TAL TECH

DOCTORAL THESIS

Leveraging Artificial Intelligence for Microkinematic Analysis of Fine Motor Skills in Parkinson's Disease Detection

Elli Valla

TALLINNA TEHNIKAÜLIKOOL
TALLINN UNIVERSITY OF TECHNOLOGY
TALLINN 2025

TALLINN UNIVERSITY OF TECHNOLOGY
DOCTORAL THESIS
26/2025

Leveraging Artificial Intelligence for Microkinematic Analysis of Fine Motor Skills in Parkinson's Disease Detection

ELLI VALLA

TAL TECH PRESS

TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies
Department of Software Science

The dissertation was accepted for the defence of the degree of Doctor of Philosophy in Computer Science on 24.04.2025

Supervisor: Prof. Sven Nõmm,
Department of Software Science,
School of Information Technologies,
Tallinn University of Technology,
Tallinn, Estonia

Co-supervisor: Prof. Aaro Toomela,
Institute of Psychology,
Tallinn University,
Tallinn, Estonia

Opponents: Professor Lucio De Paolis,
University of Salento,
Lecce, Italy

Professor Jianhua Zhang,
Oslo Metropolitan University,
Oslo, Norway

Defence of the thesis: 16.05.2025, Tallinn

Declaration:
Hereby I declare that this doctoral thesis, my original investigation and achievement, submitted for the doctoral degree at Tallinn University of Technology, has not been submitted for any academic degree elsewhere.

Elli Valla

signature

ISSN 2585-6898 (publication)
ISBN 978-9916-80-305-9 (publication)
ISSN 2585-6901 (PDF)
ISBN 978-9916-80-306-6 (PDF)
DOI https://doi.org/10.23658/taltech.26/2025

Valla, E. (2025). *Leveraging Artificial Intelligence for Microkinematic Analysis of Fine Motor Skills in Parkinson's Disease Detection* [TalTech Press]. https://doi.org/10.23658/taltech.26/2025

TALLINNA TEHNIKAÜLIKOOL
DOKTORITÖÖ
26/2025

Tehisintellekti rakendamine peenmotoorika mikrokinemaatilises analüüsis Parkinsoni tõve tuvastamiseks

ELLI VALLA

TAL TECH
KIRJASTUS

Contents

List of Publications

The present Ph.D. thesis is based on the following publications, which are listed in chronological order and are referred to in the text by Roman numerals. Publications in which I am the main author are highlighted in **bold**.

I **Elli Valla, Sven Nõmm, Kadri Medijainen, Pille Taba, and Aaro Toomela. Tremor-related feature engineering for machine learning based Parkinson's disease diagnostics. *Biomedical Signal Processing and Control*, 75:103551, 2022**

II Erik Dzotsenidze, Elli Valla, Sven Nõmm, Kadri Medijainen, Pille Taba, and Aaro Toomela. Generative adversarial networks as a data augmentation tool for CNN-based Parkinson's disease diagnostics. volume 55, pages 108–113. Elsevier, 2022

III Vassili Gorbatsov, Elli Valla, Sven Nõmm, Kadri Medijainen, Pille Taba, and Aaro Toomela. Machine learning based analysis of the upper limb freezing during handwriting in Parkinson's disease patients. volume 55, pages 91–95. Elsevier, 2022

IV **Elli Valla, Henry Laur, Sven Nõmm, Kadri Medijainen, Pille Taba, and Aaro Toomela. Deep learning based segmentation of Luria's alternating series test to support diagnostics of Parkinson's disease. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1066–1071. IEEE, 2023**

V **Elli Valla, Ain-Joonas Toose, Sven Nõmm, and Aaro Toomela. Transforming fatigue assessment: Smartphone-based system with digitized motor skill tests. *International journal of medical informatics*, 177:105152, 2023**

VI Xuechao Wang, Junqing Huang, Marianna Chatzakou, Sven Nõmm, Elli Valla, Kadri Medijainen, Pille Taba, Aaro Toomela, and Michael Ruzhansky. Comparison of one-two-and three-dimensional CNN models for drawing-test-based diagnostics of the Parkinson's disease. *Biomedical Signal Processing and Control*, 87:105436, 2024

VII **Elli Valla, Gert Kanter, Sven Nõmm, Anton Osvald Kuusk, Peeter Maran, Karl Mihkel Seenmaa, Killu Mägi, and Aaro Toomela. Enhancing cerebral palsy gait analysis with 3D computer vision: A dual-camera approach. In *2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 1352–1357, 2024**

VIII **Elli Valla, Lilian Väli, Sven Nõmm, and Aaro Toomela. Smartphone-based microkinematic feature analysis for mental fatigue detection using machine learning. *Computers in Biology and Medicine*, Submitted 2025** [1]

[1] Submitted to Cognition, Technology and Work, Springer

Author's Contributions to the Publications

I I was the first author. I developed the machine learning framework, engineered the features, performed the formal analysis, created the visualisations, and wrote the manuscript.

II My contributions included the initial concept of employing GANs for data augmentation. I oversaw the deep learning technique, analysed the classification results, and wrote the manuscript.

III I was responsible for developing the machine learning analysis pipeline, including coding and feature engineering, and conducting the necessary investigations.

IV I served as the first author. I developed the machine learning classification algorithm, supervised the deep learning segmentation process using YOLO, and wrote the manuscript.

V As the first author, I conceptualised and led the smartphone-based fatigue assessment experiment, supervised the development of the accompanying application, oversaw data collection and formal analysis, implemented the machine learning classification model, and wrote the manuscript.

VI I pioneered the conceptualisation of transforming and improving 3D data, taking care of data generation and refinement.

VII As the first author, I was responsible for the conceptualisation of the experimental framework, supervised the development of the data acquisition and machine learning pipeline and wrote the manuscript.

VIII I was responsible of designing the experimental framework, oversaw data collection, and contributed to the development of the smartphone application and the machine learning model, in addition to drafting the manuscript.

Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
PD	Parkinson’s Disease
CP	Cerebral Palsy
CNN	Convolutional Neural Networks
GAN	Generative Adversarial Networks
XAI	Explainable Artificial Intelligence
YOLO	"You Only Look Once" (object detection algorithm)
SVM	Support Vector Machines
DT	Decision Trees
RF	Random Forest
LR	Logistic Regression
AB	AdaBoost
KNN	K-Nearest Neighbours
RTS	Reaction Test (Simple)
RTA	Reaction Test (Advanced)
ASD	Archimedean Spiral Drawing test
LAS	Luria’s Alternating Series test
mAP	Mean Average Precision
PEF	Physical Exertion based Fatigue
MEF	Mental Exertion based Fatigue
SHF	Sleep Hours based Fatigue
SAF	Self-Assessed Fatigue

1 Introduction

Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder after Alzheimer's disease [9]. It affects approximately 0.5% to 1% of people aged 65 to 69 years worldwide, with a prevalence rising to 1% to 3% among those aged 80 years and older [9]. The number of patients with PD is expected to double by 2040, exceeding 12 million [10]. PD is characterised by motor symptoms such as tremor, rigidity, and bradykinesia, as well as non-motor symptoms such as cognitive changes and sleep disorders [11]. Traditional diagnostic approaches, such as clinical evaluations and neuroimaging, often encounter challenges related to subjectivity, high costs, and limited accessibility [12][13]. PD diagnosis has traditionally been based on the observation of motor symptoms, which - although crucial - are typically evaluated using rating scales that lack robust validation and standardisation. Meanwhile, non-motor manifestations such as cognitive and sensory abnormalities can emerge early, sometimes preceding the classic motor signs, yet they remain too nonspecific to serve as definitive diagnostic indicators [11]. This illustrates a broader need for refined clinical tools and biomarkers that capture both motor and non-motor dimensions of PD.

Handwriting analysis, supported by artificial intelligence (AI), or more specifically, ma-chine learning (ML), is proving to be a powerful tool in the diagnosis of PD [14][15][16][17]. This approach involves using digital devices, such as tablets and smart pens, to capture key dynamic features during writing tasks, such as pressure, speed, and pen orientation. These features are indicative of motor impairments associated with PD. The captured data are then processed and fed into ML models, which are trained to distinguish between healthy controls (HC) and patients with PD. This method improves the accuracy and reliability of the diagnosis of PD by leveraging advanced algorithms to interpret subtle variations in motor function. The general workflow is depicted in Figure 1.

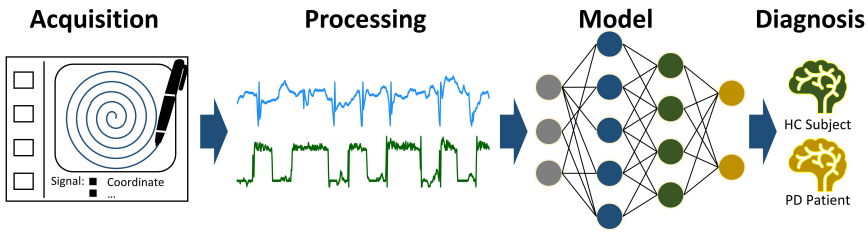


Figure 1: The workflow for AI-supported diagnosis of Parkinson's disease.

This research explores how AI-based analysis of fine motor skills can improve early detection and classification of PD. In addition to addressing clinical challenges such as subjectivity and lack of standardization, the research emphasizes usability in real-world settings. To demonstrate the broader feasibility of this methodology, a secondary application involving fatigue assessment using smartphone-based motor testing is also included. Although not the primary focus, this use case helps to illustrate the flexibility of the proposed tools in different diagnostic contexts.

To situate this research within the broader scientific landscape, the following section provides a detailed overview of existing machine learning approaches for fine motor skill analysis in the diagnosis of Parkinson's disease.

