be reduced, that time could be allocated towards patient care. In addition, longer tests might introduce more noise into the data, making it less reliable, through for example reduced comfort as the test progresses.

As the field of analysing digitised fine motor tests is relatively young, the issue of how long an optimal test should be has not surfaced a lot, and to to the best knowledge of the author, there have been no studies on the optimal length of digitised tests in this area. It seems that the length of fine motor tests has been selected in an overly arbitrary manner to be long enough to see differences and short enough for the subject not to become fatigued, with the length gravitating towards longer than necessary, as it is always possible to make a test shorter in the process of analysis, as opposed to making it longer. When using the system on subjects needing to be diagnosed, however, unnecessarily long tests make tests difficult to carry out, and thus reduce their usefulness as a diagnostic tool.

As it is not possible to provide a general solution to all sorts of tests due to the classifying problem relying on experiments, the thesis examined six variations of Luria's alternating series tests, which are covered in detail in the background and methodology chapters.

The main aim of the thesis was to answer the question of what the length of the six Luria's alternating series tests have to be in order to be as short as possible and still maintain a similar level of classifying patients based on if they have PD or not. In addition, it was aimed to develop an application for visualising data pertaining to the task, to be used to facilitate academic research in the area of fine motor tests.

In the second chapter a theoretical background for the thesis is set. The third chapter explores research relating to the thesis. In the fourth chapter, the methodology of determining the optimal length for a test is explained, and in the fifth chapter, requirements and the implementation of the developed application are covered. The sixth chapter presents the results of the thesis.

2 Background

In this chapter the theoretical background for the underlying thesis is described.

2.1 Parkinson's disease

PD is the second most prevalent neurodegenerative disease after Alzheimer's disease, and in the coming years is expected to become an ever bigger problem in society, as the the average age increases. In 2006, the prevalence of PD was estimated at about 0.3% of the global population and about 1% in people over the age of 60 [4].

PD is characterised by four main symptoms: resting tremors, rigidity, bradykinesia i.e. slowness of movement, and postural instability. In addition to these, flexed posture and motor blocks are often included among the features of the disease in addition to the main four [5].

PD is a relatively mysterious disease regarding both its causes and what would be a good way of diagnosing it [4]. In the past, studies of PD were mostly very small and focused on prevalence. In the past decade, however, studies have reached the stage of being conducted on large sample sizes, allowing studies to examine risk factors and prognosis in a lot more detail [4].

In addition, as opposed to other neurodegenerative diseases, PD has treatments that mitigate symptoms [6]. Prescription medicine can improve motor and cerebral function and, in the cases where the treatments do not work or develop strong side effects, treatments such as deep brain stimulation offer improved quality of life [6]. Thus developing a method for making PD easy to diagnose serves a strong purpose.

2.1.1 Diagnosis

The diagnosis of PD is complicated. No reliable test has been developed as of yet that is easy to carry out. Analysing a patient by clinical criteria leads only to a probable diagnosis of PD, a definite diagnosis requires a post-mortem confirmation. The clinical criteria for a diagnosis

requires at least two of the four main symptoms described above to be present, and to rule out other diseases which could cause the four main symptoms, which are collectively referred to as parkinsonism [4].

One of the challenges in diagnosing PD is that the main symptoms may not be present before most of the dopaminergic neurons, responsible for motor functions, are lost. Thus the symptoms might not present themselves until a significant progression of the disease already exists. Another challenge is the need to identify subtle motor function features which are difficult to observe unless specifically looking for it, and an early diagnosis is made even harder by comorbidities – such as depression, anxiety, fatigue and sleep disorders – being present in the early stages of Parkinson's [7].

The lack of a relatively quick and reliable test and there being only a small amount of professionals able to diagnose PD makes it a very suitable task for attempting to diagnose the disease using statistics and machine learning. Thus, many approaches have been proposed for diagnosing PD, such as measuring and analysing data from rest tremors with the gyroscopes of a smartwatch [1], using kinematic and pressure features in handwriting for differential diagnosis [2] and also models combining handwriting and speech for diagnosis [3].

2.2 Luria's Alternating Series Test

Luria's Alternating Series Test (LAST) is a graphic (drawing) task often used to assess the ability of task-switching in patients with neurogenerative diseases. [8]. In the test, the patient is first shown an alternating series of some simple predefined shapes – usually rectangles, peaks or letters – and then asked to reproduce it in some predefined way – for example copying the sequence of continuing it –, in order to detect any disorders between the patient understanding, planning and executing the sequence of actions required for reproducing the shapes. In order to complete the LASTs, constant change of motion is required from the patient, which is made difficult by disorders affecting fine motor skills. Movements drawn in LASTs by patients with such disorders usually look overly deliberate and rigid as opposed to fluent. In the original paper-and-pen LAST, the set of two predefined shapes is to be repeated 5-7 times [9].

Though the alternating tests for in-person analysis delve more deeply into the specific part of

when the disorders occur – either in the process of understanding, planning or execution – [9], the goal of digitised Luria's Alternating Sequence Tests (dLAST) aims to perform analysis using datasets, calculated features and classification algorithms, and thus do not require the same sort of analysis. In addition, the usual pen-and-paper version of LAST is carried out by a practicioner drawing the shapes in front of the patient [9], whereas in dLASTs most often the series of shapes is already on the screen at the start of the test.

2.3 Welch's t-test

Many statistical procedures use mean comparison as a central point of analysis. One of the most widely known mean comparison tests is the t-test, which can be used to determine whether the statistical means of two sets of data originating from independent populations differ enough to be statistically significant [10].

Welch's t-test was developed from Student's t-test, which is more widely used than Welch's t-test, even though Welch's t-test performs better in most conditions. Student's t-test assumes that the variances between two populations are equal, whereas Welch's t-test can be used in the cases of both unequal and equal populations [10].

The formula for calculating Welch's t-statistic is the following:

$$t_w = \frac{\overline{y_1} - \overline{y_2}}{\sqrt{{s_1}^2/n_1 + {s_2}^2/n_2}}.$$

$\overline{y_1}$ and $\overline{y_2}$ are the means of the two populations, ${s_1}^2$ and ${s_2}^2$ are the sample variances of the populations and n_1 and n_2 are the amount of samples in the populations [10].

2.4 Fisher's score

As a supervised way of deciding the best features for a given class, Fisher's score has been widely used in the recent times. [11] Fisher's scoring is a ratio of average interclass separation to the average intraclass separation. This means that the greater the score is, the more discriminating the feature. Although Fisher's score is meant for analysing features independently of each other, this is suboptimal when creating a classifier, and thus some relatively

complex approaches of selecting featuresets jointly have been made, based on Fisher's score [12]. The formula for calculating Fisher's score is presented below [13]:

$$F = \frac{\sum_{j=1}^{k} p_j (u_j - \mu)^2}{\sum_{j=1}^{k} p_j {\sigma_j}^2}.$$

μ_j and σ_j are, respectively, the mean and standard deviation of data point belonging to class j for a given feature. p_j is the amount of data points belonging to class j for a given feature. μ is the global mean of the data points belonging to all classes under examination [13].

2.5 Classification models

In this chapter four fundamental classification models that are used in the thesis will be explored.

2.5.1 Support Vector Classification

Support Vector Classification (SVC) is the classification variant of Support Vector Machines (SVM), as opposed to a regression variant. The purpose of SVMs is to construct a hyperplane or a set of hyperplanes in a high-dimensional space which have the largest distance to the nearest training data points of any class. The most easily comprehensible version of this is a two-dimensional map of points, with a line chosen to separate points that belong to the two different classes [14].

In addition, a margin is often used with the classification model in order to fit the data better. The idea of the margin is to select points within the distance of the margin to the vector, and minimise the distance of the points within the margin [14].

2.5.2 Classification and Regression Tree

Decision trees are considered to be one of the most effective and computationally efficient machine learning approaches. In addition, the tree is easily understood once completed, which makes decision trees popular within the artificial intelligence field. There are a large amount of different algorithms for creating decision trees [15].

In this thesis the algorithm of choice is the classification and regression tree (CART). The tree is created by recursively splitting subsets of the dataset, using all given features, to create two child nodes repeatedly, with the goal of creating subsets of the dataset that are as homogeneous – regarding features – as possible, and of course, accurate within the training data as well [16].

2.5.3 K-Nearest Neighbours

K-nearest neighbour (KNN) is a widely known classifier that is extensively used in fields requiring pattern recognition [17]. Even though KNN is by far the simplest classification algorithm among the four explored in this thesis, it is included because it is a good classification algorithm in many cases.

KNN chooses the classification of an unclassified sample point based on the classification of the nearest k points of the set of previously classified points [18].

2.5.4 Logistic regression

Logistic regression (LR), also known as maximum-entropy classification, is a linear classification model. LR uses a logistic function for describing the possible outcomes of a single trial [19].

As opposed to other classification models discussed, LR provides a continuous value as output, signifying the probability of the uncategorised data point being part of a class. Thus, LR is in the strict sense not a classification model, but a model of probability. However, it is often turned into a classification model by selecting a cutoff value – if an unclassified data point has a higher value than the cutoff value, the point will be categorised as one class, if it has a lower value, it will be classified as another class [20].

2.6 Classifier cross-validation

In order to make sure that a classification model can be generalised to new data without overfitting, model validation needs to be performed. The main model validation method used in this thesis is k-Fold Cross-Validation. K-fold cross-validation allows to use all test data for

both training and testing, allowing to make maximum use especially of small datasets [21].

The process of cross-validating models is the following. Firstly, the dataset is split into k equal or nearly equal segments. Then, training and testing is carried out in k iterations, every time leaving out one of the segments when training and using the segment as test data after the training, leaving out a different segment every time. Then, an average of the accuracy of each iteration is calculated to get the accuracy of the model [21].

Often the data also requires stratification, a process of rearranging data so that every fold has approximately the same proportion of classes as the whole does, to maintain accurate training and testing. In order to further increase accuracy, several passes of random fold-picking can also be used in conjunction with stratification [21].

3 Methodology

In this chapter the methodology of determining the optimal length for tests will be discussed. The process of determining optimal length is presented in simplifed form on Figure 1.

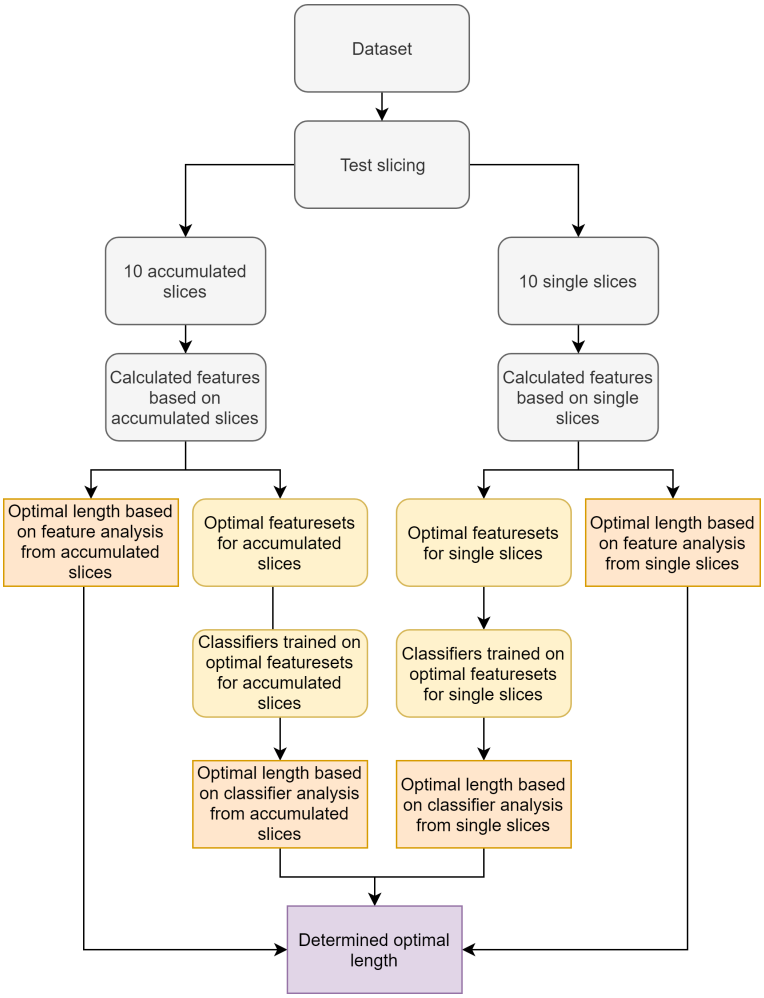


Figure 1. Process of determining optimal test length

3.1 Data collection

The dataset used in this thesis originates from tests carried out on 16 patients with PD and 12 healthy controls (HC) of approximately the same age – with a mean age of 65 – and equal gender distribution, and is the same dataset as was used in a related paper [22] that analysed

the viability of using dLASTs for diagnosing PD. For data collection, a digital pen – Apple Pencil – was used alongside an Apple iPad Pro 9.7-inch tablet equipped with an application developed specificially for collecting the data.

When conducting the test, the tablet was first positioned in front of the seated testee by the test conductor, after which the test conductor ensured that the right hand of the testee was supported by the surface of the table and then handed the digital pen to the testee. Then the patient was, in the cases of dLAST tests, shown a reference alternating series on the tablet, and asked to, depending on the test, either continue, copy or trace the reference alternating series. After one of the tests was completed, another one was immediately shown until all tests were completed.

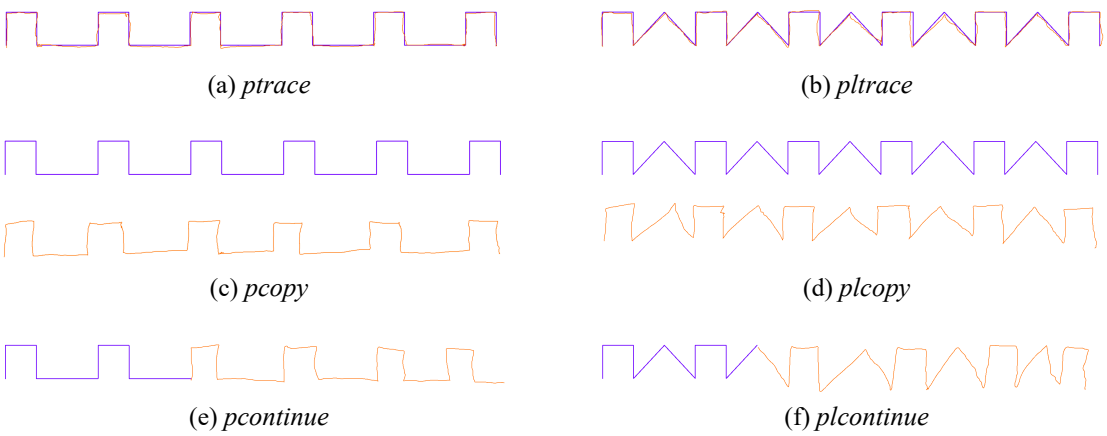
Parameters that were measured in one time instant and used in determining the optimal length were: x and y, denoting the horizontal and vertical position of the tip of the digital pen on the tablet in pixels, t, denoting the amount of seconds since midnight Jan 1, 2001, with the precision of 10 nanoseconds, and p, denoting the amount of pressure applied to the tip of the digital pen. The parameters were measured in up to 200 time instants per second [22].

In the dataset, all recorded time instances were divided into movements, meaning single continuous pen movements where the tip of the pen did not leave the surface of the tablet. No information was recorded when the tip of the pen was not on the surface of the tablet. In order to also take information and movement between the recorded movements into account, all of the recorded movements were combined into and analysed as one.

Six of the test types present in the dataset are examined in the thesis. These tests are variations of dLASTs: *pltrace*, *plcopy*, *plcontinue*, *ptrace*, *pcopy*, *pcontinue*. On Figure 2 the reference alternating series that the patient was shown has been painted blue and the line the patient drew has been painted orange.

3.2 Test slicing

In order to give an optimal assessment on what the optimal length of a test was, the existing test data was artificially limited during modelling. The area between the startpoint and endpoint



(a) *ptrace* (b) *pltrace*

(c) *pcopy* (d) *plcopy*

(e) *pcontinue* (f) *plcontinue*

Figure 2. Example patient tests on examined test types

of the test was sliced horizontally into ten equal slices, as seen in 3. This made it possible to analyse the differences between the test being the same length and the test being 10 percent of the length up until 90 percent of the length of the length of the initial test.

Two different methods of test slicing were used in analysis, referred to as single slicing and accumulated slicing. Single slicing refers to taking just the points within the slice under examination, i.e. if the slice is the 10% slice – starting from 0% and going up until 10% –, only the points in the slice are analysed. Single slicing allowed examination of the way the separation of PD patients and HCs differs in different points of the test, not accounting for other slices. For accumulated slicing, all the points from the beginning of the test up until the final point in the given slice are analysed. Accumulated slicing allowed examination of progressive separation between PD patients and HCs. The 100% accumulated slice is in the traditional sense the equivalent of analysing the whole test.

As both PD patients and HCs do not draw in perfectly straight horizontal and vertical lines, but we still want the slicing to best resemble what the test would be like if it was made shorter, the concept of expanding the range of a slice was brought in. The concept can be seen at work on Figure 4. In the example, the right side of the centre-left p (Π) would be excluded. However, if the test length was reduced to a length that would cut off at the end of the downwards line of the same p, the downwards line would still be included if we expand the range.