1.1 Literature review of machine learning approaches in fine motor skill analysis for diagnosing Parkinson's disease

In recent years, advances in ML and deep learning (DL) have revolutionised diagnostic methodologies, offering new tools for analysing motor impairments, such as those evident in handwriting and motion dynamics. Despite significant progress in the field, challenges remain in consolidating existing research, identifying gaps, and evaluating the performance of emerging technologies across diverse datasets and diagnostic modalities. To address these challenges, this review of the literature aims to map the current landscape of ML applications in fine motor skill analysis for the diagnosis of PD. Specifically, we explore the diverse data acquisition technologies, feature extraction methodologies, and ML models used in the field, highlighting their strengths, limitations, and potential for clinical application. By synthesising and critically analysing the existing body of work, we seek to provide researchers and clinicians with a comprehensive understanding of the field while identifying key opportunities for future advancement. The review was carried out systematically using the PRISMA guidelines (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). A structured approach was employed to identify, screen, and select relevant studies from multiple academic databases, including PubMed, Google Scholar, Scopus, and Web of Science. A comprehensive overview of the Boolean search queries, as well as the inclusion and exclusion criteria, can be found in the supplementary materials (Tables 19 and 20). These criteria were designed to ensure methodological rigour and relevance by capturing studies focused on ML-based approaches to motor function data for diagnostic purposes, while excluding non-diagnostic applications, non-motor data, or animal models. The search yielded 559 studies, of which 472 unique records were identified after removing duplicates. Following a rigorous screening process, 62 studies met the inclusion criteria and were analysed in detail (see Figure 2). The selected studies were categorised according to data acquisition technologies, extracted features, and applied ML models, providing a comprehensive overview of the field. An overview of the included studies is provided in the supplementary materials (Table 21), offering insight into the range of digital tools, data sources, and ML approaches employed for PD diagnosis. The

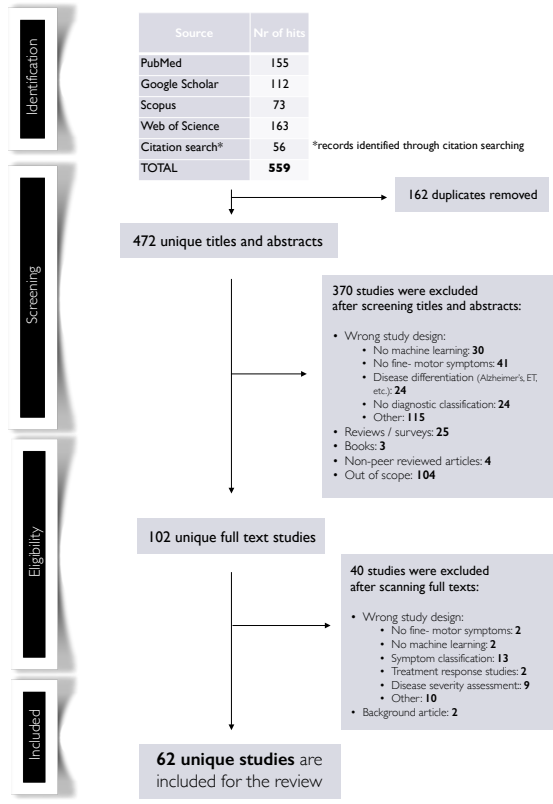


Figure 2: PRISMA flowchart

key trends and insights are visualised in Figure 3, showcasing the evolution of research focus, the prevalence of different types of devices, and the comparative performance of the ML and DL models in the diagnosis of PD. The radar chart demonstrates the performance of traditional ML and DL models in four key metrics: accuracy, sensitivity, specificity, and F1 score. DL models consistently outperform traditional ML methods, showcasing their superior diagnostic capabilities. This advantage stems from DL's ability to automatically extract complex, high-dimensional features from handwriting tasks, motion signals, and image-based data. However, despite their impressive accuracy, DL models have notable limitations. They are highly data-hungry and require large, high-quality datasets to generalise effectively. However, this poses a challenge in Parkinson's research, where data are often scarce and collected from small participant cohorts. This reliance on limited data increases the risk of overfitting, where models perform well on specific datasets but struggle to generalise to broader, real-world populations. Additionally, DL models are resource-intensive, requiring significant computational power and hardware such as graphical processing units (GPUs), which limits their practical application in clinical settings or resource-constrained environments. Another important drawback is their lack of interpretability. Unlike traditional ML models that offer clear decision boundaries and feature explanations, DL models often function as "black boxes," making it difficult for clinicians to understand or trust the underlying decision-making process. While DL models undoubtedly improve accuracy, addressing these challenges through strategies like data augmentation, explainable AI (XAI), and rigorous validation across diverse datasets is essential to ensure their reliability, robustness, and clinical utility. The horizontal bar chart in Figure 3 highlights the yearly distribution of research publications, with a peak in 2019 and 2021, reflecting the increased research activity during those years. Although publication counts have decreased slightly in recent years, the field maintains consistent output through 2024. While DL has grown substantially over the years, it has not completely displaced traditional ML methods. Instead, the two approaches have often complemented each other, with DL applied to handle complex datasets and ML used in scenarios where interpretability or computational simplicity is prioritised. In 2024, the balance seems to shift slightly towards traditional methods, possibly due to the need for trustable and interpretable outcomes in clinical settings. The pie chart in Figure 3 illustrates the proportion of device usage in all studies, with digital tablets (37%) emerging as the most commonly utilised tool, likely due to their reliability in capturing handwriting or motion data. Motion tracking technologies (26%) are also widely used, followed by smartphones (18%) and wearables (8%), reflecting a growing focus on mobile and real-time diagnostic tools. Custom hardware configurations, labeled as "Other systems" (11%), encompass pen-and-paper digitisation, surface electromyography (sEMG) based systems, and vision-based techniques, all of which significantly enhance innovation in data collection methodologies.

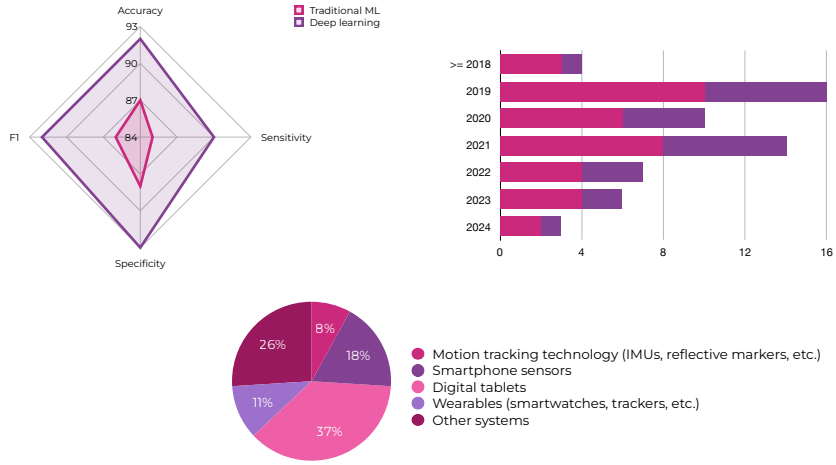


Figure 3: Overview of AI-based PD research trends.

The conducted literature review highlights the growing use of digital tools to assess fine motor skills, particularly handwriting dynamics, as a significant indicator of PD. Non-wearable sensors, such as digital tablets, remain the most prominent technology, enabling precise measurements of handwriting kinematics and geometric properties. Devices such as the Wacom Intuos tablet have been extensively used, allowing studies such as [18][19] to achieve accuracies ranging from 86% to over 91%. Smartphones have also emerged as portable and cost-effective alternatives, with applications capable of capturing spiral drawings and other fine motor assessments, achieving diagnostic accuracies around 90% [20]. Similarly, wearable sensors, particularly smartwatches, leverage built-in accelerometers and gyroscopes to monitor movement patterns, as demonstrated by [21], where smartwatch data yielded an accuracy of 89.3%. Beyond the technology used for data collection, the choice of ML techniques significantly influences diagnostic performance. Traditional ML methods, such as support vector machines (SVMs), decision trees (DTs), and Random Forest (RF), remain widely used due to their interpretability and computational efficiency. Studies like [18] have demonstrated the effectiveness of SVM classifiers, achieving accuracies as high as 91.6%. However, traditional methods rely heavily on manual feature extraction, which may limit their ability to capture the subtle, non-linear complexities inherent in handwriting and motion data. In contrast, DL methods, including convolutional neural networks (CNNs), bidirectional gated recurrent units (BiGRUs), and hybrid frameworks, have shown superior performance by automatically extracting high-dimensional features from raw data. For example, [22] achieved accuracies that exceeded 99% using fine-tuned CNN models, while [23] demonstrated the effectiveness of BiGRU for sequential handwriting analysis with accuracies surpassing 90%. Feature extraction methodologies also play a critical role in PD diagnostics, with kinematic features, such as velocity, angular velocity, and acceleration, being the most studied. Studies [24][25][26][27] emphasised the importance of kinematic measures, by analysing the tangential velocity and the duration of the stroke. Geometric features, including deviations in spiral precision and centerline accuracy, complement kinematic metrics, as seen in studies such as [28][29].

Recent research has also incorporated advanced metrics, such as entropy measures, spectral features, and non-linear dynamics, to enhance diagnostic capabilities. Advanced metrics like polar features (e.g., angular deviation and mean power frequency) have been introduced to capture fine-grained handwriting irregularities [30]. Studies [31][32] applied Shannon entropy and frequency-domain analysis to capture handwriting irregularities, achieving notable improvements in accuracy. However, reliance on DL models and high-dimensional features introduces challenges related to computational demands, data requirements, and model interpretability, limiting their immediate adoption in clinical settings. Lastly, other systems, such as pen-and-paper digitisation and vision-based methods, remain significant. In the study [33], the authors combined traditional handwriting analysis with sEMG signals to achieve an accuracy of 97.8% with ANN classifiers.

Feature selection remains a critical step in the development of reliable PD diagnostic systems, as improper or suboptimal selection can result in overfitting, poor generalisation, and reduced model robustness. Studies have applied statistical techniques such as Principal Component Analysis (PCA) and others to identify the most discriminative features [18]. However, over-reliance on these traditional methods risks overlooking higher-order, non-linear features that could enhance model performance. This shortcoming highlights the importance of systematic and rigorous feature engineering, a process often underexplored or insufficiently documented in current research. Many studies neglect to address common pitfalls, such as non-nested feature selection or suboptimal feature extraction, leading to overly optimistic results that fail to translate into real-world applications. Addressing these gaps requires a shift toward more systematic documentation of feature selection methods, along with the adoption of advanced feature engineering practices that integrate kinematic, geometric, and higher-order features to fully capture PD-related motor impairments. Despite notable progress in the application of ML and DL methods, several limitations and research gaps persist. A major challenge is data scarcity: most studies rely on small, homogeneous datasets, often collected in controlled environments, which restrict the generalisability of diagnostic models to diverse populations or real-world scenarios. This limitation increases the risk of overfitting and reduces the robust-ness of the model when deployed outside the research setting [34]. Although DL models have demonstrated superior performance over traditional ML techniques, they remain computationally intensive and rely on large data sets. Their “black-box” nature further limits interpretability, a critical factor for clinical validation and adoption. Traditional ML models are valued for their transparency and clear decision boundaries, but they often lack the capacity to capture complex, non-linear relationships in the data. In contrast, DL models, while accurate, are less interpretable and do not provide actionable information for clinicians, complicating their integration into healthcare workflows.

The limited adoption of cost-effective data acquisition tools remains a barrier to scalability. Although non-wearable sensors, such as digital tablets, are the most commonly used tools for handwriting analysis, their reliance on structured laboratory conditions limits accessibility. Methods requiring specialised hardware, such as graphic tablets, depth cameras, or sEMG devices, are resource-heavy and impractical for deployment in low-resource settings. In contrast, smart devices and wearable technologies present promis-ing alternatives for real-world monitoring due to their portability and cost-effectiveness, but their full potential remains underutilised.

Overall, while handwriting analysis and machine learning techniques hold immense potential for PD diagnostics, addressing the limitations of data scarcity, interpretability, and scalability is crucial to advancing the field. Future research should focus on developing standardised protocols, enhancing model transparency through XAI, and leverag-

ing cost-effective, non-invasive tools for widespread clinical adoption. Rigorous feature engineering and robust validation on diverse datasets will be key to ensuring the reliability and applicability of diagnostic tools in real-world healthcare settings. Building on this foundation, it is essential to critically examine the persistent challenges that hinder the development of scalable, accurate, and clinically viable AI-driven diagnostic systems. Although significant progress has been made, notable gaps remain in areas such as data availability, model transparency, and the adoption of practical tools. These limitations not only constrain the generalisability of current approaches but also impede their translation into clinical practice.

1.1.1 Identified research gaps

In the existing literature on AI-driven diagnostics for PD, several research gaps remain unaddressed, which hinder the development of accessible, accurate, and scalable diagnostic systems. These gaps not only limit the effectiveness of early diagnosis but also pose challenges in applying AI tools to real-world scenarios. This section highlights the key research gaps that this work directly addresses.

A. **Data scarcity and lack of diverse datasets:** Many studies rely on small and homogeneous datasets, which limit the generalisability and robustness of AI models. This challenge is particularly critical in early-stage PD detection, where subtle motor and non-motor symptoms are difficult to capture in limited or non-diverse data. Furthermore, data sets often lack variability in demographics, handwriting styles, and task designs, reducing their applicability to real-world scenarios.

   - ✓ Our work, particularly in **Publication II**, addresses this gap by employing generative adversarial networks (GANs) to augment datasets. This approach increases the size of the training data, improving the performance and reliability of DL models in detecting early neurological impairments, thus improving diagnostic accuracy.
   - ✓ Additionally, **Publication V** contributes by presenting a smartphone application for evaluating accessible fine motor skills, validated for fatigue detection through integration of qualitative and kinematic data. This application provides a scalable solution for continuous data collection in real-world environments, addressing data scarcity, and further strengthening AI model performance in motor function analysis. We also present novel smartphone-based motor skill test sensor data alongside subject metadata for the scientific community.

B. **Overlooked importance of robust feature engineering and selection practices:** A significant gap in the field is the lack of attention to systematic and rigorous feature engineering and selection processes. Many studies fail to adequately document their methods or address common pitfalls such as non-nested feature selection or suboptimal feature engineering practices. This often results in poor model performance, overfitting, or overly optimistic diagnostic outcomes, reducing the reliability of findings in real-world applications.

   - ✓ **Publication I** emphasises the critical role of robust feature selection, showcasing the pitfalls of non-nested approaches and advocating for best practices to avoid inflated model performance metrics. It also introduced novel feature engineering techniques, including angular metrics and high-order differ-

ential features, enriching the diagnostic capability of handwriting-based models. We demonstrated that the combination of advanced feature engineering and proper feature selection practices leads to models with improved generalisability and diagnostic precision, setting a higher standard for future research in the field.

C. **Challenges in comprehensibility and clinical adoption:** Although deep learning models achieve high accuracy, their black-box nature limits their interpretability, creating a barrier for clinical adoption. Clinicians often prefer interpretable models that align with existing medical knowledge.

   ✓ **Publications I and IV** combined high-performing ML models with comprehensible feature sets, balancing accuracy with usability for clinical decision support. By developing a segmentation framework for handwriting tests, we provided clear visual representations for both kinematic- and drawing-based diagnostics.

D. **Limited adoption of non-invasive and cost-effective data acquisition tools:** Despite the promise of modern devices such as smartphones, tablets and wearable sensors, many studies still rely on expensive and resource-intensive systems such as high-end graphics tablets or dedicated hardware. These setups restrict scalability and accessibility, particularly in remote or low-resource settings.

   ✓ **Publications V**, **VII, VIII** address this issue by paving the way for smartphone-based motor data analysis. In **Publications V and VIII**, smartphones were used for fine motor skill assessment in 41 unique users, later expanding to 131 users, demonstrating the scalability of the approach. **Publication VII** optimised marker-based systems with video-based analysis and pose estimation, extracting 3D models and essential gait parameters using only two cameras and computer vision techniques. This methodology can be adapted to smartphones, offering a more accessible and scalable solution. Gait parameters such as cadence, step length, single and double support, and walking speed were calculated in collaboration with clinicians, providing comprehensive information on gait mechanics crucial for diagnosing and treating movement disorders.

These identified research gaps highlight the need for innovation in areas such as early detection, real-world usability, scalable data acquisition, and data enhancement. By addressing these challenges, my research advances the field of AI-based diagnostics for PD, offering practical, scalable solutions that can significantly impact clinical practice.

1.2 Problem statement and research questions

Despite significant progress, existing AI-based diagnostic methods often lack transparency in feature selection, making results less interpretable for the medical community. Additionally, there is a need for tools that can seamlessly integrate into clinical practice, offering both high accuracy and ease of use. Addressing these challenges, the main objective of this thesis can be summarised as follows.

Objective

This research aims to develop and validate an AI-driven framework for analysing human motor function to detect Parkinson's disease. Focusing on advanced feature engineering techniques for handwriting analysis, the study evaluates their diagnostic accuracy and explores integration with mobile smart devices like smartphones and tablets. The ultimate objective is to deliver accessible, scalable solutions that advance early diagnosis and support rehabilitation for motor impairments and other neurological disorders.

In order to achieve this objective, several key research questions have been formulated. These questions are designed to guide the investigation into the potential of AI and fine motor skill kinematics to revolutionise the diagnosis and monitoring of neurological and cognitive impairments. By addressing these questions, the research will uncover new insights into the effectiveness of AI-based diagnostics and the integration of these tools into clinical practice. The following **research questions** will be addressed in this thesis:

RQ1: How can advanced feature engineering techniques and data augmentation methods improve the diagnostic accuracy and robustness of AI models to detect Parkinson's disease?

RQ2: How can scalable and cost-effective tools, such as smartphone-based applications, transform data collection practices for motor function diagnostics, allowing widespread accessibility and real-world applicability?

RQ3: How can machine learning frameworks be developed to enhance comprehensibility for clinicians, integrate advanced features, and maintain high accuracy, ensuring alignment with clinical workflows and fostering real-world adoption?

The answers to these research questions will contribute to the field by providing theoretical and practical advances. These contributions are crucial in developing robust AI-based diagnostic models and tools that can be used effectively in clinical settings. The research will focus not only on the development of these tools but also on their validation through experimental studies, ensuring their applicability and reliability in real-world scenarios. Figure 4 provides a research outline, highlighting the main problems, methods, and outcomes addressed in this thesis.

Figure 4: Research outline

1.3 Structure of the thesis

The main content of this dissertation is organised into chapters that address key research objectives. Sections 1 and 1.1 introduce the research, including a detailed review of the state-of-the-art in AI-driven PD diagnostics and motor skill analysis, identifying the research gaps that motivate this work. The section also concludes the problem statement, research questions, and the overall structure of the thesis, setting the stage for the investigations that follow. Section 2 presents the methodological framework for data-driven diagnostics. Section 2.1 covers the development of digital tools for data acquisition, including the smartphone application for fatigue detection, and relevant databases like SmartPhoneFatigue and DraWritePD to address **RQ2**. It discusses how these digital systems facilitate the collection of fine motor skill data for use in AI-based models. Subsection 2.3 focusses on answering **RQ1**. This chapter explores advanced data transformation and feature engineering techniques, such as the derivation of high-order kinematic features and the use of GANs for data augmentation. Additionally, it covers multi-dimensional data representation and automated segmentation for handwriting and drawing analysis, presenting the most diagnostically useful features for detecting cognitive impairments. Section 3 addresses **RQ1**, and **RQ3**. This chapter discusses ML approaches and classification models for the detection of PD. The integration of AI-based methods with traditional clinical diagnostics is critically evaluated, with a focus on improving early detection and ongoing monitoring. Section 5 summarises and concludes the contributions of the thesis. This

section revisits all research questions and reflects on how the findings addressed these inquiries. Highlights advances in AI-based diagnostics, particularly in improving accuracy and accessibility through digital tools, and discusses potential future work to enhance the models' application in clinical settings.

2 Methodological framework for data-driven diagnostics

This section outlines the methodological framework behind the ML-based approach to data-driven diagnostics. Each step contributes to transforming raw input into interpretable outcomes and is supported by one or more related publications.

2.1 Digitalisation and data acquisition tools and materials

This section highlights the development of digitalisation and data acquisition systems for the evaluation of motor skills, specifically through a smartphone-based application designed to detect fatigue, addressing **RQ2**. Bridges the technical aspects of system development with the practical workflow of the application, demonstrating how smartphone-based tools can improve data collection processes in neurological diagnostics. By meeting the clinical demand for practical and accurate assessments, this application contributes meaningfully to the advancement of digital health tools.