The algorithm for slicing is thus relatively simple. The last point in time that is placed between the horizontal start and horizontal end of the slice is assigned as the endpoint of the slice. Then the start of a slice is the selected endpoint of the previous slice. The two edge cases of the

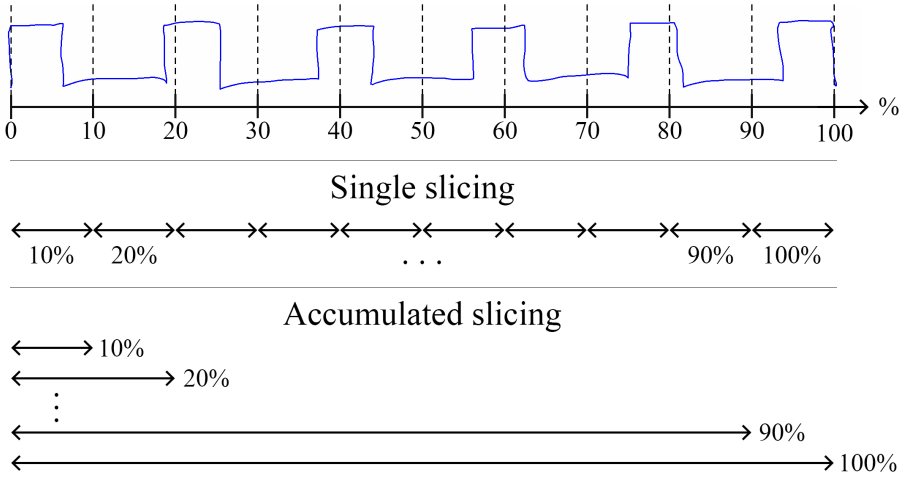


Figure 3. Accumulated and single slicing example

algorithm are when a point is on the left of the first slice and when a point is on the right of the last slice. In these events the points are assigned to be respectively on the first or the last slice.

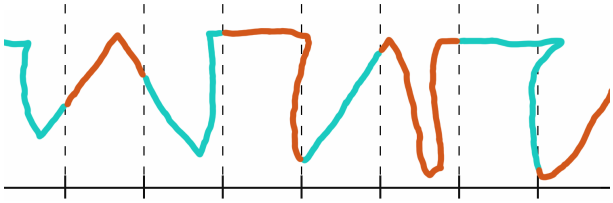


Figure 4. Example of expanding the range of a slice

3.3 Model features

Based on four parameters in every point – x, y, t and p –, 31 features were calculated. A feature is a set of values calculated from the set of initial parameters. For example, velocity can be a feature, being a set of values, calculated from the parameters of t, x and y. The features explored were selected from papers studying similar fine motor tests, consisting of papers [2] and [23].

Twenty eight of the features explored are based on parameters of motion mass, introduced in [23]. The features based on parameters of motion mass can roughly be divided into four: features based on total length and its derivatives, features based on drawing direction, features

based on pressure and completion time.

Calculating total length is done with equation (1), where l_i is the length between point i and point $i-1$. There are three variations of features with total length calculation. L_{Tx} is calculated based on l_{x_i}, i.e. just the horizontal (x) length, disregarding vertical (y) length. L_{Ty} is calculated based on l_{y_i}, i.e. just the vertical length, disregarding horizontal length. And L_{Te} is based on the Euclidean length between two successive points, i.e. the actual length $\sqrt{{l_{x_i}}^2 + {l_{y_i}}^2}$. In addition for each of the named features a temporally relative feature is calculated with the completion time t. Thus the first set of features L_{mT} is the set $\{L_{Tx}$, L_{Ty}, L_{Te}, L_{Tx}/t, L_{Ty}/t, $L_{Te}/t\}$.

$$L_T = \sum_{i=1}^{N} l_i \tag{1}$$

Calculating velocity mass is done with equation (2), where v_i is the velocity, i.e. the numerical derivative of length, in point i, calculated with three points. As with total length, there are three variations of the features – V_{Tx}, V_{Ty} and V_{Te} – and a temporally relative feature is calculated for each feature based on test completion time t. Thus the second set of features V_{mT} is the set $\{V_{Tx}$, V_{Ty}, V_{Te}, V_{Tx}/t, V_{Ty}/t, $V_{Te}/t\}$.

$$V_T = \sum_{i=1}^{N} |v_i| \tag{2}$$

Calculating acceleration mass is done with equation (3), where a_i is the acceleration, i.e. the numerical derivative of velocity, in point i, calculated with four points. The features are derived in the same way as in the previous two sets of features. The third set of features A_{mT} is thus $\{A_{Tx}$, A_{Ty}, A_{Te}, A_{Tx}/t, A_{Ty}/t, $A_{Te}/t\}$.

$$A_T = \sum_{i=1}^{N} |a_i| \tag{3}$$

Calculating jerk mass is done with equation (4), where j_i is the jerk, i.e. the numerical derivative of velocity, in point i, calculated with five points. The features are derived in the same way as in the previous three sets of features. The fourth set of features J_{mT} is thus $\{J_{Tx}$, J_{Ty},

J_{Te}, A_{Tx}/t, A_{Ty}/t, $A_{Te}/t\}$.

$$J_T = \sum_{i=1}^{N} |j_i| \tag{4}$$

The directional mass feature D_T, also known by the name of angular mass, is calculated with equation (5), where d_i is the angle between the drawing direction vectors at time instant $i-1$ i. For directional mass, a version of the feature relative to test completion time t, is also calculated. Thus the fourth set of features D_{mT} is $\{D_T, D_T/t\}$.

$$D_T = \sum_{i=1}^{N} |d_i| \tag{5}$$

The pressure mass feature P_T is calculated with equation (6), where p_i is the pressure p at time instant i. As pressure is in this case a strictly positive value, the equation can also be simplified by removing the absolute value. For P_T, a temporally relative version P_T/t is calculated as well. Thus the fifth set of features P_{mT} is $\{P_T, P_T/t\}$.

$$P_T = \sum_{i=1}^{N} |p_i| \tag{6}$$

The features based on motion mass can thus be said to be a set M_T:

$$M_T = \{L_{mT}, V_{mT}, A_{mT}, J_{mT}, D_{mT}, P_{mT}\}.$$

In addition to the features extrapolated from motion mass parameters, three more features were added to the featureset. Two additional features were selected to model pressure: the number of changes in pressure – NCP – and NCP relative to the completion time of the test NCP/t. NCP is the amount of local extrema on the graph of pressure. NCP is a feature that is meant to measure the amount of jitter in pressure levels. The final feature examined in this thesis is the completion time of the test itself, t.

3.4 Initial feature filtering

In order to reduce the amount of features that needed to be analysed, features that did not perform well for classification needed to be removed. To determine which features did not

perform well, for every feature in every slice in each test type, Welch's t-test's p-values and Fisher's scores were calculated using test data for the test type. This was done separately for both single and accumulated slices.

The method used to select features for exclusion from further analysis was to, separately for every combination of test type and slicing method, calculate for every feature j the maximum Fisher's score FS_jmax it had across all slices and the minimum p-value PV_jmin it had for every slice, and then form two sets, in one of which are features that are in the bottom 60% regarding maximum Fisher's score, and in another features that are in the top 60% regarding minimum p-values. Finally, the set of features to exclude were formed by the union of the two sets.

This method was used in order to include features that had either high Fisher's scores or low p-values. As Fisher's scores and p-values function differently, but both methods are often used for feature selection, this method aims to exclude only the features shown by both methods of feature selection to be relatively poor at separating PD patients and HCs.

3.5 Feature analysis

It is difficult to determine the optimal length of a test by looking at single values calculated based on the features. Using statistical means, medians and minimum and maximum amounts for the calculated Fisher's scores assigns no meaning to the actual progress of the test. Using linear regression could be used to gain an idea of whether the data leans upward or downward, but cannot show local maximus or minimums.

Therefore, it was decided to employ analysis based on observation of the graphs of p-values and Fisher's scores of the features included in the initial feature analysis. The goal of the analysis was to determine optimal length assessments for the test types and, based on the optimal length assessments, also a featureset containing features with the best classification power to be used for training a classifier.

Two optimal length assessments were selected for every combination of test type and slicing method. Usually, one length was selected to be the global maximums of Fisher's scores and global minimums of p-values and the other an earlier length signifying respectively the local

maximums and minimums of Fisher's scores and p-values, that did not have a significant difference in scores from the global maximums and minimums of the longer length. As slices are being analysed, the length in this case is measured in slices.

Then, for both length assessments in every combination of test type and slicing method, a featureset for training the classifiers was selected based on the highest scoring features in the optimal slice selected based on both Fisher's score and p-values. Three features were selected into every featureset, as more features would produce too much overfitting for the dataset of 28 people total.

3.6 Classifier training and analysis

Classifiers were trained to provide an alternate assessement for optimal lengths. The specific featuresets used for training were selected in feature analysis. Once the featuresets were obtained, four or eight classifiers were trained – depending on the featuresets chosen – for every combination of test type, slice and slicing method. Four classifiers were trained in the case that the featuresets chosen for both of the optimal lengths were the same. Four different classifier models were trained: SVC, CART, KNN and LR.

Firstly, the dataset was split into training and testing groups using a stratified k-fold method with a k of 4. A seed was set for the splitting in order for the splits to be the same in all test cases. Then, the classification model was fitted to the training data. During this, a cross-validating grid search took place for finding the best hyperparameters for the given training data, optimising for accuracy score with a k-fold of 4. Then, the accuracy score of the trained classifier was recorded for the testing groups. The process after the splitting was repeated for a total of four times to obtain the results for all splits. Then, the mean of the four results was taken to obtain the final cross-validated score. To reiterate, this training took place for every classifier separately in every slice for the 1-2 datasets proposed for a given slicing method, for every test type.

The analysis of the classifier results took place in a manner similar to feature analysis. Graphs of the accuracy scores of the trained classifiers were analysed to give another assessment on

the optimal length for the test types.

3.7 Determining optimal length

After classifier analysis the test lengths that were considered optimal in feature analysis and the test lengths that were considered optimal in classifier analysis were analysed together, and a final determination of optimal lengths was proposed.

The determined optimal lengths of the test types were also converted into the number of repetitions the alternating series would have to be for the resulting test to be of the determined optimal length.

4 Application

In order to facilitate analysis in this thesis and also to facilitate future analysis, an application was developed to show detailed results.

4.1 Requirements

For guiding the development of the application, several requirements to the application were set forth by the author:

4.1.1 Functional requirements

The following are the functional requirements imposed on the application by the author:

- allow for playback of the test in realtime,
- allow for slower than realtime playback for more detailed analysis,
- display realtime feature values in regards to the current time instant of the test,
- display the whole population's Fisher's scores along with p-values in a separate category.

4.1.2 Nonfunctional requirements

The following are the nonfunctional requirements imposed on the application by the author:

- allow analysis of a wide range of medical conditions with few modifications to the program's code, provided that the data is in the same format,
- allow for simple feature expansion for use in future research,

- allow compiling the application into an executable for easy use.

4.2 Implementation

The application used PyQt as the main user interface (UI) framework, which implements the cross-platform Qt framework in the Python programming language [24]. Even though Python is an interpreted language as opposed to compiled, it can still be compiled into cross-platform executables [25]. This was another reason Python and PyQt were selected.

For plotting graphs in the UI, the pyqtgraph library was used. As realtime plots are in constant need of being updated, pyqtgraph is used due to its speed advantage over another choice, matplotlib [26]. In cases where graphs would have to be more complex, however, e.g. graphing the trained classification class areas along with training points, these plots would need to be created with the matplotlib library due to to pyqtgraph missing complex functionality. However, both libraries can coexist with one another in PyQt and thus adding classification plots for future work would not be difficult in terms of adding it to the UI.

A minor detail is that the author decided to omit displaying the time the test was taken, which would be the only place where an absolute value of time would be used. In all other places only relative time is used, i.e. measuring the difference in relation to for example the previous or the first datapoint. In the dataset under consideration, an iOS timestamp was used. In addition, all other input data values – x, y and p – are absolute, but are treated in the same way as t, meaning that even if the test is carried over to screens of a different size and the pressure metric is changed, the application can still be used as an analytical tool in the same way, even if the different test conditions may provide a different conclusion.

4.3 Data visualisation

The main objective of the application is to provide a means of test data visualisation, and to

also display feature scores for the given test type associated with the test.

4.3.1 Main application section

On Figure 5, the main section of the application is presented. On the upper part of the application the test drawing can be seen, 5.36 seconds after starting the test, along with time controls allowing to change the playback speed or advance the time in steps of 1 and .1 seconds with the option to also go to the beginning or end of the test. By dragging the slider, the user can go to a specific instant of time in the test. The same section with the final test results can be observed on Figure 6.

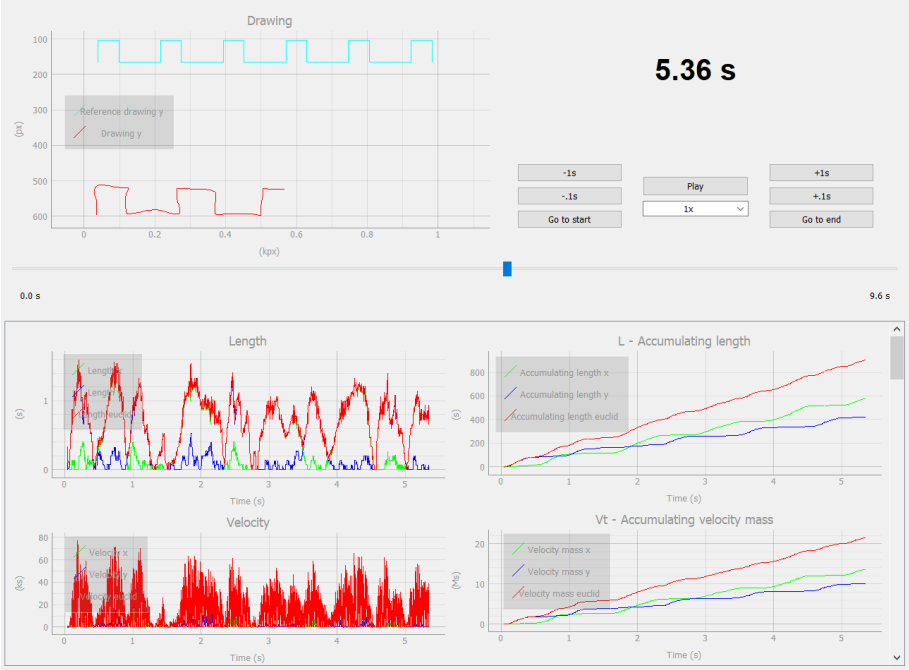


Figure 5. Main page of the developed application with test halfway through

The bottom part of the main section comprises of graphs plotting features discussed in the previous chapter on the right hand side, and also plain length, velocity, jerk and pressure on the left hand side. When the test is played back or when a different instant of time is selected with the slider, the graphs adjust to display only the test data up until that point, i.e. realtime playback is accompanied by realtime feature display. This is useful for when the user wants to examine how a certain small movement in the test affects the values of the calculated features. The graphs can be further examined by zooming in with the scrollwheel, or by manually increasing the size of the subsection the graphs are in.

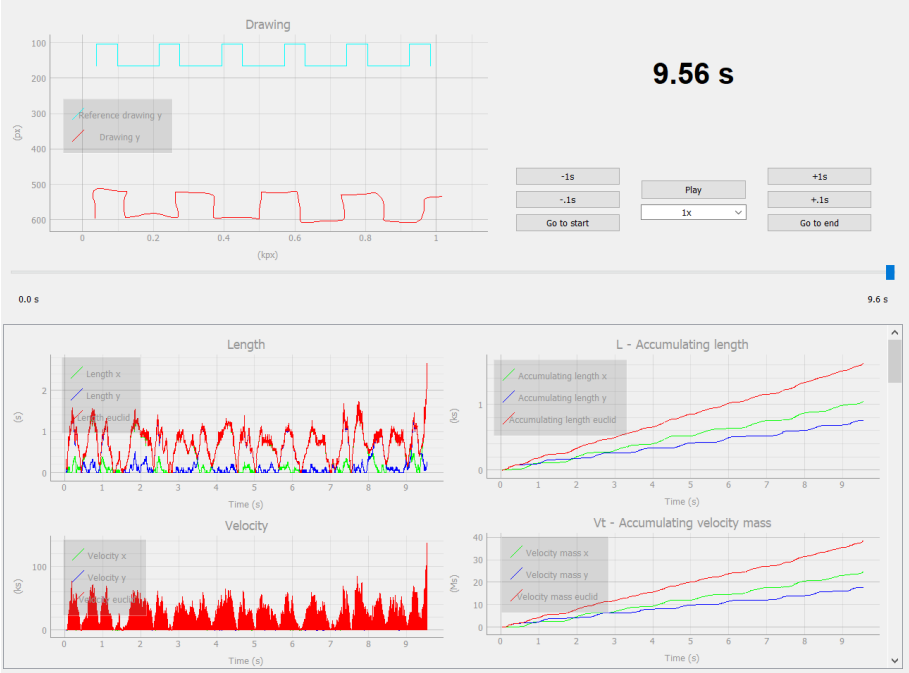


Figure 6. Main page of the developed application with finished test

4.3.2 Fisher's score and Welch's t-test section

On Figure 7, the secondary section of the application is presented. As opposed to the main page, no realtime plotting takes place in this page, only previously calculated Fisher's scores and p-values are displayed for the features examined in the thesis. The main use of this page is to examine the viability of a feature and a test type when attempting to classify two groups, PD patients and HCs in the case of the thesis.

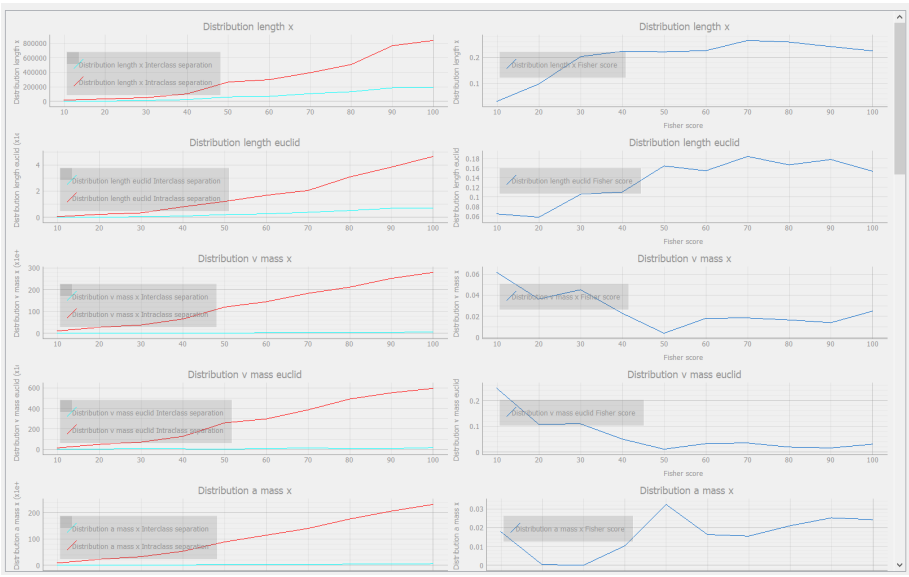


Figure 7. A page of the developed application showing Fisher's scores

5 Results

In this chapter the most significant results will be discussed.