2.1.1 Development of the smartphone application for fine motor skill assessment

To bridge accessibility gaps and improve the practicality of assessing neurological conditions, we developed a smartphone-based suite of motor skill tests aimed at evaluating fine motor skills and cognitive impairments. The application offers a user-friendly and scalable solution for real-time remote monitoring and assessment of motor performance. Using widely available mobile technology, it enables frequent assessments in real-world settings, reducing barriers to early detection and continuous monitoring of neurological disorders. The methodologies described in this section are based on the work detailed in **Publication V** and **VIII**. The workflow of the smartphone application is visualised in Figure 5, which illustrates the step-by-step process, from user registration to completion of tests, data collection and feature extraction. The flowchart outlines how users interact with the app and participate in a series of motor skills tests. The pre-test questionnaire is designed to gather essential background data from users prior to their first motor skill assessment. It includes demographic and contextual variables such as gender, age, height, weight, dominant hand, education level, and lifestyle indicators like daily activity type and fatigue perception. In subsequent sessions, the questionnaire adapts to include more task-specific factors: users are asked to rate their mental and physical effort, interest, anxiety, and sleep hours on a scale from 0 to 10. This structured input supports a more personalised and context-aware interpretation of fine motor test results. The four distinct fine motor skill tests are designed to assess various aspects of motor control, including precision,

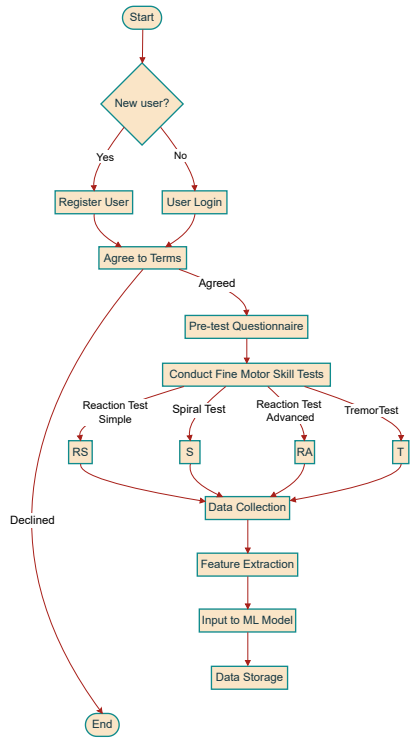


Figure 5: Sequential flowchart of user activities and testing in the fatigue detection app.

coordination, and reaction time.

The Reaction Test Simple (RTS) is the first test within the application and is designed to evaluate the user's response times, accuracy, and mistakes. In this test, the user is expected to tap on black dots that appear at various locations on the screen in a randomised manner, each differing in size. The total count of these black dots that the user must hit is fifteen. The user is provided with an animated tutorial that demonstrates the appropriate method for performing the test. The user workflow in this test is shown in Figure 6. The application records several parameters during the test: each screen tap, the coordinates of these taps, the accuracy of tapping directly on the black dots, the elapsed time in milliseconds between taps, and the dimensions of the screen of the user's smartphone. Moreover, the application also tracks the duration from the moment the user initiates the test to the point where the fifteenth black dot is tapped.

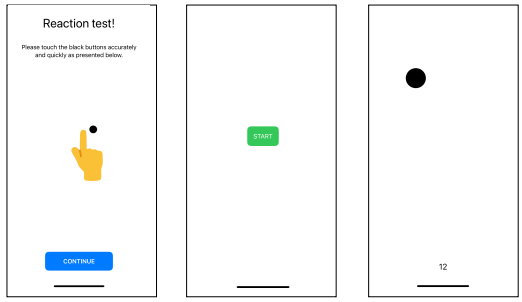


Figure 6: Screen views of the first reaction test (RTS) in the application. From the left: tutorial, start button, final view.

The Archimedean Spiral Drawing Test (ASD) is the second test within the application and is designed to have the user draw a spiral while maintaining the line within specified boundaries. Instructional guidance for this test is provided to the user through an ani-

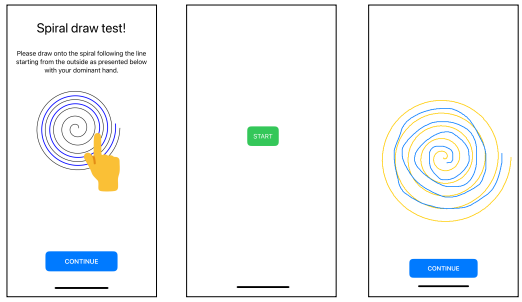


Figure 7: Screen views of the Archimedean spiral test in the application.

mated tutorial, which demonstrates the correct technique to perform the spiral drawing task. The user workflow in this test is shown in Figure 7. Several key metrics are recorded during this test. These include the height and width of the drawable area on the screen (depending on screen size), the coordinates of each point of the line drawn by the user, and an assessment of whether each point coincides with the pre-defined background line. In addition, the total duration taken by the user to complete the spiral drawing is mea-

sured. Another feature of the test is the real-time calculation of the percentage of the drawing that aligns with the background line, which is incorporated into the resulting data object after the completion of the test.

The Reaction Test Advanced (RTA) is the third test within the application and is designed to challenge users to tap on dots that correspond to a colour indicated at the bottom right of the screen.

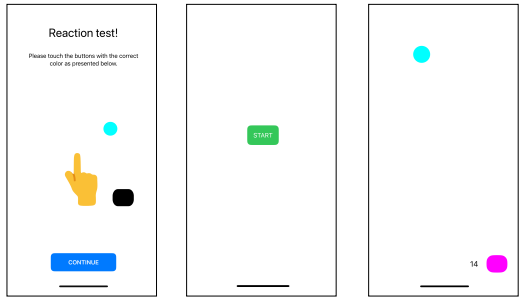


Figure 8: Screen views of the advanced reaction test (RTA) in the application.

The dots appear at various locations on the screen in a randomised manner, each differing in size and colour. This test features four pre-selected colours - purple, blue, yellow, and black. The user's task is to accurately tap on a dot when its colour matches the indicated colour. An animated tutorial is provided to instruct users on the proper execution of this test. The user workflow in this test is shown in Figure 8. This test records a variety of metrics: the height of the screen, the coordinates of each tap, the accuracy of tapping on the correct dot, the elapsed time since the last tap, and the time elapsed since the first appearance of a correctly coloured dot. In addition, the total duration taken by the user to complete the test is also captured. The test starts when the user taps the green 'START' button (shown in the second section of Figure 8) and finishes when the last correct dot is tapped.

The Tremor Test is the last test within the application and is designed to measure the hand tremors of the user.

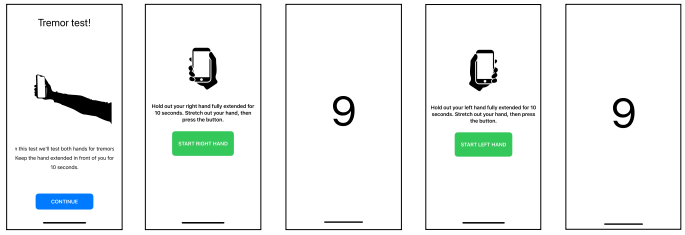


Figure 9: Screen views of the Tremor test in the application.

Users are expected to extend one hand outward while initiating the test by pressing the start button on the screen with their other hand. This test is repeated with both hands. The instruction for this test is conveyed through an image that demonstrates the correct

method to perform the tremor test. The user workflow in this test is shown in Figure 9. During this test, the smartphone's accelerometer sensors actively measure the movements of the hand in all directions for 10 seconds. The test is to be conducted identically with both hands to ensure consistent data collection starting with the left hand. The first half of the test starts with left-hand measurements when the user taps the green 'START LEFT HAND' button and finishes when 10 seconds have passed. The second part of the test for the right hand is identical to that of the left hand. Once users complete the motor skills tests, the application processes the data through a feature extraction pipeline. This pipeline transforms the raw data into key performance metrics such as reaction time distributions, tremor amplitude, and drawing smoothness. These extracted features are then fed into machine learning models that assess the level of fatigue of the participant. The integration of qualitative data from the pre-test questionnaire with the kinematic features collected during the tests allows the ML models to make more accurate and informed predictions. These predictions are stored for further analysis or clinical review, contributing to the ongoing improvement of diagnostic models. The back-end system of the application plays a critical role in ensuring data integrity and user engagement. It enforces time intervals between test attempts, preventing users from attempting multiple tests in quick succession, and ensuring consistency in data collection. The backend also provides users with feedback, allowing them to compare their current performance with previous results, providing information on potential improvements or deterioration in motor skills. In addition, the back-end supports the retrieval of test data over customisable date ranges, enabling longitudinal monitoring of user performance. This allows for a more nuanced analysis of motor function trends, which can be crucial in detecting early signs of fatigue or neurological impairment. The integration of these tests into a smartphone-based platform reduces the barriers to frequent and remote assessment of motor skills, making it accessible to a larger population. This innovative approach improves data availability and leverages the widespread use of smartphones to facilitate real-time monitoring and early detection of motor function anomalies associated with different neurological and cognitive impairments and other conditions. Figure 10 illustrates the general workflow for this smartphone-based fatigue assessment tool, which transitions from tablet-based to smartphone-based digitised motor skills tests, integrates metadata questionnaires, and applies feature engineering techniques. In addition, Figure 11 demonstrates an enhanced workflow for smartphone-based fatigue detection. Data collection is conducted on both iOS and Android devices and integrates fine motor skill tests (reaction, spiral draw, tremor tests) with qualitative metadata from self-assessed questionnaires. The extracted features are then fed into machine learning models for fatigue classification. This expanded approach facilitates an accurate fatigue analysis, as shown through performance evaluation metrics.

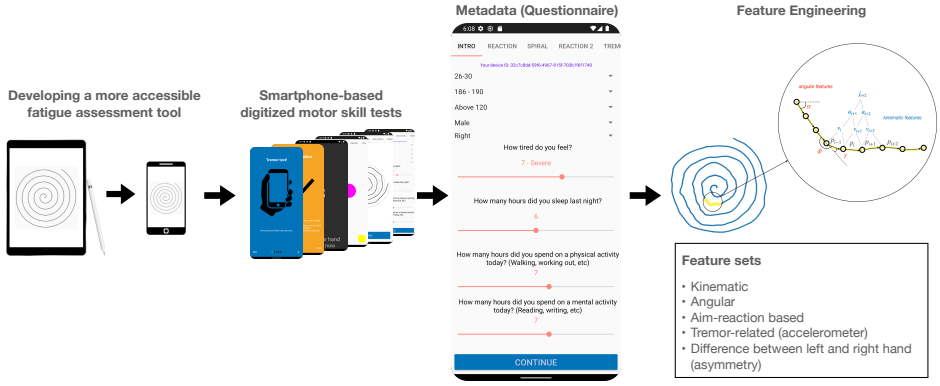


Figure 10: General workflow for a smartphone-based fatigue assessment tool. The workflow includes transitioning from tablet-based to smartphone-based digitised motor skill tests, integrating metadata questionnaires, and applying feature engineering techniques. Extracted feature sets include kinematic, angular, aim-reaction-based, tremor-related (via accelerometer), and asymmetry features.

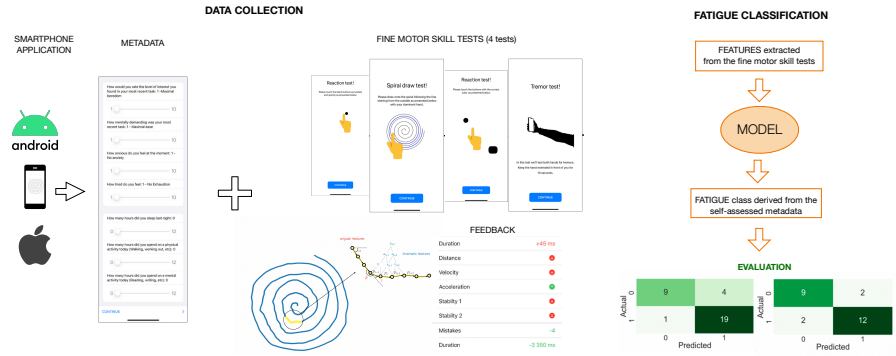


Figure 11: Enhanced smartphone application workflow for fatigue detection: Data collection on both iOS and Android devices integrates fine motor skill tracking and qualitative questionnaires, expanding the dataset for improved machine learning-based fatigue analysis.

With a well-established system in place, in the next section we turn our attention to examining the data collected through this platform to better understand its impact and potential.

2.1.2 A dataset for assessing fine motor skills using smartphone-based digital tasks

We developed a comprehensive dataset consisting of two versions of fine motor skill assessment tests, administered via mobile applications named *SmartPhoneFatigue* and *SmartPhoneFatigueV2*. Data collection involved 41 participants who completed 157 motor skills tests and self-reported their fatigue levels. The study was carried out under the supervision of the Tallinn University Ethics Board (decision number 12, dated 12.05.2021).

Following data cleaning procedures, including the removal of faulty tests and the detection of outliers, 131 trials were deemed suitable for analysis. These trials were carried out on 166 unique devices, with 94 participants completing the tests twice and 35 completing them once. Detailed non-PII participant information is provided in **Publication V**. The **Publication VIII** describes the second iteration of the database expanded to include 347 completed tests, capturing the dynamics of user interactions through computed features such as Euclidean distance, jerk, angular velocity, and cumulative slope angles. This work highlights the importance of a cross-platform application to ensure inclusive and representative data collection. The purification process also included the removal of all instances with missing values and a rigorous visual inspection of the smallest distances in spiral-drawing tasks. This refined the data set to 343 records, which was further segmented according to test completion time, resulting in 218 records.

2.1.3 Drawing and handwriting tests for Parkinson's disease diagnostics (DraWritePD)

This dataset, comprising contributions from 24 patients with PD (mean age 74.1 ± 6.7 years) and 34 healthy control subjects of the same age and sex (mean age 74.1 ± 9.1 years), was collected before my doctoral studies by our research group. The participants completed a battery of 12 different drawing and writing tests. Sample images of Archimedean spiral drawings from healthy individuals and Parkinson's patients, as well as Luria alternating series (LAS) patterns illustrating task performance and drawing trajectories, are shown in Figures 12 and 13, respectively.

(a) HC *(b) PD* *(c) PD*

Figure 12: Sample drawings of an Archimedean spiral performed by a healthy control subject (a) and the PD patients (b, c) from DraWritePD dataset.

Data acquisition for this research was performed using an Apple iPad Pro (2016) tablet and an Apple Pencil. The tablet, with a 26.77 cm (10.5 inches) diagonal screen, captures the Apple Pencil signal at a frequency of 240 points per second. From a software perspective, the data were collected using a custom iOS application developed by our research team. The dynamic features recorded by the tablet included x-coordinate (mm), y-coordinate (mm), timestamp (sec), pressure (force applied to the surface: [0,..., 6.0]), altitude (rad) and azimuth (rad). This dataset has since been utilised for classification and feature engineering tasks within my research. The data acquisition process adhered strictly to privacy laws, with the study approved by the Research Ethics Committee of the University of Tartu (No. 1275T-9).

2.1.4 Parkinson's disease handwriting (PaHaW) dataset

Data acquisition of the *PaHaW* dataset is described in detail in [35] and [36]. For the sake of self-sufficiency, the main properties of this data set important for the present studies are described in this section. The age and gender distribution of the *PaHaW* dataset is similar to that of *DraWritePD*. The data set consists of 37 patients with PD and 38 healthy controls (HC) with the same age and sex. HC subjects have a mean age of 62.4 years (stan-

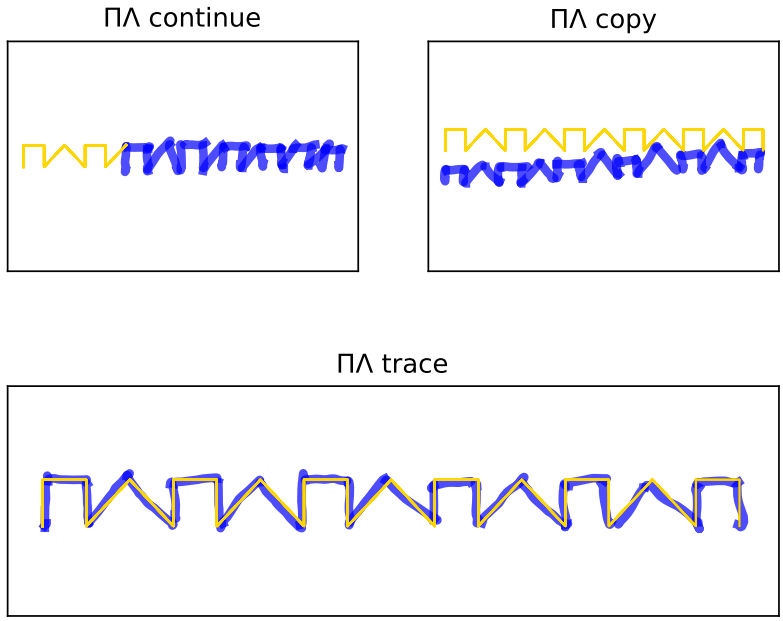


Figure 13: LAS patterns - The subject was assigned three distinct tasks. The initial pattern is depicted in yellow, while the blue line represents the trajectory of the subject's drawings.

dard deviation 11.3), while patients with PD have a mean age of 69.3 years (standard deviation 10.9) [35]. During the acquisition of *PaHaW* dataset, each subject was asked to complete a handwriting task according to the prepared pre-filled template at a comfortable speed. Subjects were allowed to repeat the task in the event of an error or a mistake in handwriting [35]. The handwriting signals were recorded using a Wacom Intuos digitising tablet overlaid with a blank sheet of paper, the sampling rate was set to 100 samples per second. The tablet captured the following dynamic features: x-coordinate; y-coordinate; timestamp; button status; pressure; altitude, and azimuth. All features were converted to the same units as in DraWritePD. The battery of the tasks presented in *PaHaW* dataset differs much from the one employed in *DraWritePD*. However, the Archimedean spiral drawing test is present in both datasets and was thus used in this research.

2.2 Deriving key attributes from data

In this research, we developed a system for digitising traditional paper-and-pencil tests to assess motor skills using smartphones and tablets, enabling enhanced data accessibility and supporting home monitoring. The system captures detailed motion data through the device's sensors, including touch coordinates, timing of interactions, and motion dynamics. Additionally, it records nuanced information such as azimuth and altitude angles of the stylus, providing a richer dataset for analysis. The collected data are processed to extract a wide range of features, such as positional coordinates, timing, and pressure, which are critical for evaluating motor control, tremors, and other fine motor functions. These features are crucial in diagnosing conditions such as Parkinson's disease, providing deeper insight into motor impairments through detailed kinematic and spatial data.

The sample of raw signals extracted from tablet-based drawing tests, which are used

in the assessment of motor function in PD, is presented in Table 1.

Table 1: Dynamic sequential stylus input features over three adjacent time points

t	x	y	p	a	l
533033966.322112	446.2969	-431.0742	0.723517	0.518733	1.059078
533033966.352263	449.875	-439.7695	0.739844	0.490138	1.059078
533033966.374942	454.7188	-448.125	0.800081	0.444148	1.059078

Note: The abbreviations x, y denote the x- and y-coordinate features; and a, l and p are the azimuth, altitude, and pressure features, respectively; timestamp is represented by t.

These features include the x- and y-coordinates of the pen's position, azimuth (a), altitude (l), and pressure (p) values, along with a timestamp (t) that records the exact time each point was captured. This data provides critical information on the drawing dynamics, reflecting the user's motor precision and control. Figure 14 illustrates a schematic of sample input, depicting the six independent features recorded for each data point as the user traced a spiral on the tablet. The diagram also indicates the drawing direction, offering a visual representation of how these features are collected in real-time during the test. The arrows highlight the progression of the pen's motion, while the values x and y represent the pen's position on the screen, and a, l, and p represent the pen's orientation and pressure applied at each moment. These raw signals provide a comprehensive dataset to assess motor control in patients with PD, contributing to a detailed analysis of their performance. By capturing these parameters, the system can detect subtle changes in motor behaviour, such as tremor frequency or variations in pressure and altitude, which are key indicators of motor dysfunction in PD. The combination of coordinate, orientation, and pressure data offers a rich foundation for further analysis and feature extraction, helping to diagnose and monitor motor symptoms related to PD.

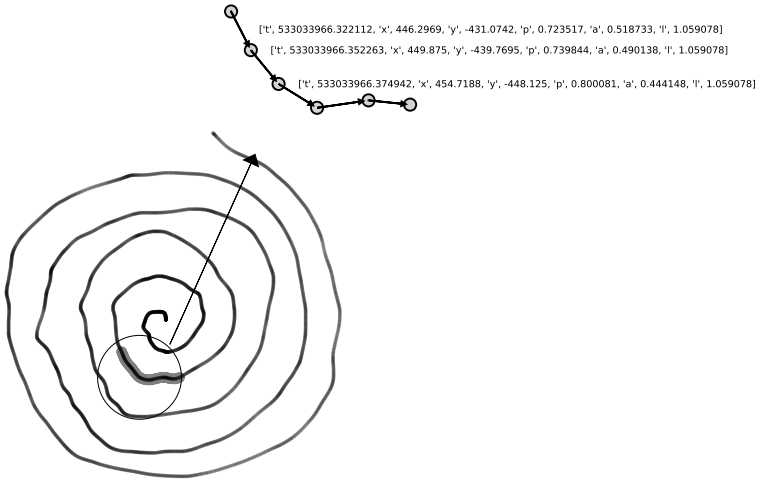


Figure 14: Schematic diagram of the sample input, illustrating the collection of six independent features for each data point. The arrows indicate the drawing direction. The abbreviations x and y represent the $x-$ and $y-$ coordinate features, while a, l, and p correspond to azimuth, altitude, and pressure, respectively. The timestamp is denoted by t.