5.1 Initial feature filtering

As the first step of finding an optimal length, features that were not discerning enough are removed. For this, the proposed filtering method was used, with which an X of 60 was decided to be used, in order to filter out many of the features showing a far lesser classification power compared to the ones with the best Fisher's scores. The full list of features excluded and included in initial feature filtering, can be viewed in Appendix 2.

5.1.1 Single slicing

From Table 1 it can be inferred that in all cases, except for *pcopy*, the included features have a higher FS_fmax value than all the excluded features, and in some cases, e.g. *ptrace*, by a large margin. In addition, in all test type cases – except in the case of *plcontinue* – every included feature also has a lower or equal PV_fmin value among the features than the highest PV_fmin value of the excluded features. However, even in the case of *plcontinue*, the difference is not too significant, as only two features have a higher PV_fmin, having their FS_fmax at 0.9366 and 1.5876 respectively, while the highest PV_fmin has its Fisher's score at 0.4883. Thus the difference between the included and excluded features is relatively significant and none of the excluded features offer as much classification power in regards to Fisher's scores and p-values as any of the included features.

It was noted that in several cases a feature was not excluded because of being in the top 40% in one of the feature, proving the point mentioned in the methodology – that the two scoring methods provide different rankings for the same features.

The filtering method excluded 96 of the total 186 test type and feature combinations. Of 31 features, 13 were completely excluded, i.e. excluded from all test types: L_{Tx}, L_{Te}, V_{Tx}, V_{Ty},

Table 1. Minimum and maximum feature scores in initial filtering for single slicing

Test type	Included features		Excluded features	
	$minFS_fmax$	$maxPV_fmin$	$maxFS_fmax$	$minPV_fmin$
pcontinue	0.3931	0.0037	0.3299	0.0038
plcontinue	0.5173	0.0040	0.3868	0.0026
pcopy	0.5157	0.0041	0.5289	0.0050
plcopy	0.5250	0.0009	0.4732	0.0013
ptrace	0.9419	0.0001	0.5164	0.0006
pltrace	0.5517	0.0012	0.4364	0.0012

Table 2. Removed and remaining feature counts for single slicing

Test Type	Features excluded	Features included
pcontinue	15	16
plcontinue	14	17
pcopy	17	14
plcopy	18	13
ptrace	17	14
pltrace	15	16

V_{Te}, A_{Tx}, A_{Ty}, A_{Te}, J_{Tx}, J_{Ty}, J_{Te}, P_T/t and NCP/t. The amount of features excluded for each test type is displayed in Table 2. This also corresponds to how different the maximum p-value and Fisher's score rankings are for features. In the case of *plcopy*, the theoretical maximum of excluding 60% of features was achieved, as the sets of the bottom 60% of features ranked separately by FS_fmax and PV_fmin were equal. In other test types however, the ranking differed which resulted in less features being excluded.

As can be seen in Table 3, despite a lot of the features modelled on the basis of motion mass parameters having been filtered out, the temporally relative variations of the original motion mass parameters L_T, V_T, A_T and J_T all remain in the included features of each test type, with the exception of J_{Ty} being only in the included features of four test types. This suggests that on these test types the temporally relative versions of the original motion mass parameters have a better discriminating power than the non-temporal versions, which can also be observed in the the feature Fisher's score and p-value tables in Appendices 2-7.

Table 3. Amount of test types a feature is included in after initial filtering for single slicing

Features	Amount of test types the features are included in
L_{Tx}/t, L_{Ty}/t, L_{Te}/t, V_{Tx}/t, V_{Ty}/t, V_{Te}/t, A_{Tx}/t, A_{Ty}/t, A_{Te}/t, J_{Tx}/t, J_{Te}/t, NCP	6
D_T	5
J_{Ty}/t	4
t	3
L_{Ty}, D_T/t, P_T	2

Table 4. Minimum and maximum feature scores in initial filtering for accumulated slicing

Test type	Included features		Excluded features	
	$minFS_fmax$	$maxPV_fmin$	$maxFS_fmax$	$minPV_fmin$
pcontinue	0.4885	0.0027	0.4520	0.0012
plcontinue	0.4355	0.0031	0.2336	0.0123
pcopy	0.3982	0.0069	0.3227	0.0105
plcopy	0.5829	0.0006	0.4287	0.0020
ptrace	0.8663	0.0003	0.4031	0.0017
pltrace	0.4240	0.0017	0.3645	0.0024

5.1.2 Accumulated slicing

From Table 4 the exact same can be referred as from Table 1: that the difference between included and excluded features is relatively significant. In accumulated slicing, of all the test types, *pcontinue* is still the only test to have a maximum PV_fmin in included features higher than the lowest PV_fmax of excluded features, just as in the single version. Apart from this, all other test types show that all included features are better than all excluded features regarding both maximum Fisher's scores and minimum p-values across all slices.

The filtering method removed 97 of the total 186 test type and feature combinations. Of 31

Table 5. Removed and remaining feature counts for accumulated slicing

Test Type	Features excluded	Features included
pcontinue	16	15
plcontinue	12	19
pcopy	16	15
plcopy	18	13
ptrace	17	14
pltrace	16	15

Table 6. Amount of test types a feature is included in after initial filtering for accumulated slicing

Features	Amount of test types the features are included in
L_{Tx}/t, L_{Ty}/t, L_{Te}/t, V_{Tx}/t, V_{Ty}/t, V_{Te}/t, A_{Tx}/t, A_{Ty}/t, A_{Te}/t, J_{Tx}/t, J_{Ty}/t, J_{Te}/t	6
NCP	5
D_T, t	4
L_{Ty}	2
L_{Te}, A_{Tx}, D_T/t, P_T	1

features, 11 were completely excluded: L_{Tx}, V_{Tx}, V_{Ty}, V_{Te}, A_{Tx}, V_{Te}, J_{Tx}, J_{Ty}, J_{Te}, P_T/t and NCP/t. A summary of the amounts of features excluded based on test type are displayed in Table 5. In the case of *plcopy*, the theoretical maximum of excluding 60% of features was achieved just like when using single slicing.

Just as in single slicing, a large part of the features modelled on the basis of motion mass parameters have been filtered out, and – as can be seen in Table 6 – the temporally relative elements of L_T, V_T, A_T and J_T are included in most test types, and in this case of accumulated slicing, in all of the test types. There are minor variations in the initial filtering for single slicing and accumulated slicing, but overall the features included are quite similar.

5.2 Feature analysis

An example of what is being looked for in the thesis can be observed on Figure 8. From this figure it can be inferred that if only 30 percent of the test was analysed, only 0.036 would be lost in Fisher's score compared to if the whole test was analysed. Of course it can also be seen that an even better Fisher's score – or classification power – can be achieved if the test was analysed at the 80% point instead of the 100% point. Overall, two things were observed regarding the starts and ends of the tests in the dataset.

Firstly, the beginning of the test is more random, and thus in most cases has low classification power. However, even the 0.8 Fisher's score depicted on Figure 8 is a relatively good classification score, as during initial feature filtering, Fisher's scores of about 0.3 corresponded to p-values of approximately 0.01. Often used statistical significance values are 0.1, 0.05 and 0.01 [27], thus in this case even the first ten percent of a test can be considered statistically

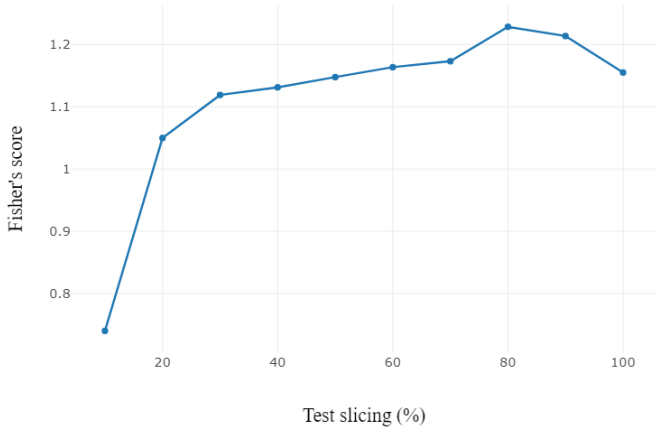


Figure 8. Fisher's scores of feature J_{Te}/t on *ptrace* test type

significant.

Secondly, just as on Figure 8, a slight decrease in Fisher's scores was observed towards the end of the test in some of the features. However, most of the figures followed a trend of either staying about the same as the previous slicing or increasing, as is the case on Figure 9.

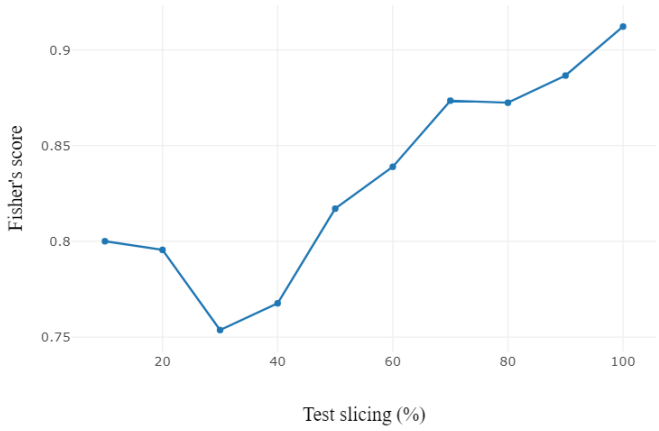
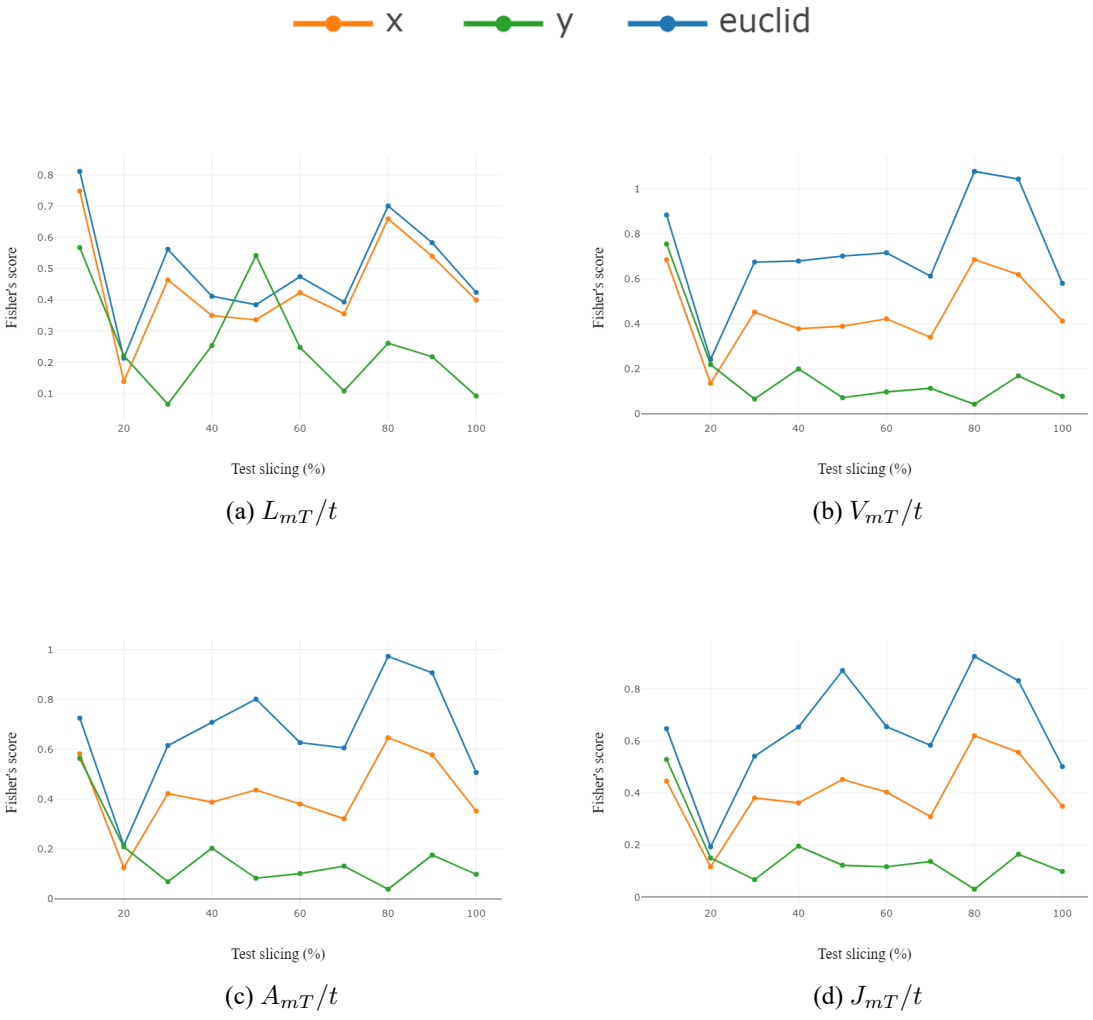


Figure 9. Fisher's scores of feature V_{Te}/t on *pltrace* test type

5.2.1 Pcontinue

For single slicing of *pcontinue*, the relative variants of the main motion mass features – L_{mT}/t, V_{mT}/t, A_{mT}/t and J_{mT}/t – proved to be best regarding both p-values and Fisher's

scores. Other features included in the single slicing variant – L_{Ty}, D_T, D_T/t, P_T, NCP and t – were very volatile, changing classification power by a large margin from one slice to the next, in both p-values and Fisher's scores, as seen respectively in Appendix 3 Table 1 and Appendix 3 Table 2. In addition, they had a low average and maximum Fisher's score, none of them reaching a peak of 0.6 Fisher's score.



(a) L_{mT}/t (b) V_{mT}/t

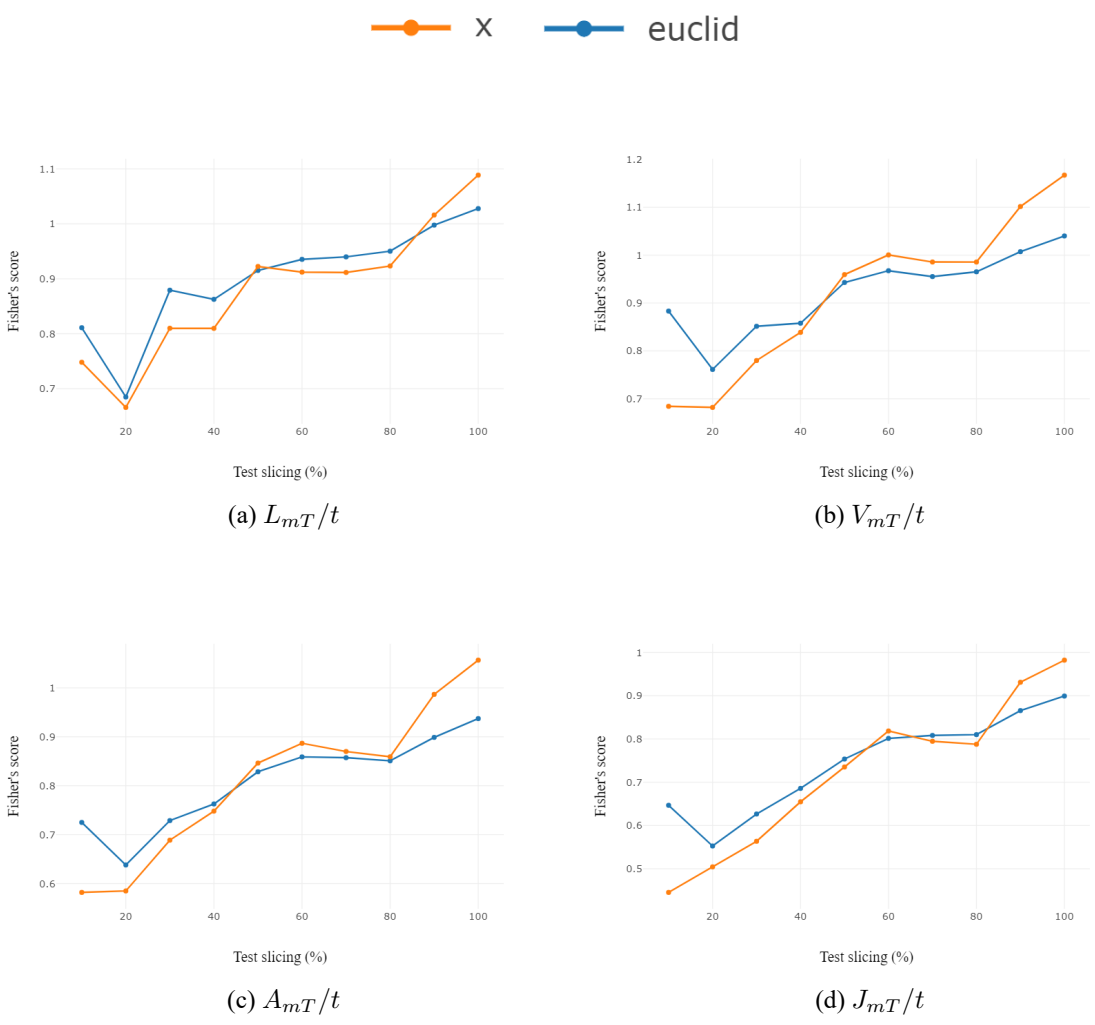
(c) A_{mT}/t (d) J_{mT}/t

Figure 10. Fisher's scores of x, y and e variants of relative motion mass features for single slicing in *pcontinue*

It can be observed on Figure 10 and in Appendix 3 Table 1 that in all cases and in all points – except the one point in the case of LmT/t – the Euclidean variants of relative motion mass features perform better. This is also evident in the p-values, seen in Appendix 3 Table 2. The optimal single slicing lengths chosen for the test are based on the local extrema on the 50% slicing and the global extrema on the 80% slicing. Due to L_{mT}/t having relatively low extrema, it will be excluded. Thus the optimal featureset chosen for both lengths is $\{V_{Te}/t$, A_{Te}/t, $J_{Te}/t\}$. For the 50% slicing, the mean Fisher's score of the features is 0.79. For the 80% slicing, the mean Fisher's score is 0.99. Thus for single slicing, an average of 0.20

Fisher's score would be lost per feature when using the 50% slicing over the 80% slicing. The respective means for p-values are 0.0012 and 0.0005.