The freezing of the hand during writing is referred to as *freezing episode*. According to [37] the freezing episode is defined as *a sudden, variable, and often unpredictable transient break in movement*. This definition needs to be formalised to allow for automatic

detection. The authors of [38] have adapted the definition for the case of hand movement during writing in the following way: *handwriting freezing was defined as an involuntary stop or clear absence of effective writing movements during at least* 1 *second*. The latest allows for being implemented in the form of programming code. Of course, one has to consider that, while the hand may freeze, small jigging in the coordinates may occur. The proper setting of threshold values may easily solve such problems. Once freezing episodes are detected, timestamps of the points where they begin and end provide the information about the ending and beginning points of the intervals to extract. In this research, the length of this interval was experimentally found to be 1 of a second. Figure 15 depicts freezing episodes, numbered and marked by yellow lines with green arrows.

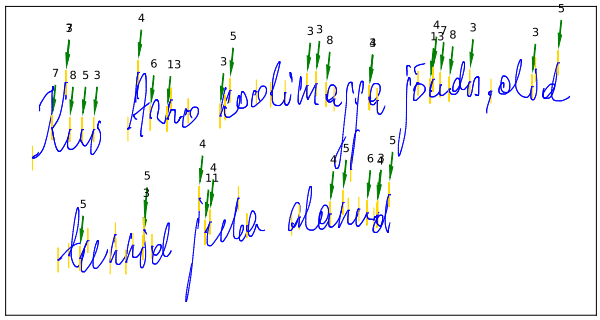


Figure 15: Freezing episodes in the sentence writing test from the DraWritePD data set.

It is important to note that the number of freezing episodes may vary between PD patients and HC subjects. To avoid problems caused by unbalanced data sets, proper sampling was applied to guarantee an equal number of freezing episodes and appropriate proportions of episodes from the different parts of the sentence. Then the feature engineering procedure described in the next section 2.3 was applied to these intervals. After each test, all the freezing episodes were described by the tuple of kinematic, pressure, and motion mass parameters. Each tuple inherited the label of the test it had been computed from, consequently forming the dataset to be used for ML analysis.

When evaluating the feasibility of smartphones for fine motor skill assessment, a parallel approach was taken to capture essential features. The Table 2 outlines the key attributes extracted during smartphone-based motor skill assessments, as detailed in **Publications V** and **VIII**. The features include touch coordinates, timestamps, and accelerometer measurements, which are critical for evaluating real-time motor performance under cognitively impaired conditions. For example, touch events are recorded with precise coordinates (x_i,y_i), and a Boolean value identifies whether the touch overlapped with the target area. The target area is a predefined region on the screen, typically a geometric shape such as a circle or rectangle, defined by specific dimensions and a centre position $(x_t\ ,y_t\ )$. The touch point is the exact location where the user makes contact, recorded as (x_i,y_i) with a timestamp t_i. Accuracy is evaluated by determining whether the touch point falls within the target area. For example, in a circular target, this involves checking if the distance between the touch point and the target centre is less than or equal to the target's radius. If the touch point overlaps with the target area, it is classified as a "hit." This evaluation is complemented by metrics such as reaction time, calculated as the time elapsed between the appearance of the target and the recorded touch, as well as temporal consistency metrics, such as the time elapsed between consecutive touches ($\Delta t_i = t_i - t_{i-1}$)

and the time from the first correct colour rendering to the touch.

Additionally, accelerometer data $(a_{x,k}, a_{y,k}, a_{z,k})$ capture motion during the test, providing insights into movement dynamics. The tremor features are derived by calculating the asymmetry between the absolute accelerations of the left and right hand:

$$f_{\text{tremor}} = \text{abs}(a_{\text{left}} - a_{\text{right}})$$

where the absolute acceleration is defined as:

$$\text{abs} = \sqrt{x^2 + y^2 + z^2}.$$

By combining data from touch events, timestamps, accelerometer readings, and derived tremor features, smartphone-based assessments provide a comprehensive understanding of motor function, reaction speed, and consistency. These tools are practical for continuous monitoring in real-world settings and serve as a valuable complement to traditional motor skill evaluation methods.

Table 2: Subset of defined features and their descriptions for the analysis of user interactions with a smartphone screen.

Feature	Notation	Description
Touch coordinates	$\mathbf{T}_i = (x_i, y_i)$	For each touch event i, the coordinates are recorded as (x_i, y_i).
Timestamps	t_k	Time series data t_k representing the time at each step k.
Total duration of the test	$T_{\text{total}} = t_{\text{end}} - t_{\text{start}}$	Total time T_{total} from the start time t_{start} to the end time t_{end}.
Accuracy of touches	$A_i = \begin{cases} 1 & \text{if } \mathbf{T}_i \text{ is on target} \\ 0 & \text{otherwise} \end{cases}$	Binary indicator A_i representing the accuracy of each touch i. A_i can be defined as a `wasHitOnTarget` feature, which is True if the area of the touch overlaps with at least one pixel of the rendered target.
Elapsed time between touches	$\Delta t_i = t_i - t_{i-1}$	The time difference Δt_i between consecutive touches i and $i-1$.
Elapsed time since stimulus appearance	$\Delta t_{i,\text{stimulus}} = t_i - t_{\text{stimulus}}$	The time difference $\Delta t_{i,\text{stimulus}}$ between the touch i and the appearance of a stimulus t_{stimulus}. This includes `timeFromFirstCorrectColorRender` feature, which measures the difference between the first correct color render and the touch time.
Screen dimensions	$\mathbf{S} = (H, W)$	Height H and width W of the user's smartphone screen or drawable area.
Real-time line alignment percentage	$P_{\text{align}} = \frac{\sum_{j=1}^{N} L_j}{N} \times 100\%$	Percentage P_{align} of the drawing that aligns with the predefined line, where N is the total number of points.
Accelerometer data	$\mathbf{A}_k = (a_{x,k}, a_{y,k}, a_{z,k})$	For each time step k, the accelerometer data is recorded as $(a_{x,k}, a_{y,k}, a_{z,k})$, representing the acceleration in x-, y-, and z-directions.
Derived tremor features	$f_{\text{tremor}} = \text{abs}(a_{\text{left}} - a_{\text{right}})$	Tremor data includes features such as absolute acceleration $\text{abs} = \sqrt{x^2 + y^2 + z^2}$. The *asymmetry* between the left-hand and right-hand absolute accelerations, defined as the absolute value of the difference between left (a_{left}) and right (a_{right}) accelerations.

Building on this foundation of clean and accurate data, the next step involves transforming raw signals into more meaningful and discriminative features.

2.3 Advanced data transformation and feature engineering

Feature engineering involves the derivation of informative and discriminative attributes from raw data, which improves the effectiveness of ML models. This section explores methods applied in the diagnostics of PD through handwriting analysis, focussing on tremor-related kinetic analysis and using high-dimensional data representations.

2.3.1 High-order kinematic features for enhanced motor function analysis

Building on the raw signals described in the previous section, such as pen position (x- and y-coordinates), timestamps, pen pressure, and orientation (altitude and azimuth), we

derive a range of kinematic features that offer deeper insights into motor impairments, particularly for the analysis of PD. These features go beyond basic motion descriptors, capturing both the dynamic aspects of hand motion and higher-order changes over time, which are crucial for identifying subtle motor dysfunctions that may be missed by traditional observational methods. To enhance the granularity of motor function assessment, we extended the feature set by incorporating higher-order kinematic features, specifically, jerk (the rate of change of acceleration), snap (rate of change of jerk), crackle, and pop. These features are computed from the pen tip's position vector, $\mathbf{p}_i = (x_i, y_i)$, sampled at discrete time steps t_i during handwriting tasks. Introduced in advanced dynamics contexts [39], these successively derived quantities offer a more detailed representation of motion, capturing subtle variations often missed by lower-order measures. When applied to handwriting tasks like spiral drawing, these microkinematic descriptors help capture subtle motion irregularities indicative of early Parkinsonian symptoms. Beyond positional data, we also examined the force exerted on the drawing surface. Higher-order derivatives of pressure, namely yank (the rate of change of pressure), tug (rate of change of yank), snatch, and shake, were calculated. These pressure-based microkinematics offer insight into variations in tremor amplitude and irregular force control, key markers of im-paired motor function in PD. By introducing higher-order derivatives for both position and pressure signals, this work extends conventional kinematic analysis into the domain of microkinematics. This enriched feature set enables machine learning models to detect nuanced motor symptoms with improved sensitivity, contributing to more precise diagnostics and progression monitoring in PD.

Another critical aspect of our analysis involves pen orientation and inclination, which are often underutilised in related studies. These angular features, azimuth (pen orientation) and altitude (pen inclination), were incorporated alongside kinematic data to capture more nuanced aspects of motor control during drawing tasks. To further enhance the feature set, we consider three additional angles derived from the pen trajectory: the slope angle α, the rotational angle ϕ, and the yaw angle γ, which are defined in 3. Given the slope k of the position vector, the angle α was calculated based on the change in coordinates between two consecutive points in the drawing.

Table 3: Angular features derived from pen trajectory

Angle	**Formula**
Slope (k_i)	$k = \frac{y_i - y_{i-1}}{x_i - x_{i-1}}$
Slope angle (α_i)	$\alpha = \arctan k_i$
Rotational angle (ϕ_i)	$\phi_i = \pi + \alpha_{i-1} - \alpha_i$
Yaw angle (γ_i)	$\gamma_i = \alpha_i - \alpha_{i-1}$

These expressions capture directional changes and rotational movements of the pen, enhancing the feature set for motor assessment. The yaw angle represents the directional change of the point vector, and these angular parameters were extended with their respective higher-order derivatives, enriching the analysis with up to the third derivative. These micro-changes in pen movement, though difficult to observe visually, are crucial for analysing tremor-like symptoms often seen in PD. Tremors typically manifest as irregular, less smooth movements, reflected by greater accelerations and frequent directional changes. By capturing such tremor-induced deviations in the trajectory, we can directly link these angular and kinematic features to motor impairments associated with PD.

Furthermore, we adopted the concept of "motion mass parameters" introduced in prior research [40], which have demonstrated strong discriminative power in distinguish-

ing PD patients from healthy controls. These parameters aggregate the absolute values of the kinematic and pressure-based features at each observation point throughout the test. The velocity mass V_N, for instance, is defined as:

$$V_N = \sum_{k=1}^{N} |v_k|$$

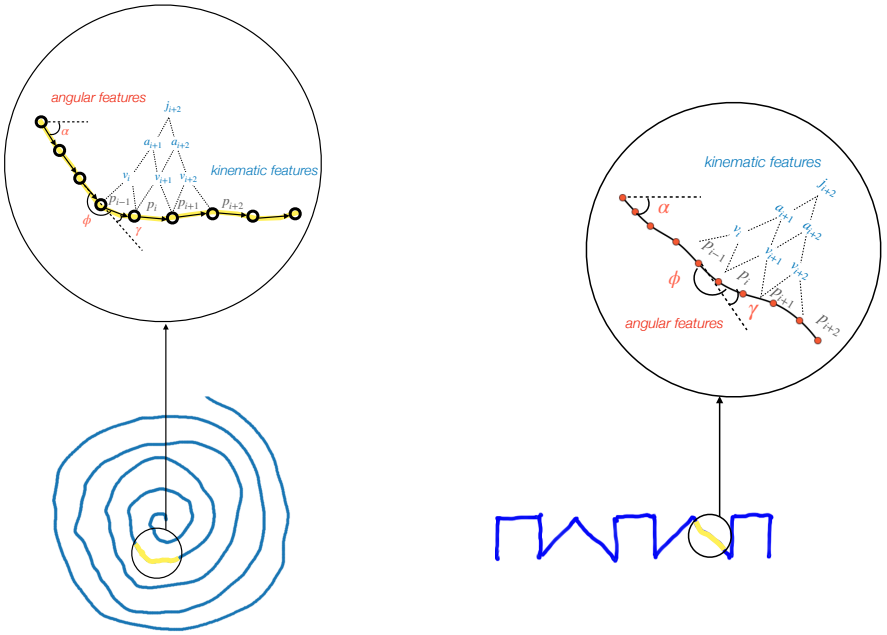
In the same way, the mass parameters for acceleration, jerk, and their higher-order derivatives, such as snap, crackle, and pop, were defined. These mass parameters were also applied to pressure data and angular changes, capturing the total "motion mass" during the drawing tasks.

A comprehensive overview of the engineered features is presented in Tables 4 and 5, which summarise both vector- and scalar-valued descriptors extracted from raw handwriting data collected during motor assessments. The differential and angular types of features discussed in this section are also visually illustrated in Figure 16. This figure depicts the extraction process for both kinematic features (velocity, acceleration, jerk, and higher-order derivatives) and angular features (yaw, rotational angle, and slope). Subfigure 16a shows the Archimedean Spiral Drawing test (ASD) [41], while subfigure 16b presents the Luria Alternating Series test (LAS) [42].

Table 4: Subset of vector-based features derived from stylus trajectory and pressure signals.

Feature set	Feature	Description	Mathematical definition
Spatial-temporal	displacement	Euclidean distance between two points	$d_i = \|\mathbf{p}_i - \mathbf{p}_{i-1}\|$
Kinematic	velocity $\mathbf{v}$	First derivative of position	$\mathbf{v}_i = \frac{d\mathbf{p}}{dt} \approx \frac{\mathbf{p}_i - \mathbf{p}_{i-1}}{t_i - t_{i-1}}$
	acceleration $\mathbf{a}$	Second derivative of position	$\mathbf{a}_i = \frac{d\mathbf{v}}{dt} \approx \frac{\mathbf{v}_i - \mathbf{v}_{i-1}}{t_i - t_{i-1}}$
	jerk $\mathbf{j}$	Third derivative of position	$\mathbf{j}_i = \frac{d\mathbf{a}}{dt} \approx \frac{\mathbf{a}_i - \mathbf{a}_{i-1}}{t_i - t_{i-1}}$
	snap σ	Fourth derivative of position	$\sigma_i = \frac{\mathbf{j}_i - \mathbf{j}_{i-1}}{t_i - t_{i-1}}$
	crackle χ	Fifth derivative of position	$\chi_i = \frac{\sigma_i - \sigma_{i-1}}{t_i - t_{i-1}}$
	pop ξ	Sixth derivative of position	$\xi_i = \frac{\chi_i - \chi_{i-1}}{t_i - t_{i-1}}$
Pressure-derived	pressure$_diff$	Change in pressure between steps	$\Delta f_i = f_i - f_{i-1}$
	yank ψ	First derivative of pressure	$\psi_i = \frac{df}{dt} \approx \frac{f_i - f_{i-1}}{t_i - t_{i-1}}$
	tug τ	Second derivative of pressure (rate of yank)	$\tau_i = \frac{d\psi}{dt} \approx \frac{\psi_i - \psi_{i-1}}{t_i - t_{i-1}}$
	snatch ζ	Third derivative of pressure	$\zeta_i = \frac{d\tau}{dt} \approx \frac{\tau_i - \tau_{i-1}}{t_i - t_{i-1}}$
	shake η	Fourth derivative of pressure	$\eta_i = \frac{d\zeta}{dt} \approx \frac{\zeta_i - \zeta_{i-1}}{t_i - t_{i-1}}$
Geometric	altitude_diff	Change in pen elevation	$\Delta\ell_i = \ell_i - \ell_{i-1}$
	azimuth_diff	Change in pen azimuth angle	$\Delta a_i = a_i - a_{i-1}$
	α_diff	Change in slope angle	$\Delta\alpha_i = \alpha_i - \alpha_{i-1}$
	ϕ_diff	Change in rotational angle	$\Delta\phi_i = \phi_i - \phi_{i-1}$
	γ_diff	Change in yaw angle	$\Delta\gamma_i = \gamma_i - \gamma_{i-1}$

The higher-order kinematic and pressure-based features introduced in this work directly address **RQ1**. By extending traditional velocity- and acceleration-based approaches to include derivatives such as snap, crackle, pop, this research enriches the feature space and improves the model's ability to detect subtle motor impairments characteristic of PD. Moreover, the incorporation of angular features, along with their respective higher-order derivatives, ensures a multi-dimensional assessment of handwriting movements that captures tremor-like symptoms more precisely. This advanced feature engineering framework not only bolsters diagnostic precision but also enhances the robustness of ML models against handwriting style variations, addressing the objectives of **RQ1** and paving the way for scalable, clinically viable AI solutions. Notably, the introduction of these high-order derivatives and angular features in **Publication I** marks a step forward in feature engineering techniques for evaluating motor impairments in patients with PD.



(a) (b)

Figure 16: Visual representation of angular and differential-type kinematic features extracted from stylus trajectories. Angular features (shown in red) include the slope angle α, rotational angle ϕ, and yaw angle γ, which capture abrupt directional changes, oscillatory motion, and rotational instability in pen movement. Kinematic features (shown in blue) are derived from the differential representations of the position vector $\mathbf{p}$, including velocity $\mathbf{v}$, acceleration $\mathbf{a}$, and jerk $\mathbf{j}$, reflecting the speed, smoothness, and control of motion. These features are computed using consecutive positional points sampled during the drawing of structured tasks such as (a) the Archimedean Spiral Drawing (ASD) test and (b) the Luria Alternating Series (LAS) test. Such quantitative representations are instrumental for capturing subtle motor abnormalities, especially in early-stage Parkinson's disease. When processed through machine learning models, these features support fine-grained classification and aid in distinguishing pathological handwriting from that of healthy individuals

Table 5: The sample subset of scalar features.

Feature Set	Feature	Description	Mathematical definition
Spatial-temporal features	duration	Time interval between first and last timestamp	$T = t_N - t_1$
Kinematic features	velocity_mass	Mass of velocity vector	$V_N = \sum_{i=2}^{N} \lvert\mathbf{v}_i\rvert$
	acceleration_x_mass	Accumulated x-directional acceleration	$A_x = \sum_{i=2}^{N} \lvert a_{x,i}\rvert$
	jerk_median	Median jerk magnitude	$\text{median}(\mathbf{j})$
	snap_mass	Accumulated snap magnitude	$S_N = \sum_{i=2}^{N} \lvert\boldsymbol{\sigma}_i\rvert$
Pressure features	shake_median	Median of shake	$\text{median}(\eta_i)$
	pressure_diff_min	Minimum pressure difference	$\min(\Delta f_{ij})$
	tug_mass	Accumulated tug magnitude	$\sum_{i=2}^{N} \lvert\tau_i\rvert$
Geometric features	ϕ_mass	Accumulated rotational angle	$\sum_{i=2}^{N} \lvert\phi_i\rvert$
	α_accel_min	Minimum angular acceleration	$\min(\ddot{\alpha}_i)$
	γ_std	Yaw variability	$\text{std}(\gamma)$

2.3.2 Leveraging generative adversarial networks for data augmentation

To address the challenge of limited data in PD diagnostics, this section outlines the use of generative adversary networks (GANs) [43] for data augmentation, a technique used to artificially increase the size and diversity of the data set, thus improving the robustness and generalisability of diagnostic models. This contribution, published in Publication II, is illustrated in Figure 17. The process begins with the collection of raw spiral test data, a clinical tool to assess motor dysfunction in patients with PD. The data then undergoes filtering and preprocessing steps, including normalisation, noise reduction, and feature extraction, to ensure quality and consistency. After preprocessing, the data are divided into three subsets: training, validation, and testing data sets. The training set is used to build the models, while the validation and testing sets are reserved for performance evaluation, ensuring the models avoid overfitting.