For accumulated slicing, when considering the relative parts of the feature tuples of L_{mT}, V_{mT}, A_{mT} and J_{mT}, in all four cases the variation of the motion mass parameter with just the horizontal component x had both a higher maximum Fisher's score and a lower minimum Fisher's score than the combined Euclidean distance component e, with the figure appearances maintaining roughly the same shape, as can be seen on Figure 11 and in Appendix 3 Table 3. Thus, for all cases in the mentioned feature tuple, when the slice selected was 10% of the test, the Euclidean variant had better discriminating power, and when the slice selected was 100% of the test, the variant with just the horizontal component x was better. In the event that the slice selected was between 10% and 100%, there was no certain crossover point – which of the variations was better depended on the selected feature tuple and slice.



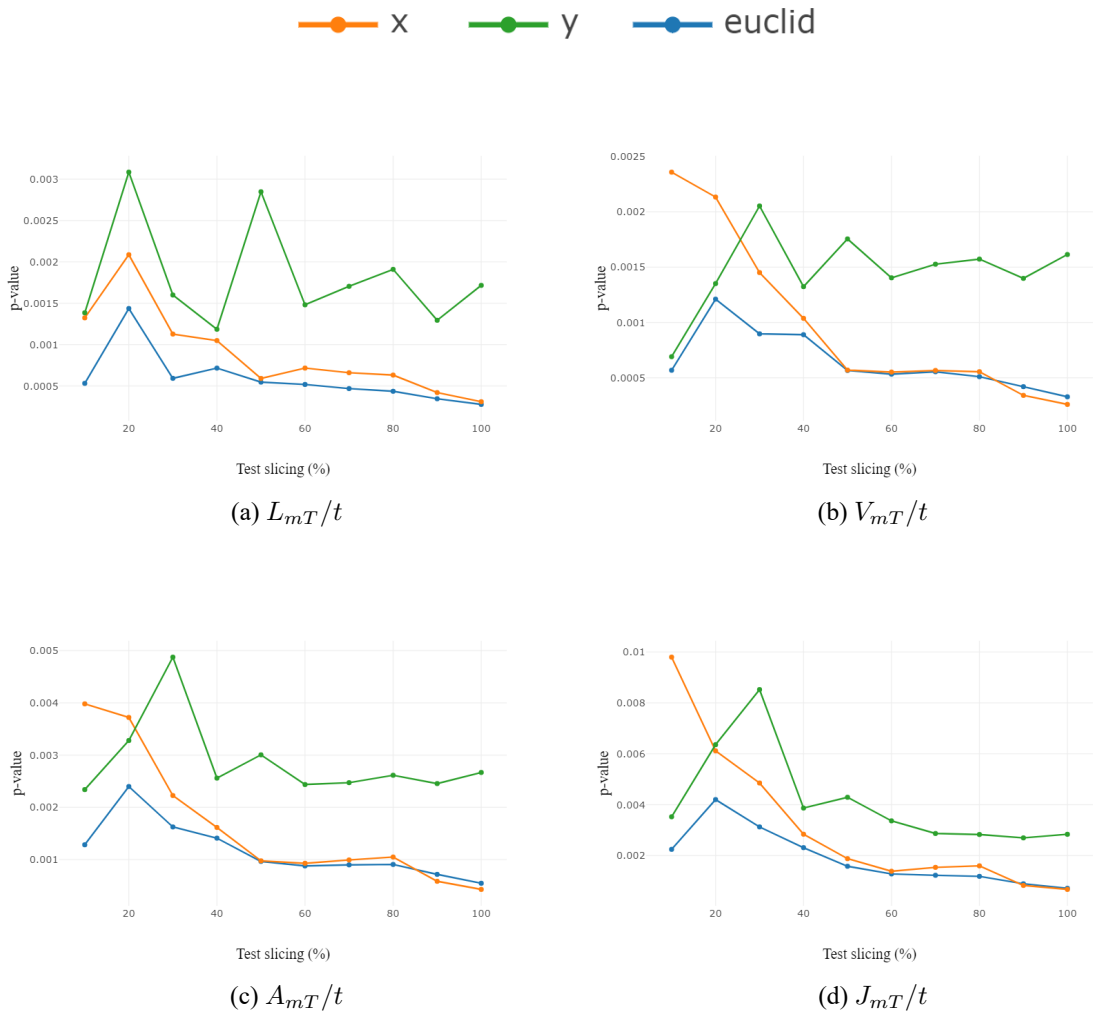
(a) L_{mT}/t (b) V_{mT}/t

(c) A_{mT}/t (d) J_{mT}/t

Figure 11. Fisher's scores of x and e variants of relative motion mass features for accumulated slicing in *pcontinue*

With that notion in mind, it was observed that if the slicing was chosen to be less than 50%, the Euclidean variant had a larger Fisher's score, but in the case that the slicing at hand was more than 50%, the variation with the horizontal component had better a larger Fisher's score. An example of the Fisher's scores in such a case can be seen on Figure 11.

Features that were not part of the the main temporally relative motion mass parameters were L_{Ty}, NCP and t. They had an FS_jmax of 0.53, 0.49 and 0.59, respectively. Furthermore, the scores of t and L_{Ty} decreased quickly with larger test slicings, down to 0.35 and 0.1. This means the temporally relative motion mass features are the most useful for classification in test type *pcontinue*.



(a) L_{mT}/t (b) V_{mT}/t

(c) A_{mT}/t (d) J_{mT}/t

Figure 12. P-values of x, y and e variants of relative motion mass features for single slicing in *pcontinue*

The features further excluded were L_{Ty}/t, V_{Ty}/t, A_{Ty}/t and J_{Ty}/t. Analysing the vertical part separately proved to not provide a lot of additional information for classification in this test type, as can be seen in Appendix 3 Table 3. Though the y variants of the features proved

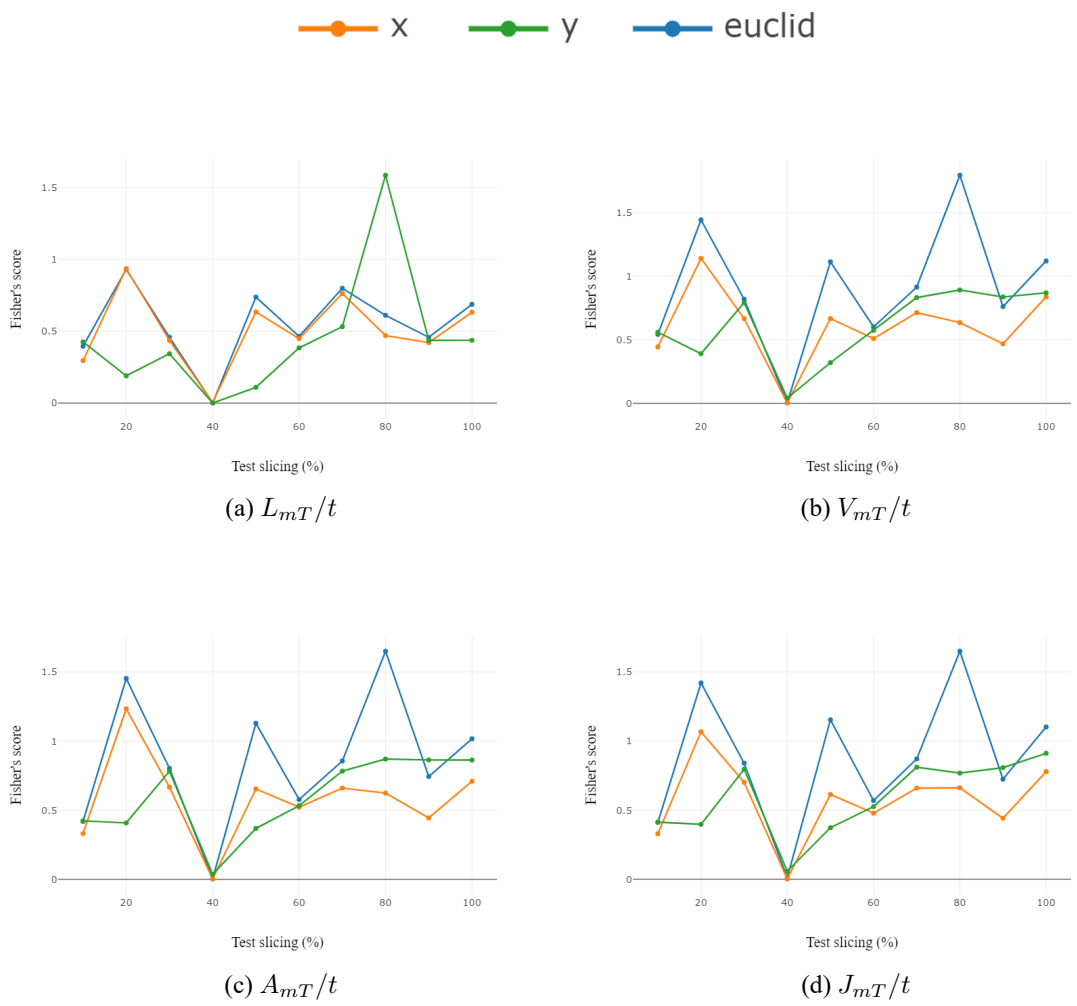
to be statistically significant, e and x variants had on average a Fisher's score 0.3 higher at a 60% slicing than that of the y variant. In addition, the slicing percent of the test had little effect for the removed features making features insignificant when looking at the optimal length of the test.

P-value analysis on the excluded parts was in alignment with Fisher's score analysis, along with the analysis on the Fisher's scores as can be seen on Figure 12 and in Appendix 3 Table 4. Two optimal slicing lengths are proposed based on the local maximum on the 60% and the global maximum on the 100% slicing. The 60% slicing optimal featureset chosen for *pcontinue* is $\{L_{Te}/t, V_{Tx}/t, A_{Tx}/t\}$. The 100% slicing optimal featureset chosen is $\{L_{Tx}/t, V_{Tx}/t, A_{Tx}/t\}$. The mean Fisher's score of selected features on the 60% slicing was 0.941. The mean Fisher's score of selected features on the 100% slicing was 1.10. So if the 60% slicing was used, it would lose on average 0.159 of Fisher's score per feature. The respective means for p-values are 0.00067 for 60% and 0.00033 for 100%.

5.2.2 Plcontinue

As the test itself was the same – continuing a given pattern – not a lot change was expected for the single slicing in the data of features from *pcontinue*. However, the results for the temporally relative variants of the four main motion mass parameters were a lot more volatile than in *pcontinue*, as can be observed on Figure 13 and in Appendix 4 Table 1. The volatility was also mirrored in the p-values, seen in Appendix 4 Table 2. The peaks of all other features included in the single slicing of *plcontinue* again remained below 0.6, with the exception of D_T and P_T, with the respective peaks of 0.778 and 0.601. Thus when selecting features they cannot compete with the several peaks on Figure 14.

The optimal single slicing lengths chosen for *plcontinue* are based on the local extrema on the 20% slicing and the global extrema on the 80% slicing. As the e components have more favourable Fisher's scores and p-values, the optimal featureset chosen for both lengths is again $\{V_{Te}/t, A_{Te}/t, J_{Te}/t\}$. For the 20% slicing, the mean Fisher's score of the features is 1.44. For the 80% slicing, the mean Fisher's score is 1.70. The respective means for p-values are 0.00071 and 0.00022.



(a) L_{mT}/t (b) V_{mT}/t

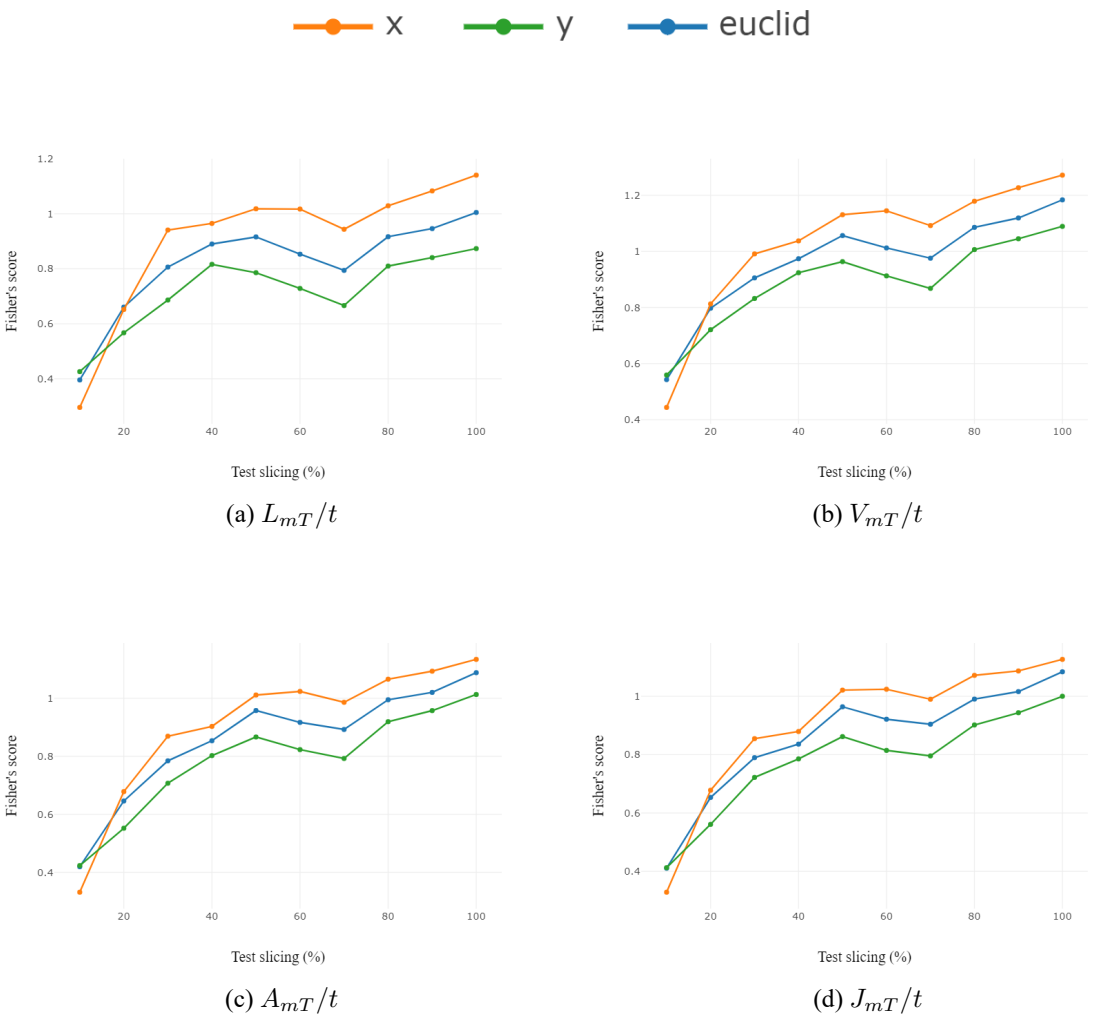
(c) A_{mT}/t (d) J_{mT}/t

Figure 13. Fisher's scores of x, y and e variants of relative motion mass features for single slicing in *plcontinue*

Deviation from the results of *pcontinue* can also be seen for accumulated slicing, on the motion mass paramaters Figure 14 and in Appendix 4 Table 3, the y variants of the temporally relative motion mass features improved by a large margin, as opposed to its results in *pcontinue* not improving over slices and its Fisher's scores staying under 0.6. This improvement also impacted the relation between x and e variants. For the *plcontinue* test, every slice above 20% favours x variants. This is also confirmed by the p-values seen in Appendix 4 Table 4.

Four previously included features – L_{Ty}, L_{Te}, P_T and t – were immediately excluded on the basis of low initial FS_fmax and a quick downward trend. Two features showed promise: D_T and NCP, shown in Figure 15. If there were a bigger dataset of patients and thus more features could be used in classifiers, these features would definitely be worth including in order to provide an alternate dimension to the features exploring length and its derivatives.

Two optimal slicing lengths are proposed for accumulated slicing based on the local maximum

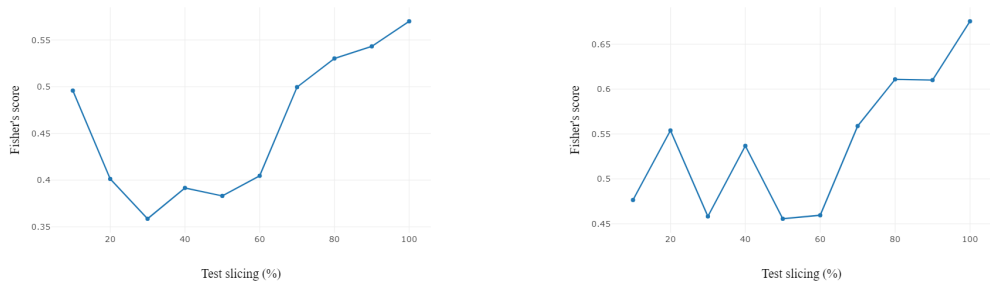


(a) L_{mT}/t (b) V_{mT}/t

(c) A_{mT}/t (d) J_{mT}/t

Figure 14. Fisher's scores of x, y and e variants of relative motion mass features for accumulated slicing in *plcontinue*

on the 50% and the global maximum on the 100% slicing. The 50% slicing optimal featureset chosen for *plcontinue* is $\{L_{Te}/t, V_{Tx}/t, J_{Tx}/t\}$. The 100% slicing optimal featureset chosen is $\{L_{Tx}/t, V_{Tx}/t, A_{Tx}/t\}$. The mean Fisher's score of selected features on the 50% slicing



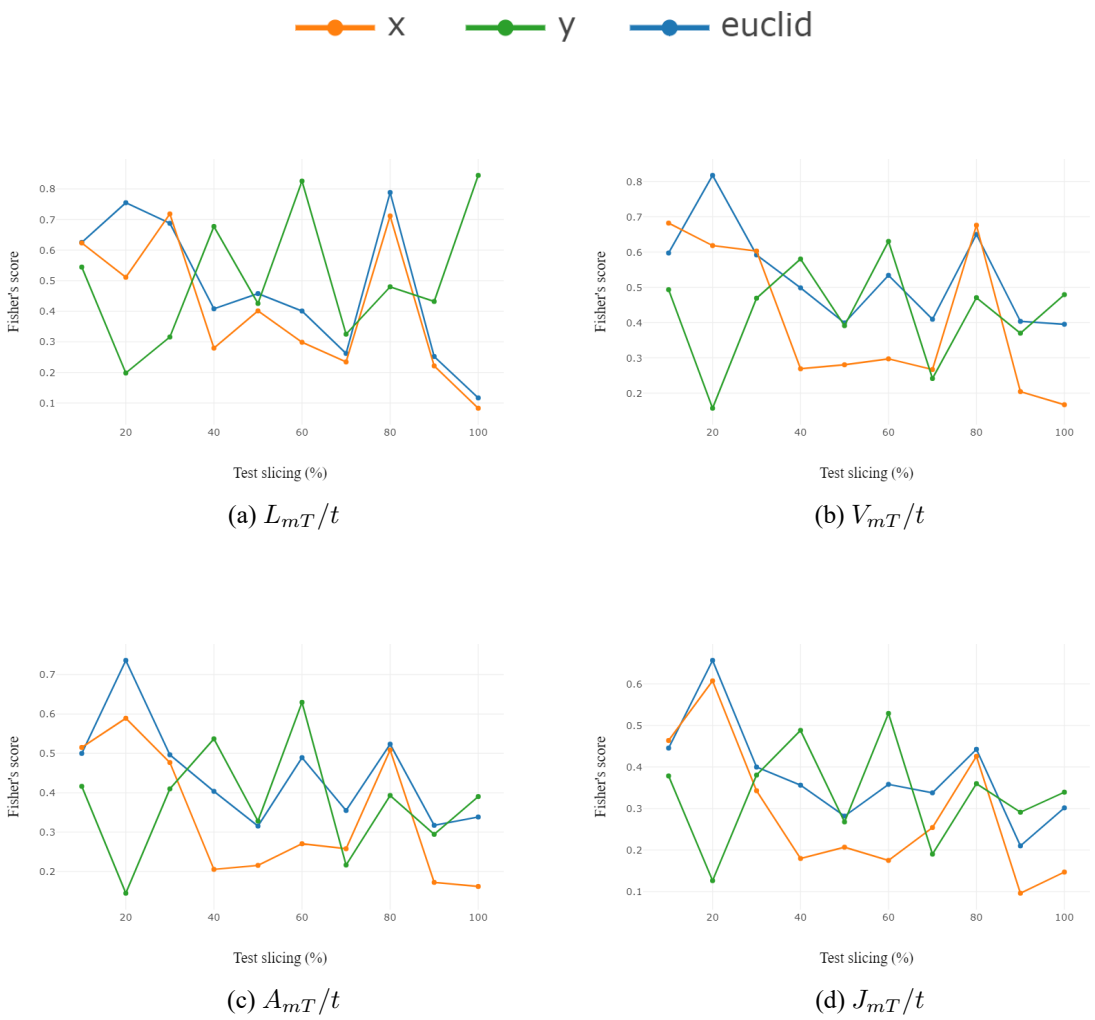
(a) D_T (b) NCP

Figure 15. Fisher's scores of features D_T and NCP in *plcontinue*

was 1.06. The mean Fisher's score of selected features on the 100% slicing was 1.18. The mean p-values are respectively 0.0026 and 0.0017.

5.2.3 Pcopy

For the single slicing of *pcopy*, the Fisher's scores of four main temporally relative motion mass features decreased significantly, as can be seen on Figure 16 and in Appendix 5 Table 1, especially when compared to the previous test type. The Fisher's score peaks of most of the features included were still significantly lower than the main relative motion mass features, though D_T/t had an FS_jmax of 0.84 at the 20% slice, as seen on Figure 17. The findings based on Fisher's scores were also reflected in the p-values, as seen in Appendix 5 Table 2.



(a) L_{mT}/t (b) V_{mT}/t

(c) A_{mT}/t (d) J_{mT}/t

Figure 16. Fisher's scores of x, y and e variants of relative motion mass features for single slicing in *pcopy*

The test lengths proposed for the single slicing variant of *pcopy* are based on the 20% global maximums and the 60% local maximums. The featureset chosen for the 20% global maximum

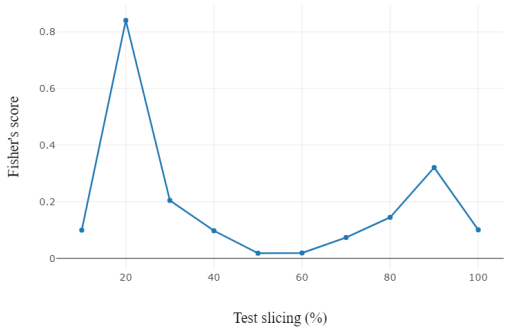
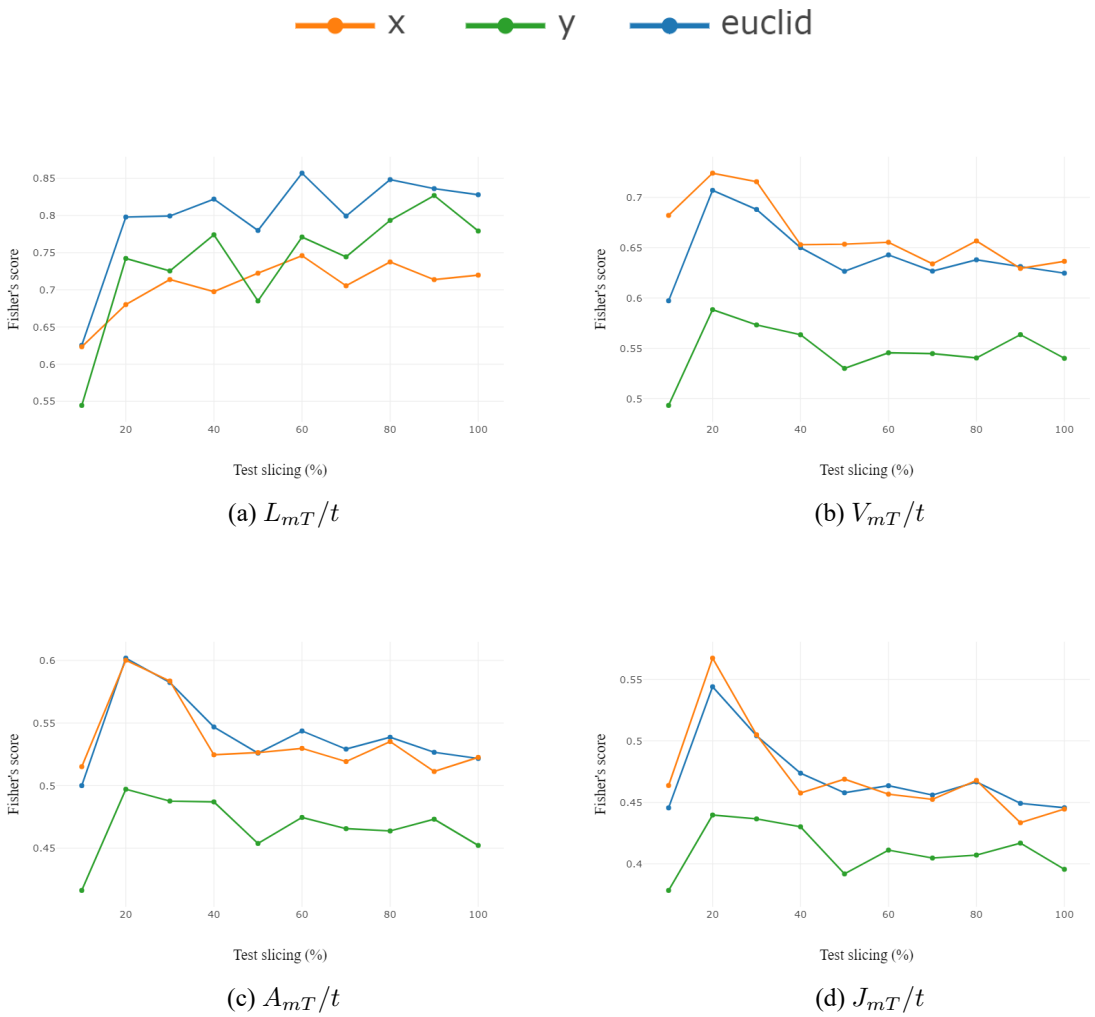


Figure 17. Fisher's scores of D_T/t for single slicing in *pcopy*



(a) L_{mT}/t (b) V_{mT}/t

(c) A_{mT}/t (d) J_{mT}/t

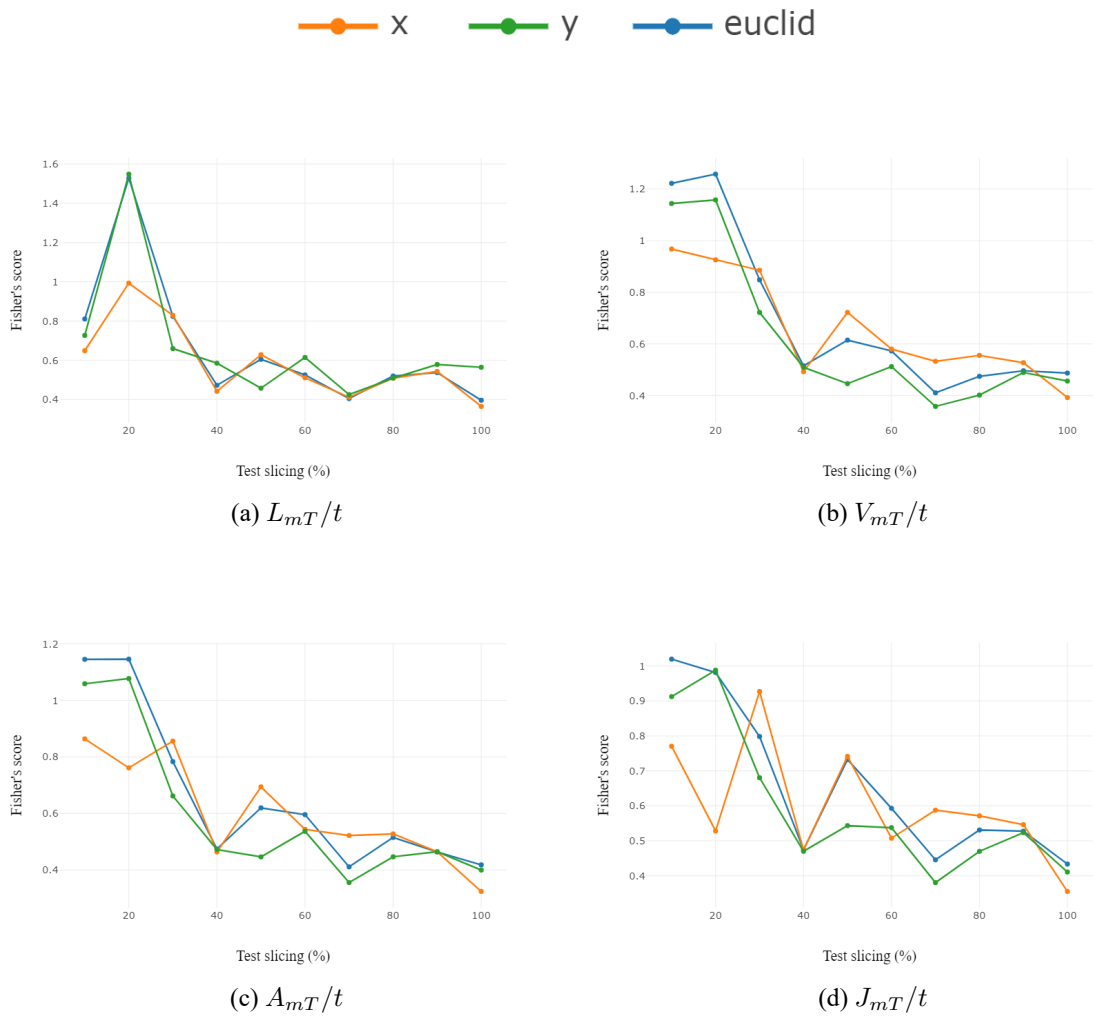
Figure 18. Fisher's scores of x, y and e variants of relative motion mass features for accumulated slicing in *pcopy*

of *pcopy* is $\{L_{Te}/t, V_{Te}/t, D_T/t\}$. The featureset chosen for the 60% local maximum was $\{L_{Ty}/t, V_{Ty}/t, A_{Ty}/t\}$. The mean Fisher's score for the 20% global maximum featureset was 0.80. The mean Fisher's score for the 60% local maximum featureset was 0.70. The respective mean p-values were 0.0009 and 0.0023.

For accumulated slicing, the features of the main relative mass features declined in Fisher's score as well, as can be seen on Figure 18. However, there were no other features contending with the main features, all of them having Fisher's score peaks of less than 0.55, as seen in Appendix 5 Table 3. The same relative results could also be seen in the p-values of the features, displayed in Appendix 5 Table 4.

The optimal lengths based on feature analysis for accumulated slicing of *pcopy* are based on the maximums of 20% and 60% slicings. The best featureset for both turns out to be the same, both based on p-values and Fisher's scores: $\{L_{Te}/t, V_{Tx}/t, A_{Te}/t\}$. The mean Fisher's score of the 20% slice featureset on the 20% slice is 0.71. The mean Fisher's score of the 80% slice featureset on the 80% slice is 0.68. The respective p-value means are 0.0019 and 0.0021.

5.2.4 Plcopy

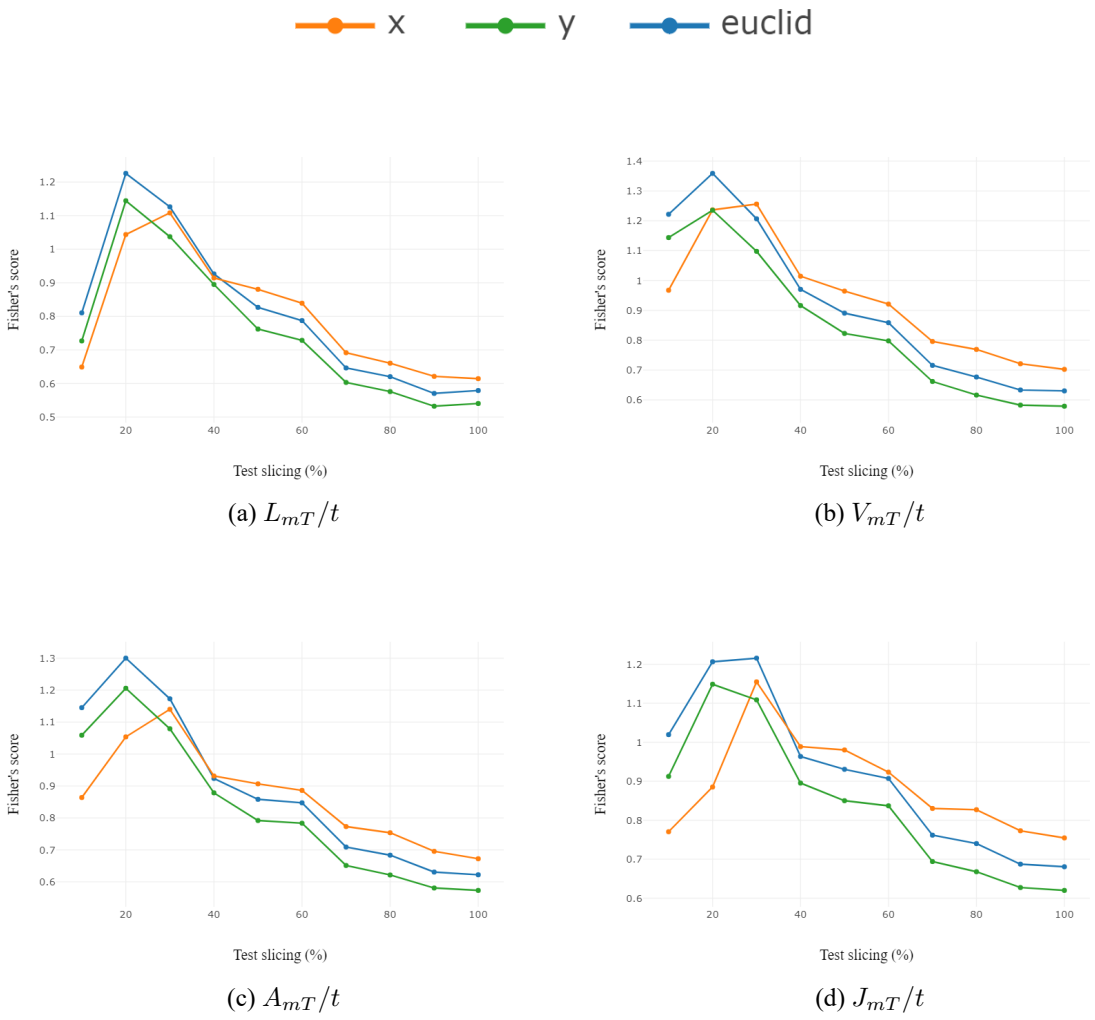


(a) L_{mT}/t (b) V_{mT}/t

(c) A_{mT}/t (d) J_{mT}/t

Figure 19. Fisher's scores of x, y and e variants of relative motion mass features for single slicing in *plcopy*

For the single slicing of *plcopy*, a maximum amount of features were excluded in the initial filtering process. Out of the 13 remaining features, 12 were the main four temporally relative motion mass feature sets, with their Fisher's scores shown on Figure 19. The last feature selected was NCP, though it had an FS_jmax of just 0.53, as can be seen in Appendix 6 Table 3. This means its classification power is relatively low compared to the main relative motion mass features. The same is true for p-values, as seen in Appendix 6 Table 2.



(a) L_{mT}/t (b) V_{mT}/t

(c) A_{mT}/t (d) J_{mT}/t

Figure 20. Fisher's scores of x, y and e variants of relative motion mass features for accumulated slicing in *plcopy*

The test lengths proposed based on feature analysis for single slicing are based on the global maximums on the 20% slice and the local maximums on the 50% slice. The optimal feature-set chosen for the 20% global maximum is $\{L_{Ty}/t, V_{Te}/t, A_{Te}/t\}$. The optimal featureset chosen for the 50% slice is $\{V_{Tx}/t, A_{Tx}/t, J_{Tx}/t\}$. The mean Fisher's score for the feature-set optimal for the 20% slice is 0.68. The mean Fisher's score for the featureset optimal for the 50% slice is 0.55. The respective p-values are 0.0019 and 0.0043.

For accumulated slicing of *plcopy*, the same downward trend exists as did for single slicing, which can be observed on Figure 20 and in Appendix 6 Table 3. Due to the nature of accumulated slicing, it can be observed that the peaks of single slicing are more evened out, and in the cases of V_{mT}, A_{mT}/t and J_{mT}/t, has even higher peaks than the single slicing. This is not just with Fisher's scores as well, as it can be observed from Appendix 6 Table 4 that the same has happened with p-values.

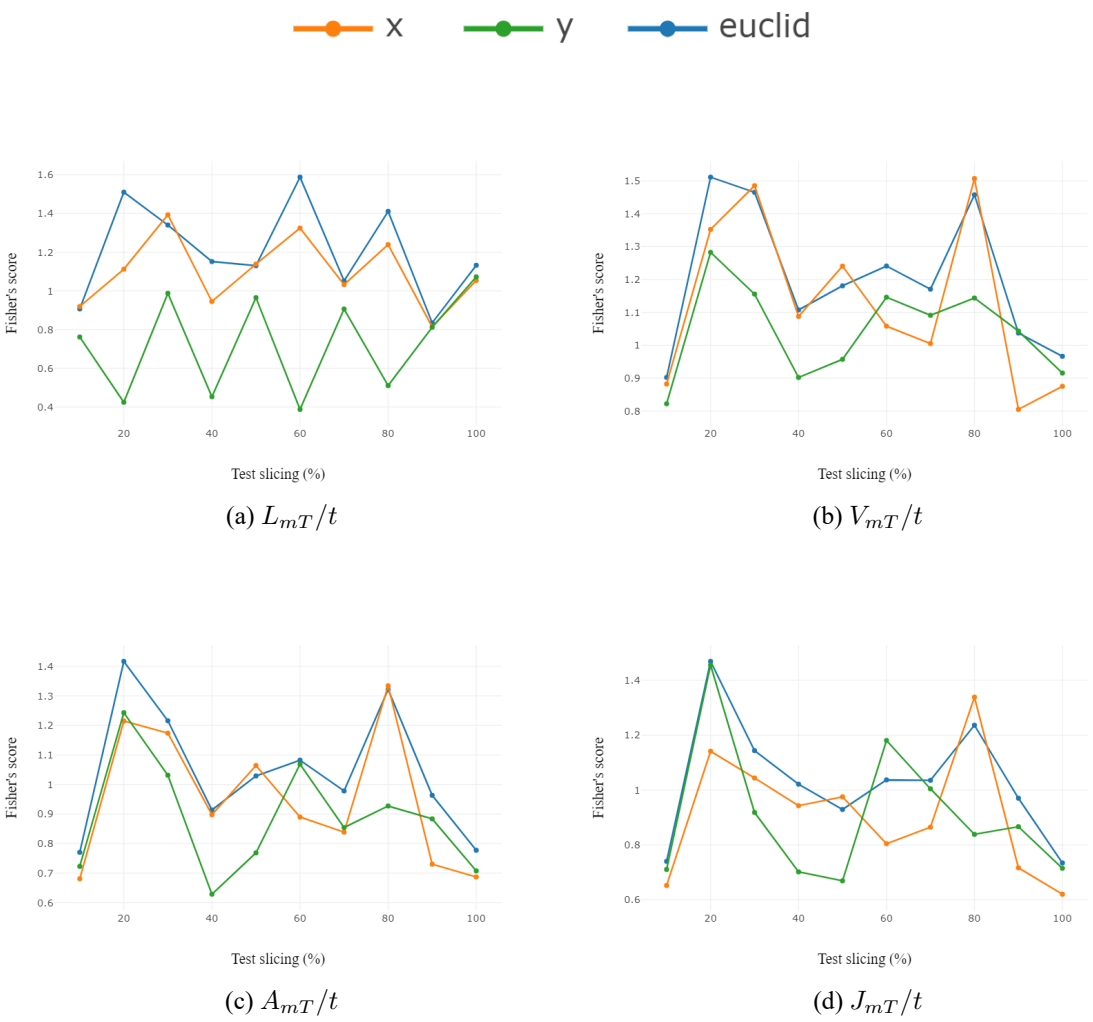
The optimal test lengths based on feature analysis for accumulated slicing in *pcopy* are based on the global maximums in the 20% slice and local maximums in the 60% slice. The featureset chosen for the 20% global maximums is $\{L_{Te}/t, V_{Te}/t, A_{Te}\}$. The featureset chosen for the 60% local maximums is $\{V_{Tx}/t, A_{Tx}/t, J_{Tx}\}$. The mean Fisher's score for the featureset chosen for 20% is 0.83 in the 20% accumulated slice. The mean Fisher's score for the featureset chosen for the 60% slice is 0.91 in the 60% accumulated slice. The respective mean p-values are 0.0018 and 0.0009.

5.2.5 Ptrace

For single slices of *ptrace*, two features were included in addition to the main four temporally relative motion mass feature sets: D_T and NCP. It can be observed on Figure 21 that the four main temporally relative motion mass features are relatively stable, meaning that all of the single slicings of the test perform well enough for use in classification.

The two optimal test lengths based on feature analysis for single slicing in *ptrace* are based on the global and local maximums of 20% and the local maximums of 80%. As can be seen from Appendix 7 Table 1 and Appendix 7 Table 2, the Fisher's scores and p-values have different lists of the best three features, though the differences are not huge. As a compromise between the two lists, the 20% will include two of the best features from the p-value list – L_{Te}/t and J_{Ty}/t, and two of the best features from the Fisher's score list – L_{Te}/t and V_{Te}/t. Thus the featureset for the 20% slice is $\{L_{Te}/t, V_{Te}/t, J_{Ty}/t\}$.

Regarding the featureset for chosen slice of 80%, the top three best features were the same. The featureset for the 80% slice is $\{L_{Te}/t, V_{Tx}/t, A_{Tx}/t\}$. The mean Fisher's score for the featureset assigned to the 20% slice is 1.42. The mean Fisher's score for the featureset



(a) L_{mT}/t (b) V_{mT}/t

(c) A_{mT}/t (d) J_{mT}/t

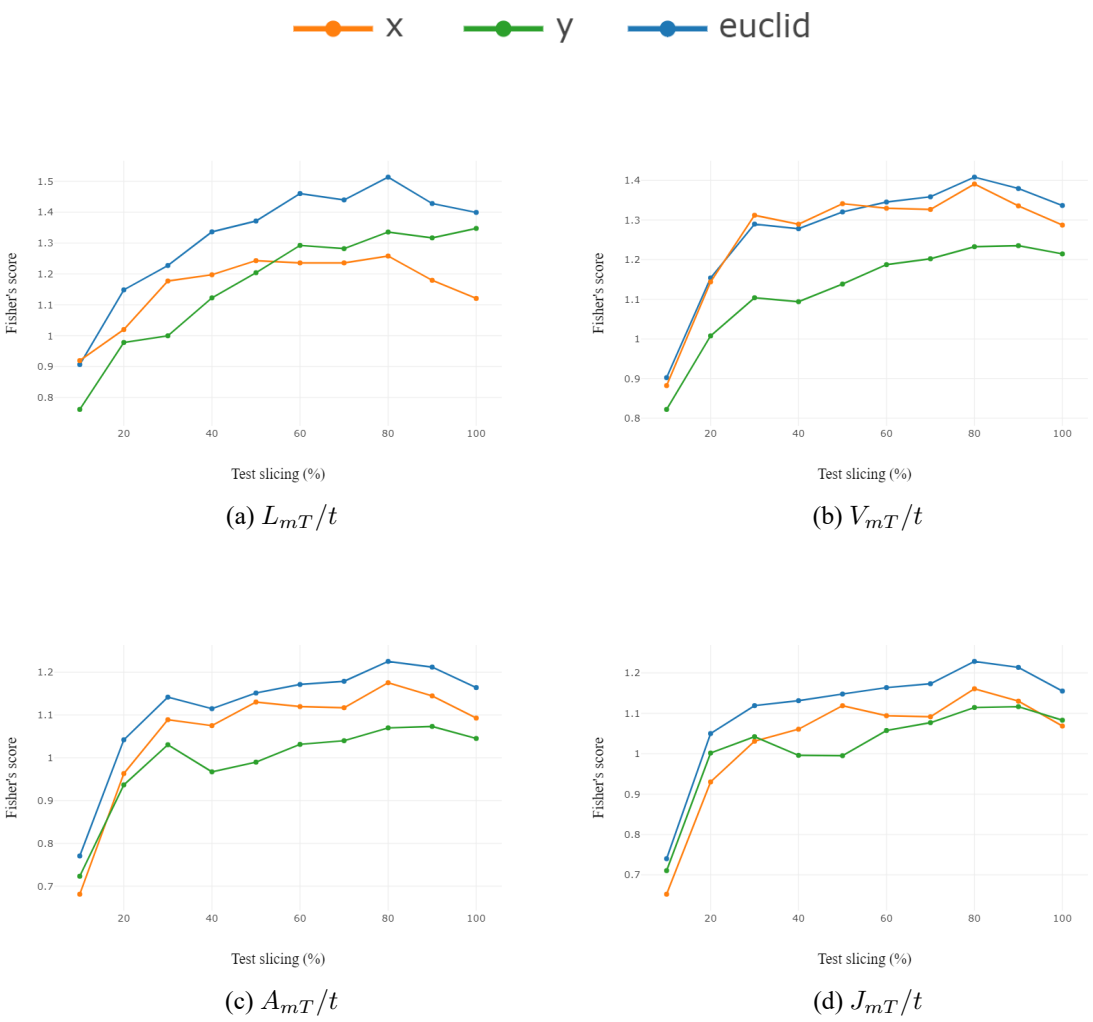
Figure 21. Fisher's scores of x, y and e variants of relative motion mass features for single slicing in *ptrace*

assigned to the 80% slice is 1.36. The respective mean p-values are 0.00007 and 0.00012.

For accumulated slices of *ptrace*, the same two features were included along the main temporally relative motion mass features: D_T and NCP. As they had an FS_jmax of 1.050 and 0.866 respectively, they are statistically significant features. However, as can be seen on Figure 22 and in Appendix 7 Table 3, the temporally relative motion mass features offered significantly higher Fisher's scores.

The values of the Fisher's scores being relatively constant and on the upwards trend until the 90% slice showed that a longer test does not offer a significantly increased amount of discriminating power.

Two optimal slicing lengths for accumulated slicings are proposed based on the local maximum on the 30% and the global maximum on the 80% slicing. For the 30% slicing, Fisher's



(a) L_{mT}/t (b) V_{mT}/t

(c) A_{mT}/t (d) J_{mT}/t

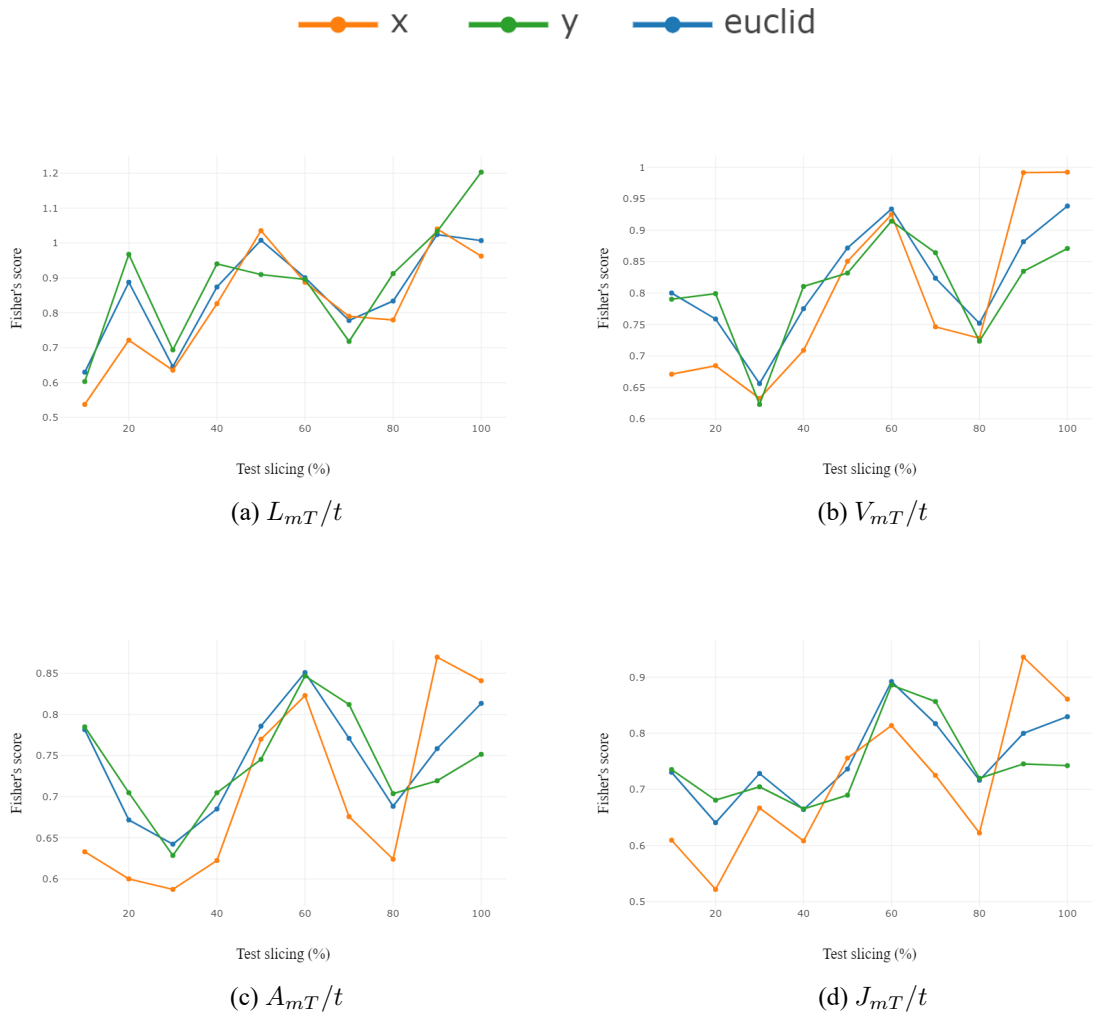
Figure 22. Fisher's scores of x, y and e variants of relative motion mass features for accumulated slicing in *ptrace*

scoring attaches a relatively high value to L_{Te}/t, V_{Tx}/t and A_{Te}/t, and attaches a relatively low value to D_T and NCP, as can be seen in Appendix 7 Table 4. However, p-values for D_T and NCP are very high and relatively low values to L_{Te}/t, V_{Te}/t and A_{Te}/t. Thus, a compromise will be made for choosing the optimal featureset for the 30% slicing, choosing the optimal featureset to be $\{L_{Te}/t, V_{Tx}/t, D_T\}$.

For the 80% slicing Fisher's scoring and p-values gave the same best three features. Thus the 80% slicing optimal featureset chosen is $\{L_{Te}/t, V_{Te}/t, J_{Te}/t\}$. The mean Fisher's score of selected features on the 30% slicing was 1.23. The mean Fisher's score of selected features on the 80% slicing was 1.38. The respective mean p-values were 0.00015 and 0.00011.

5.2.6 Pltrace

For single slicing on *pltrace*, initial feature filtering included, in addition to the four main temporally relative motion mass featuresets, D_T, D_T/t, NCP and t. However, as can be observed in Appendix 8 Table 1, the additional features are either relatively low in Fisher's score or fluctuate too severely. As can be seen from 23, the Fisher's scores of the features are relatively stable in this test type as well, meaning any of the selected slices is suitable for classification tasks.



(a) L_{mT}/t (b) V_{mT}/t

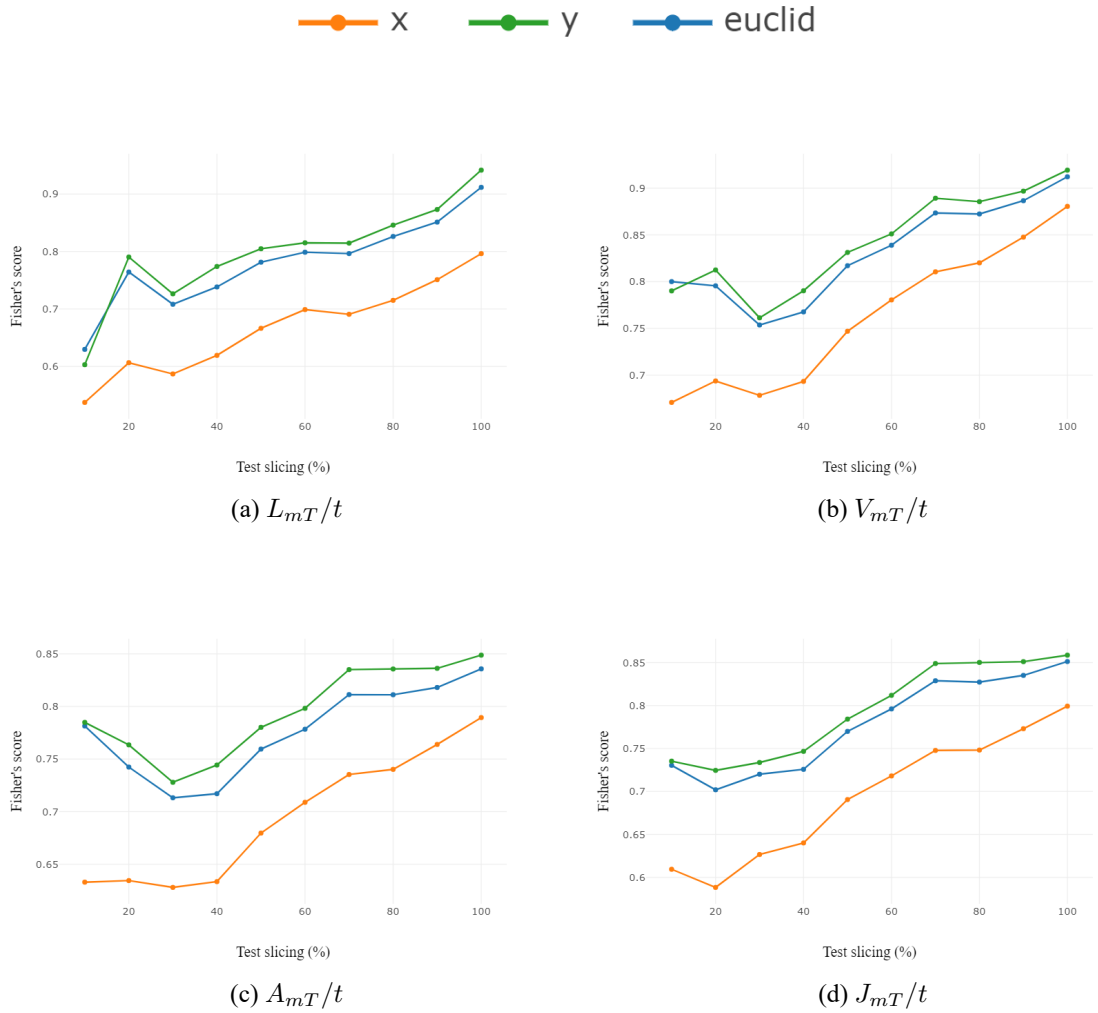
(c) A_{mT}/t (d) J_{mT}/t

Figure 23. Fisher's scores of x, y and e variants of relative motion mass features for single slicing in *pltrace*

For single slicing, the optimal lengths proposed are based on the maximums on the 60% slice and the maximums on the 90% slice. The featureset decided for the 60% slicing is $\{L_{Te}/t$, V_{Te}/t, $J_{Te}/t\}$. For the 90% slicing, as among p-values the p-value of NCP was the highest, as can be seen from Appendix 8 Table 2, it was decided to incorporate it into the featureset. The next two best features were also in the two best features by Fisher's score ranking. Thus the featureset decided for the 90% slicing is $\{L_{Tx}/t$, V_{Tx}/t, $NCP\}$. The mean Fisher's score for the featureset created for 60% slicing was 0.91. The mean Fisher's score fore the

featureset created for 90% slicing was 0.88. The respective mean p-values were 0.00080 and 0.00045.

For accumulated slicing, in addition to the temporally relative motion mass features, three features were included for *pltrace* in the filtering process: D_T, NCP and t. It can be inferred from Figure 24 and from Appendix 8 Table 3 that all Fisher's scores of the temporally relative motion mass features have lowered significantly in comparison with their values in *pltrace*, seen on Figure 24. Despite this, D_T, NCP and t have a relatively low value as well in this test type, having a FS_jmax of 0.439, 0.550 and 0.424 respectively. What can also be inferred from Figure 24 is that the x variants of the relative motion mass features have relatively a very low discrimination power in this test type, which also reduces the discrimination power of the e variant to below that of the y variant.



(a) L_{mT}/t (b) V_{mT}/t

(c) A_{mT}/t (d) J_{mT}/t

Figure 24. Fisher's scores of x, y and e variants of relative motion mass features for accumulated slicing in *pltrace*

Although for *pltrace*, the four features with the best Fisher's scores were quite clearly L_{Ty}/t,

V_{Ty}/t, A_{Ty}/t and J_{Ty}/t, however the choice of optimal slicing was more difficult to determine due to the lack of clear local and absolute maximums. In the case of the 70% slice, L_{Ty}/t would be left out of the featureset and the mean Fisher's score of the features in the slice would be 0.858. In the case of the 100% slice, A_{Ty}/t would be left out and the mean Fisher's score would be 0.906. The trade-off of 0.048 for making a test 30% shorter was considered to be enough to select the 70% variant.

Two optimal slicing lengths are proposed based on the local maximum on the 10% and the local maximum on the 70% slicing. For both slicings the p-values showed that NCP had relatively a very low p-value of 0.0008, as seen in Appendix 8 Table 4. Thus it was incorporated into the featuresets. The chosen optimal featuresets for both slicings were the same: $\{V_{Ty}/t$, A_{Ty}/t, $NCP\}$. The mean Fisher's score of selected features on the 10% slicing was 0.69. The mean Fisher's score of selected features on the 70% slicing was 0.75. The respective mean p-values were 0.0012 and 0.0009.

Selecting the minimum slicing length as a proposed slicing length seems unintuitive, as it is a very small length compared to the initial test. It also changes the aspects the *pltrace* test was designed to assess. However, data supports the slice's ability to discriminate between PD patients and HCs based on even just a short slice, meaning its selection must be viable.

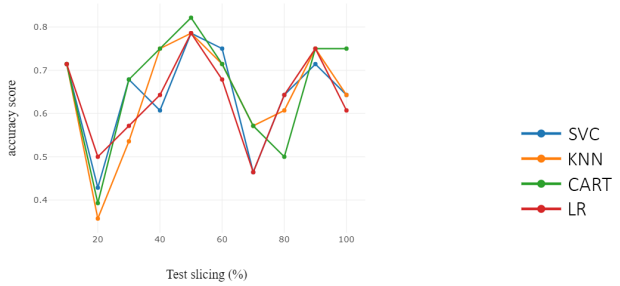
5.3 Classifier analysis

Depending on the test type, a minimum of two and a maximum of four classifiers were created, based on the amount of featuresets selected for the test type. Every featureset was limited to three features in order to avoid overfitting. As stated in the methodology chapter, all classifiers using single slicing featuresets as their featuresets are trained and tested only on single slicing data, and all classifiers using accumulated slicing featuresets as their featuresets are trained and tested only on accumulated slicing data.

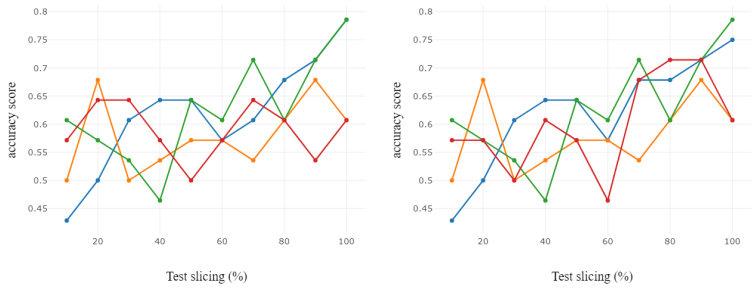
5.3.1 Pcontinue

For *pcontinue*, Figure 25 and Appendix 9 reveal that both of the accumulated featuresets provided quite similar results in terms of accuracy score and its change through different

slicings. It was observed that these featuresets provided a relatively high increase at the 100% slice point. As there was a significant increase at the slice of 70% as well, a test length of 70%-100% was initially considered to be the optimal length for *pcontinue* based on the classifier results.



(a) Featureset optimal for 60% and 100% slices in single slicing



(b) Featureset optimal for 60% slice in accumulated slicing

(c) Featureset optimal for 100% slice in accumulated slicing

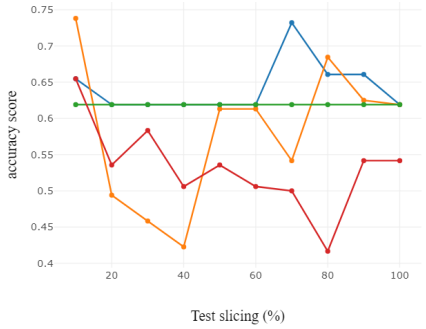
Figure 25. Classifiers trained for *pcontinue* with the chosen featuresets

The single slicing classifier, however, showed that analysing just the 50% single slice yielded an improved accuracy score over the 100% accumulated slice, in the case of both classifiers. Based on the three classifier results, a test length of 40-70% of the initial test length is proposed to minimise length and maintain similar classification power.

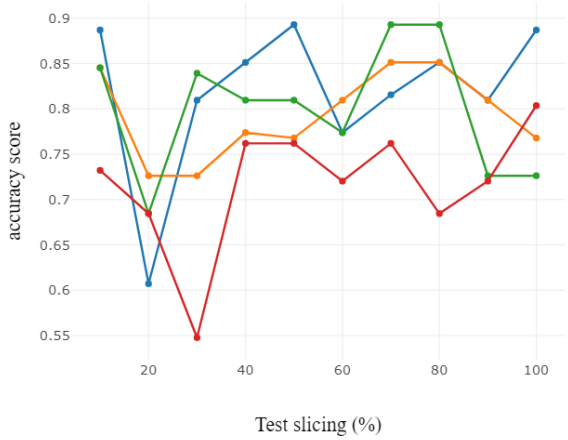
5.3.2 Plcontinue

For *plcontinue*, especially when looking at Figure 26c and Appendix 9, it is apparent that there were classifiers based on accumulated slicing performing at the same accuracy level on 10%, 50%, 70% and 80% as is the accuracy level of the 100% slice. Even on Figure 26b the 100% slice is outperformed by earlier slices, even if the accuracy scores are relatively low. For the single slicing classifier, the 20% and 30% slicings had nearly the same accuracy score

(a) Featureset optimal for 20% and 80% slices in single slicing



(b) Featureset optimal for 50% slice in single slicing



(c) Featureset optimal for 100% slice in single slicing

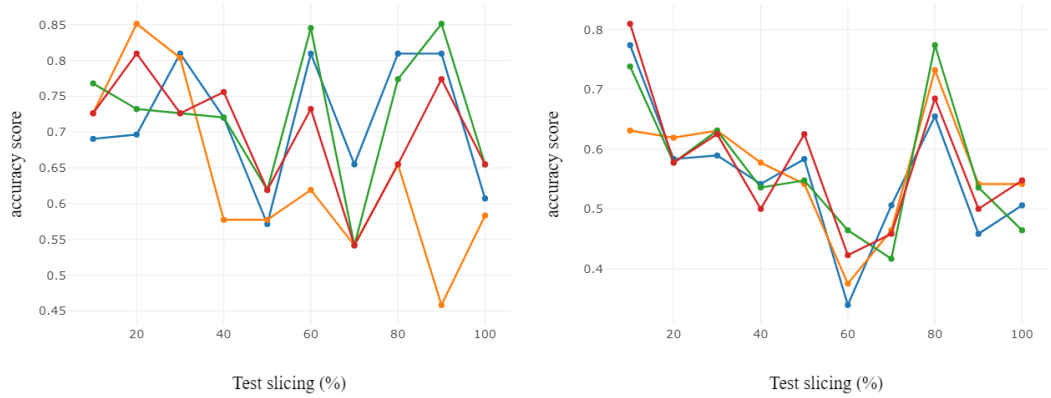
Figure 26. Classifiers trained for *plcontinue* with the chosen featuresets

as the 100% slice as well.

Based on the results of the three classifiers, it is safe to say that a test length of about 30-50% of the initial test length is viable for *plcontinue* without losing a significant amount of classification power.

5.3.3 Pcopy

Again, it can be seen on Figure 27 that the accuracy scores of the 10% and 20% slices are nearly as good as the best accuracy scores in the range of 70-100% slices. This can be observed both with single slicing and accumulated slicing. In addition, in the case of the 20% single slicing featureset, the there are a total of five slices that nearly the maximum more than



(a) Featureset optimal for 20% slice in single slicing(b) Featureset optimal for 60% slice in single slicing

(c) Featureset optimal for 20% and 60% slices in accumulated slicing

Figure 27. Classifiers trained for *pcopy* with the chosen featuresets

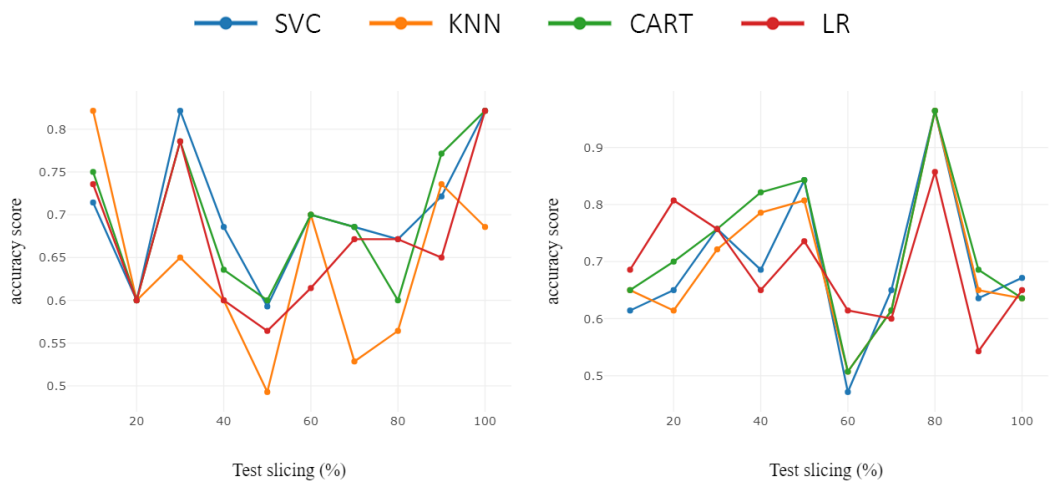
twice.

As all three classifiers showed that half or even less than half of the test can be analysed without losing a significant amount of accuracy, the optimal test length based on the trained classifiers is chosen to be 30-50% of the initial test length.

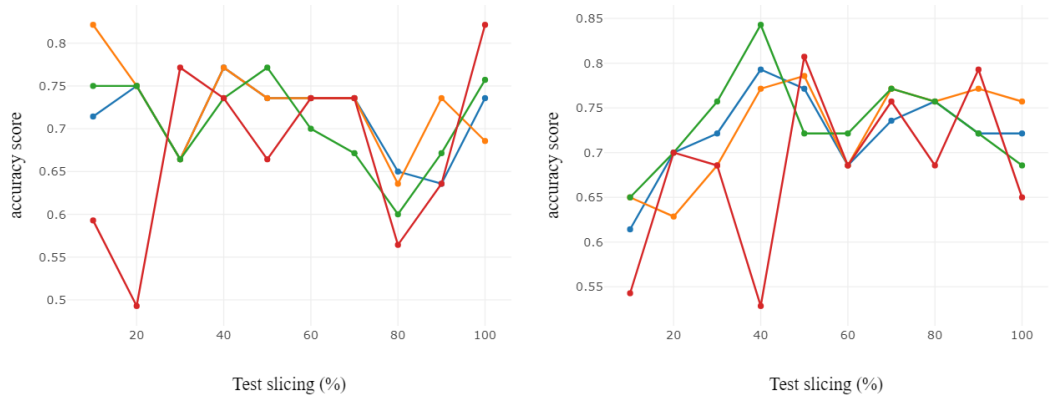
5.3.4 Plcopy

In the four featuresets used when training classifiers for *plcopy*, all but the second, i.e. the 50% featureset for single slicing, had no significant difference between the maximum accuracy scores in the first half and the maximum accuracy scores in the second half, as can be observed on Figure 28. However, a counter-example to the peak on the second featureset was the significant difference on the 60% featureset for accumulated slicing between the maximum accuracy score of the first half and the second half, in which the first half had a higher

maximum accuracy score.



(a) Featureset optimal for 20% slice in single slicing(b) Featureset optimal for 50% slices in single slicing



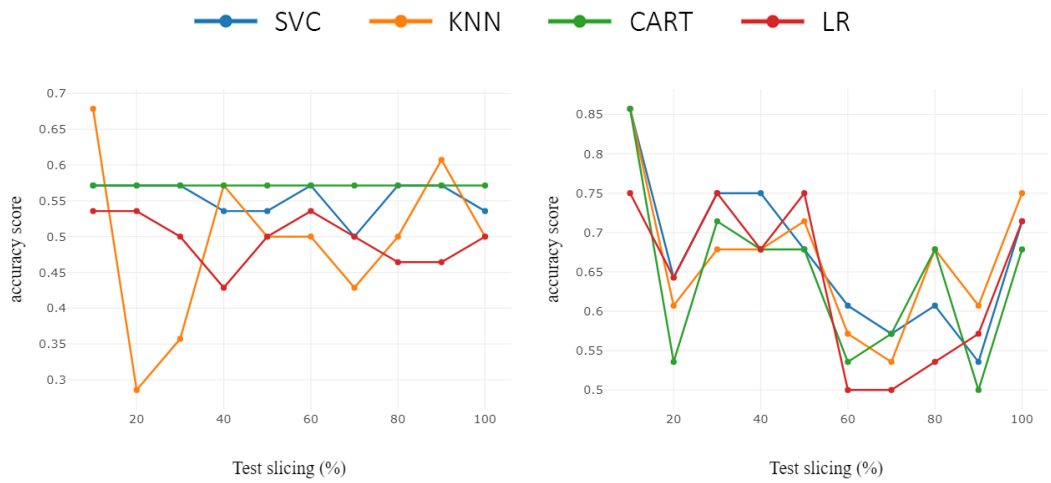
(c) Featureset optimal for 20% slice in accumulated slicing(d) Featureset optimal for 60% slice in accumulated slicing

Figure 28. Classifiers trained for *plcopy* with the chosen featuresets

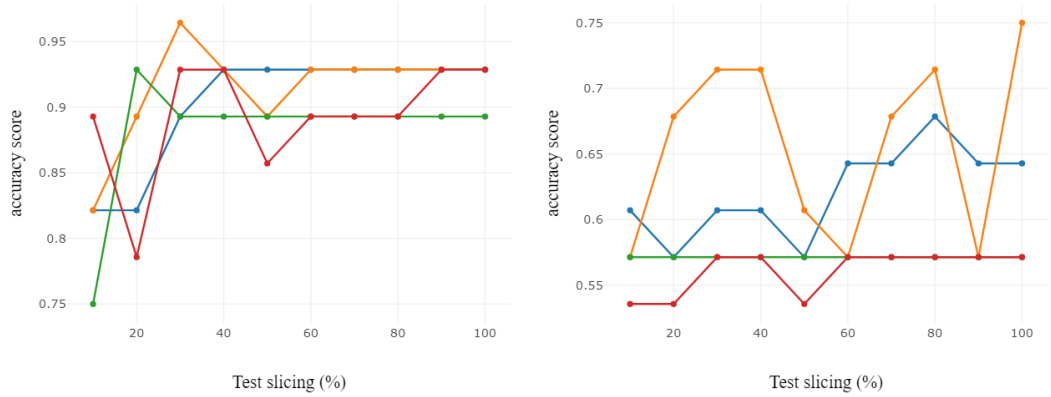
The classifiers trained exhibited that the maximum accuracy scores of a featureset did not improve enough to warrant a higher test length than half the current test length. Thus the suggested test length based on classifier is chosen as 30-50% of the initial *plcopy* test.

5.3.5 Ptrace

As can be seen on Figure 29, again most of the featuresets provide slices in the first half that have classifiers with an accuracy score higher than that of the ones in the second half. The only classifier among the featureset not in accordance with this is the 80% accumulated slicing featureset, which is not significant in the terms that were set, i.e. the difference is less



(a) Featureset optimal for 20% slice in single slicing(b) Featureset optimal for 80% slice in single slicing



(c) Featureset optimal for 30% slice in accumulated slicing(d) Featureset optimal for 80% slice in accumulated slicing

Figure 29. Classifiers trained for *ptrace* with the chosen featuresets

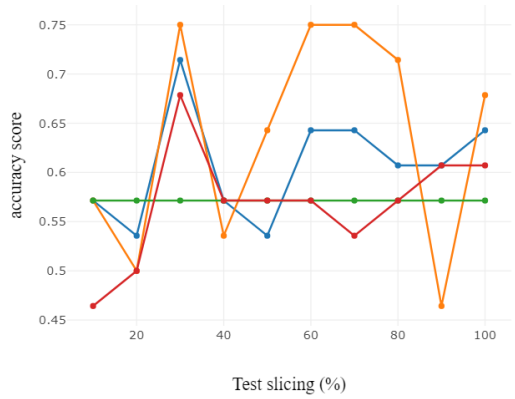
than 0.05 in accuracy score.

From the *ptrace* classifier graphs and Appendix 9 it can be observed that accuracy scores of classifiers of most of the featuresets suggest that a test as short as 40-60% can have nearly as good of a classification power as analysing the whole test. Thus an optimal test length of 40-60% of the initial test length is proposed for *ptrace* based on the classifier results.

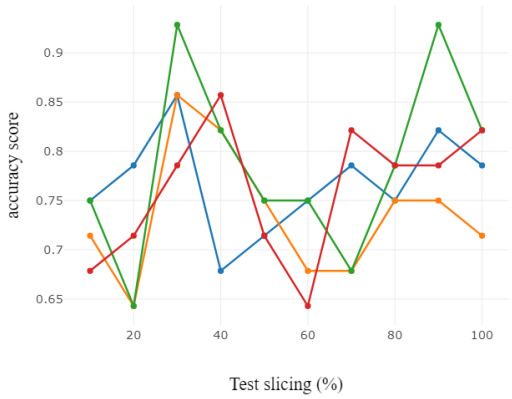
5.3.6 Pltrace

The scores of the classifiers trained for the *pltrace* test type, which can be seen on Figure 30, were similar to previously seen scores. Both featuresets coming from single slicing had a maximum or near-maximum at the 30% slice, displaying yet again that there existed classifiers in the first half of the test that either were maximums or were not significantly lower

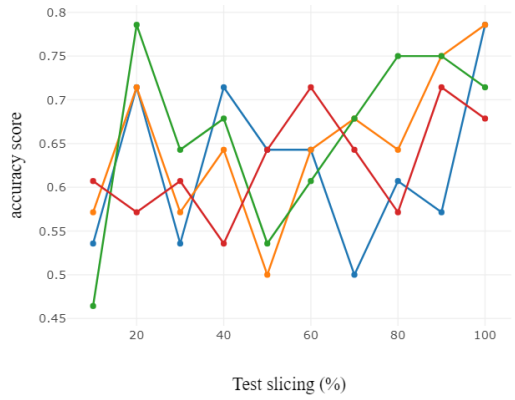
than maximums. The same was observed in the case of the featureset for accumulated slicing. As the accuracy scores of the classificators using the single slicing featuresets dropped significantly by the slices of 40% and 50%, and there is a classificator using the accumulated slicing featureset for every slice which performs favourably, a test length of 30-50% of the initial length of the test was considered to be optimal.



(a) Featureset optimal for 60% slice in single slicing



(b) Featureset optimal for 90% slice in single slicing



(c) Featureset optimal for 10% and 70% slices in accumulated slicing

Figure 30. Classifiers trained for *pltrace* with the chosen featuresets

5.4 Optimal lengths

The shorter optimal lengths for *pcontinue* through feature analysis came out to be the 50% slice for single slicing and the 60% slice for accumulated slicing. Classifier analysis showed 40-70% as the optimal result. As the analyses aligned, the final optimal length range chosen

for *pcontinue* was 40-70%.

For *plcontinue*, in feature analysis the optimal slicing lengths chosen were 20% and 80% for single slicing and 50% and 100% for accumulated slicing. However, there was also a smaller peak at 50% for the better scoring features when single slicing. Classifier analysis showed a test length of 30-50% to be optimal. Therefore analysises align quite well. Therefore a total optimal length of 30-50% of the initial test was considered best.

In the case of *pcopy*, the optimal slicing lengths chosen in the process of feature analysis were 20% and 60% for both single slicing and accumulated slicing. In classifier analysis a test length of 30-50% was chosen to be optimal. As the 20% length had both higher mean Fisher's scores and lower mean p-values in both single slicing and accumulated slicing, 60% is not included in the test length. Thus the optimal test length is chosen to be 30-50%.

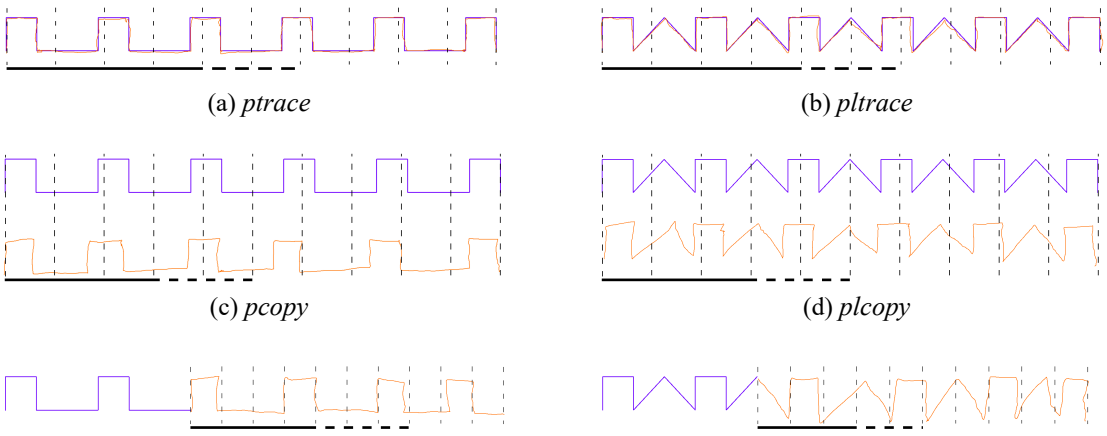
For *plcopy*, the optimal single slicing lengths chosen were 20% and 50%. The optimal accumulated slicing lengths chosen were 20% and 60%. In both cases, 20% was the global maximum. Classifier analysis found the optimal length of the test to be 30-50% of the test. The optimal test length range was hence selected to be 30-50% for *plcopy*.

Feature analysis determined the optimal lengths for the *ptrace* test type to be 20% and 80% for single slicing, and 30% and 80% for accumulated slicing. Classifier analysis considered the a test length of 40-60% to have enough classification power. As there is a significant peak in the case of 60% in both single and accumulated slicings, 60% is included in the final length range. As the lower end of the test length decided by classifier analysis is in alignment with the lengths proposed in feature analysis, the final optimal length range chosen for *ptrace* is 40-60%.

Pltrace's feature analysis selected the optimal lengths to be 60% and 90% in single slicing, and 10% and 70% in accumulated slicing as optimal test lengths. Classifier analysis found the slicing of 30%-50% to be informative enough for optimal classification. However, both in the cases of single and accumulated slicing, the 30% slice has significantly lower feature scores. Thus the final optimal length range chosen for *pltrace* is 40-60% of the initial test length.

The selected optimal test lengths in relation to example tests can be seen on Figure 31. As

pcontinue and *plcontinue* had slightly varying length between the tests taken, example tests with the median of both of them – 4 repetitions – is displayed. The solid line at the bottom indicates the minimum optimal test length and the dotted line represents the area between the minimum and the maximum test length. Thus it can be said that *ptrace* and *pltrace* require 2-4 repetitions, *pcopy*, *plcopy* and *pcontinue* require 2-3 repetitions and *pcontinue* requires 1-2 repetitions.



(a) *ptrace* (b) *pltrace*

(c) *pcopy* (d) *plcopy*

(e) *pcontinue* (f) *plcontinue*

Figure 31. Optimal test lengths for each test type

6 Discussion

When comparing the results of feature analysis and classifier analysis of this thesis to the results of a paper studying the best features from *pcopy* based on a very similar dataset [28], some similarities were found. Even though many of the features were different, for example in the given paper temporally relative features of motion mass – on average the best discerning features in this thesis – and NCP/t were not explored. However, the other features were included. Comparing the Fisher's scores of the results presented in both works – based on analysing the whole test –, it was seen that the presented values of directional mass D_T, NCP and drawing duration t differed by 0.1, 0.04 and 0.005 Fisher's score respectively. The other feature scores were not present in the given paper. The reason the Fisher's scores of the same features are not exactly the same is most likely due to the slight change of the dataset since the paper was presented in 2008, however the calculated features likely share a lot of similarity judging by the three compared Fisher's scores of the same features.

As a pilot research due to a relatively new field and a sample size of just 16 PD patients and 12 HCs, there are several research avenues that the thesis opens. The thesis found that the length of Luria's alternating series tests can be reduced by up to 70% and at least by 30% without losing a significant amount of the ability to diagnose Parkinson's disease. Of course, this analysis can also be carried out on further tests, for example spiral tests, with the aim of reducing their length. Performing this analysis is also greatly facilitated by the developed application, the most major addition required is the need to implement a slicing algorithm suitable for the chosen test, if necessary.

In addition, 31 features were calculated, but only three of them could be used at once in a classifier due to a small sample size. A larger sample size of patients could be used to explore classification models using a larger amount of features. Using more features could also allow to cut the length of tests even further. However, as the patients have to be correctly diagnosed for the test, a large sample size is not easy to obtain.

As the sample size is small, the classifiers will also rarely discern between patients with Parkinson's disease and Parkinson-like symptoms. However, provided the dataset used to train the model is big enough and also has enough controls with Parkinson-like symptoms

who do not have Parkinson's disease, it might be possible to separate Parkinson's disease and the symptoms resembling it.

Finally, the same process could also be used with data from patients having other diseases that display fine motor function symptoms, or even different stages of Parkinson's disease. The developed application is not designed to diagnose specifically Parkinson's disease, but instead show differences in fine motor skills between two pre-selected groups and then categorise test results that have not been classified. Thus, given the data is in the same format, it can also be used when analysing other diseases, and to explore whether it is even possible to diagnose a disease with the use of fine motor tests.

7 Summary

The main aim of the thesis was to determine the optimal lengths of the digitised versions of the six chosen Luria's alternating series tests, with the lengths being as short as possible and the tests maintaining a similar amount of classification power for Parkinson's disease as for the full test length in the original data.

The thesis found that based on the data provided, the length of a test could be reduced by up to 70% for half of the test types, and by at least 30% in all test types, without reducing Fisher's scores, p-values and classification power by a significant amount. This means that as opposed to the 5-7 repetitions of two shapes usually suggested for the regular pen-and-paper variant of Luria's alternating series tests, its digitised version only needs, depending on the test type, 1-4 repetitions.

To facilitate further research with larger datasets and research into other diseases which affect fine motor skills, an application was developed for displaying detailed results of data for a single test and Fisher's score and p-values of data for all slices in a given test type. In addition, the application allows for analysing the classification power of other kinds of digitised fine motor tests as well in the case that the data is in the format, with minor modifications required only in the case that the developed distance slicing algorithm does not apply for the given test type.

References

[1] R. López-Blanco, M. A. Velasco, A. Méndez-Guerrero, J. P. Romero, M. D. del Castillo, J. I. Serrano, E. Rocon, and J. Benito-León, "Smartwatch for the analysis of rest tremor in patients with Parkinson's disease," *Journal of the Neurological Sciences*, vol. 401, pp. 37 – 42, 2019.

[2] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, 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.

[3] P. Sharma, S. Sundaram, M. Sharma, A. Sharma, and D. Gupta, "Diagnosis of Parkinson's disease using modified grey wolf optimization," *Cognitive Systems Research*, vol. 54, pp. 100 – 115, 2019.

[4] L. M. de Lau and M. M. Breteler, "Epidemiology of Parkinson's disease," *The Lancet Neurology*, vol. 5, no. 6, pp. 525 – 535, 2006.

[5] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[6] M. T. Hayes, "Parkinson's disease and parkinsonism," *The American Journal of Medicine*, 2019.

[7] G. DeMaagd and A. Philip, "Parkinson's disease and its management: Part 1: Disease entity, risk factors, pathophysiology, clinical presentation, and diagnosis," *P & T : a peer-reviewed journal for formulary management*, vol. 40, pp. 504–32, 08 2015.

[8] P. Stępień, J. Kawa, D. Wieczorek, M. Dąbrowska, J. Sławek, and E. J. Sitek, "Computer aided feature extraction in the paper version of Luria's alternating series test in progressive supranuclear palsy," in *Information Technology in Biomedicine* (E. Pietka, P. Badura, J. Kawa, and W. Wieclawek, eds.), (Cham), pp. 561–570, Springer International Publishing, 2019.

[9] A. R. Luria, *Higher Cortical Functions in Man*. Springer, 1995.

[10] Z. Lu and K.-H. Yuan, “Welch’s t-test,” in *Encyclopedia of Research Design* (N. J. Salkind, ed.), pp. 1620–1623, Thousand Oaks, CA: Sage, 2010.

[11] K. Tsuda, M. Kawanabe, and K. Muller, “Clustering with the fisher score,” in *Advances in Neural Information Processing Systems*, Neural information processing systems foundation, 1 2003.

[12] Q. Gu, Z. Li, and J. Han, “Generalized Fisher score for feature selection,” *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence, UAI 2011*, 02 2012.

[13] C. C. Aggarwal, *Data Mining: The Textbook.* Cham: Springer, 2015.

[14] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, pp. 273–297, Sep 1995.

[15] A. Trabelsi, Z. Elouedi, and E. Lefevre, “Decision tree classifiers for evidential attribute values and class labels,” *Fuzzy Sets and Systems*, vol. 366, pp. 46 – 62, 2019.

[16] W. Chen, X. Xie, J. Wang, B. Pradhan, H. Hong, D. T. Bui, Z. Duan, and J. Ma, “A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility,” *CATENA*, vol. 151, pp. 147 – 160, 2017.

[17] J. Gou, H. Ma, W. Ou, S. Zeng, Y. Rao, and H. Yang, “A generalized mean distance-based k-nearest neighbor classifier,” *Expert Systems with Applications*, vol. 115, pp. 356 – 372, 2019.

[18] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, pp. 21–27, January 1967.

[19] “1.1. Generalized Linear Models – scitkit-learn 0.21.2 documentation.” https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression. Accessed: 2019-05-17.

[20] K. Roy, S. Kar, and R. N. Das, “Chapter 6 - selected statistical methods in QSAR,” in *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment* (K. Roy, S. Kar, and R. N. Das, eds.), pp. 191 – 229, Boston: Academic Press, 2015.

[21] S. Yadav and S. Shukla, “Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification,” in *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, pp. 78–83, Feb 2016.

[22] S. Nomm, K. Bardos, A. Toomela, K. Medijainen, and P. Taba, “Detailed analysis of the Luria’s alternating series tests for Parkinson’s disease diagnostics,” in *ICMLA*, pp. 1347–1352, IEEE, 2018.

[23] S. Nõmm, A. Toomela, J. Kozhenkina, and T. Toomsoo, “Quantitative analysis in the digital Luria’s alternating series tests,” in *2016 14th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pp. 1–6, Nov 2016.

[24] “Riverbank | software | pyqt | what is pyqt?.” https://riverbankcomputing.com/software/pyqt/intro. Accessed: 2019-05-09.

[25] “Pyinstaller quickstart – pyinstaller bundles python applications.” https://www.pyinstaller.org/. Accessed: 2019-05-09.

[26] “Pyqtgraph - scientific graphics and gui library for python.” http://www.pyqtgraph.org/. Accessed: 2019-05-17.

[27] “Tests of significance.” http://www.stat.yale.edu/Courses/1997-98/101/sigtest.htm. Accessed: 2019-05-12.

[28] K. Bardõš, “Analysis of interpretable anomalies and kinematic parameters in Luria’s alternating series tests for Parkinson’s disease modeling,” Master’s thesis, Tallinn University of Technology, Estonia, 2018.

Appendix 1 - Online location of appendices

Due to the large volume of resulting data tables constituting the appendices, they are presented online: https://github.com/takoss/optimal-length-appendix.