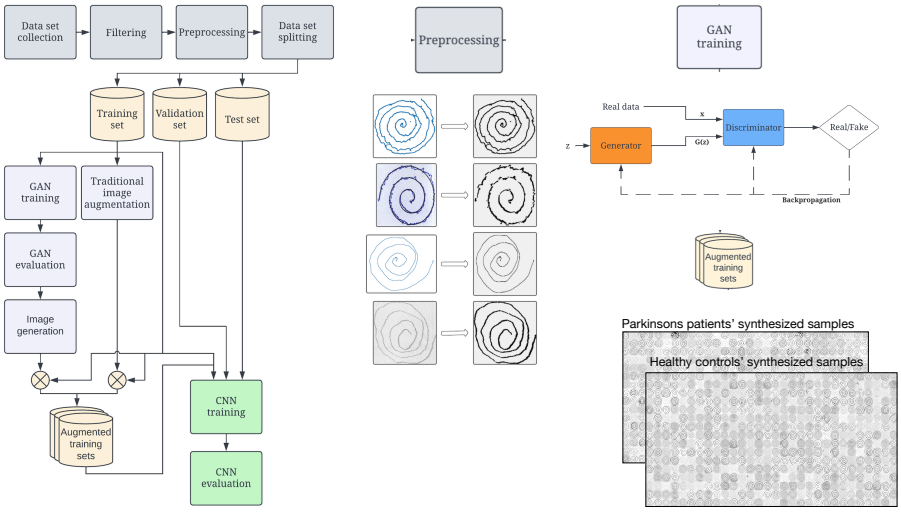


Figure 17: GAN-based data augmentation workflow

Traditional augmentation techniques such as rotation, scaling, and flipping are initially applied to the training set to artificially expand the diversity of input images. GAN training is then introduced, using real spiral test images from the training set to train a generator, which creates synthetic images, and a discriminator, which evaluates the authenticity of these images. The adversarial interaction between the generator and the discriminator continues until the generator produces spiral images that are nearly indistinguishable from the real ones. Once trained, the GAN generates synthetic spiral images that reflect the variability and complexity seen in real patient data, including characteristics representative of both PD patients and healthy controls, enabling the creation of enhanced training sets. Synthetic images are combined with those produced by traditional augmentation methods to form a more diverse and robust training dataset. The augmented data set is rigorously evaluated to ensure that the synthetic images enhance, rather than degrade, the quality of the training set. Finally, this enriched data set is used to train a CNN model designed to classify images as PD or non-PD. CNN is then validated and tested to ensure its accuracy, robustness, and ability to generalise effectively in real-world diagnostic scenarios. By generating synthetic yet realistic images and combining them with traditional augmentation methods, this workflow improves diagnostic model performance

while minimising the reliance on costly and time-consuming data collection. The result is a cost-effective and efficient approach to address the critical issue of limited data in PD research, ultimately enhancing diagnostic accuracy and reliability.

GANs consists of two parts: a generator G and a discriminator D. G uses a noise variable z as input from the distribution p_z (latent space) and produces an output in the form of an image that is in the generated distribution p_g. D takes in an image x and outputs the probability that x came from the real data distribution p_{real} rather than p_g. The adversarial training process involves both networks competing with each other. D is trained to maximise the probability that the correct label is assigned to an image. Simultaneously, G is trained to minimise the probability that the discriminator correctly labels the fake images. They play the following two-player minimax game with value function $V(D,G)$:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{real}}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1-D(G(z)))]$$

In theory, during training, the GAN model converges when both D and G reach a Nash equilibrium, i.e., a saddle point, which is the optimal point for Equation 1. The Nash equilibrium is achieved when players in a non-cooperative game lack any incentive to change their strategy. With GANs, this means that $p_g = p_{\text{real}}$ and D cannot tell the difference between real and generated samples. Finding this saddle point is a difficult task.

For image generation, we used four different GAN architectures: StyleGAN2-ADA [44], StyleGAN2-ADA + LeCam [45], StyleGAN3 [46], and Projected GAN [47]. These models were selected because they were specifically created with limited training data in mind and have been shown to generate good-quality images under these conditions. Each of the GAN models was trained using transfer learning from a model trained on the Flickr-Faces-HQ dataset (FFHQ) [48]. We used only the train split, which was amplified with horizontal, vertical flips, and horizontal + vertical flips during GAN training. The generated images are of size 256x256 pixels. Each model was trained on a single NVIDIA A100 GPU for three days. For model evaluation, we used the kernel inception distance (KID) [49], which measures the dissimilarity between probability distributions and is unbiased when used with small datasets, unlike the Fréchet inception distance (FID), which is used more frequently as a quality metric of GAN [50]. The best KID score for each of the trained GAN architectures can be seen in Table 6. Projected GAN achieves a KID score that is an order of

Table 6: Best KID scores of each trained GAN model

GAN	KID (↓)	
	HC	PD
StyleGAN2-ADA	0.01416	0.01054
StyleGAN2-ADA + LeCam	0.01826	0.02517
StyleGAN3	0.02148	0.02113
Projected GAN	0.001264	0.0009285

magnitude lower than the other models. To illustrate its training dynamics, the evolution of KID across training steps is shown in Figure 18, demonstrating both rapid convergence and stable performance over time.

The comparison between the original spirals (left) and the GAN-generated synthetic spirals (right) is illustrated in Figure 19. The figure demonstrates the ability of GANs to produce realistic spiral patterns, which closely mimic the visual characteristics of the original data while introducing slight variations in line smoothness and structure. These synthetic examples enhance dataset diversity and improve model robustness when incorporated into training.

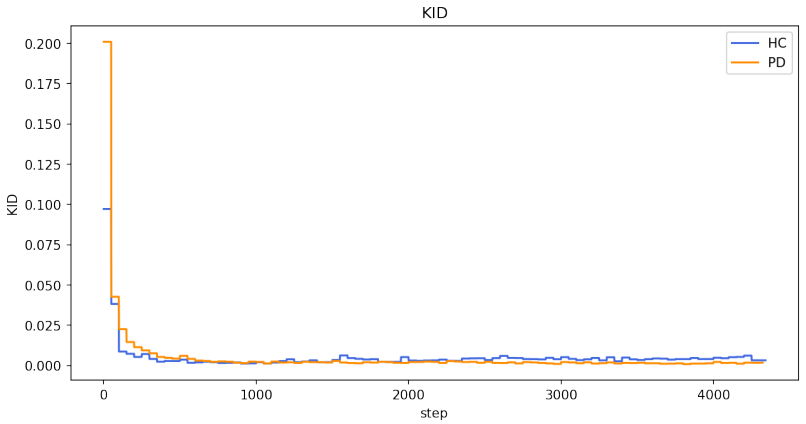


Figure 18: KID scores during Projected GAN training for HC and PD classes. The model converges rapidly within the first 500 steps and maintains low, stable values throughout.

In conclusion, GANs play a pivotal role in augmenting the dataset, creating synthetic examples that help overcome the limitations posed by small sample sizes. We assess the utility of these synthetically generated images by incorporating them into our training sets and comparing performance against sets augmented with traditional image manipulation techniques. This approach directly addresses **RQ1**, which focuses on improving the diagnostic accuracy and robustness of AI models through advanced data augmentation methods. By using GAN-generated synthetic spiral images, the dataset becomes more diverse and representative, thereby enhancing the ability of machine learning algorithms to detect PD-related patterns despite limited real-world data.

The described augmentation strategy not only strengthens model performance but also underscores the critical role of innovative data-enhancement techniques in developing scalable, clinically impactful solutions for PD diagnostics.

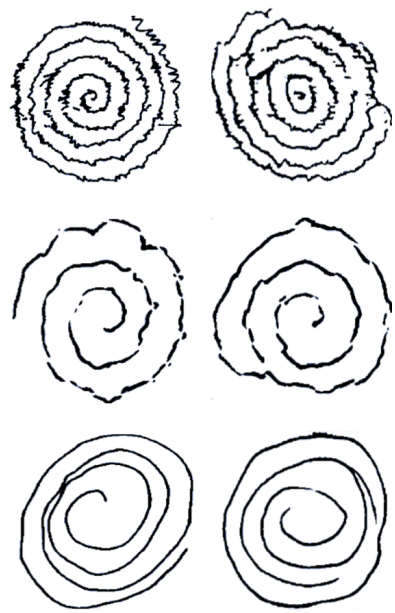


Figure 19: Comparison of the original (left) and the GAN-generated synthesised (right) digital spirals.

2.3.3 Multi-dimensional data representation and transformation

In this section, we extend the concept of embedding dynamic handwriting features in static images, enhancing the representation of handwriting tasks used in the diagnosis of PD. By encoding additional kinematic and pressure-related features into higher-dimensional data representations, we improve the depth of analysis. Subsequently, these enriched datasets are processed using CNNs in one dimension (1D), two dimensions (2D), and three dimensions (3D) to facilitate automated classification and diagnosis. Figure 20 represents a 3D visualisation of data generated from digital drawing tests, specifically focussing on a spiral drawing task that is often used in motor function analysis for PD. In the

given image, the voxelised representation, where individual data points are visualised as cubes, encodes two distinct kinematic parameters to enrich the analysis. The color gradient, ranging from purple to yellow, indicates changes in velocity, which is the rate of displacement over time. Higher velocities are represented by brighter, yellow hues, while lower velocities are shown in darker, purple hues. Such color coding allows for a detailed assessment of velocity variations across different points of the drawing task, providing visual insight into motor control performance, such as smoothness or irregularities that may be symptomatic of PD.

Beyond the color representation, the height (or z-coordinate) of each voxel is determined by the pressure applied at each point during the task. This additional dimension captures the force with which the pen was pressed against the drawing surface. Higher voxels represent greater pressure, while lower ones correspond to lighter touches. By incorporating a third dimension, the model provides a more comprehensive view of how both velocity and pressure fluctuate over the course of the drawing, which is particularly useful in detecting tremor, rigidity, or bradykinesia - common motor symptoms in PD. The voxelised 3D representation not only offers an intuitive way to visualise complex kinematic data but also facilitates the application of advanced machine learning techniques. Specifically, 1D, 2D, and 3D CNNs are employed to classify motor function based on these rich multi-dimensional feature sets. CNNs can detect subtle patterns in velocity, pressure, and temporal evolution, helping to distinguish PD patients from healthy controls. The integration of color and voxel height enhances the representational power of the data, ensuring that maximum information is passed to classification models, thereby improving diagnostic accuracy. This methodology exemplifies the synergy between high-order kinematic feature extraction and advanced computational modeling.

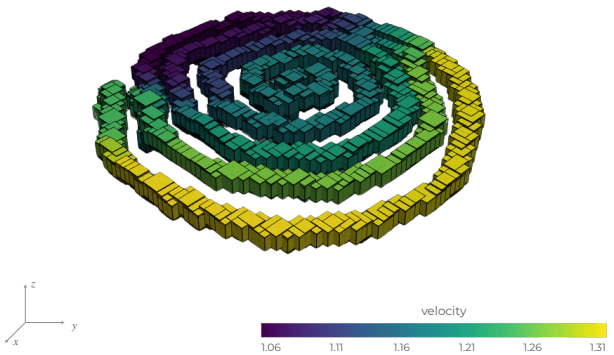


Figure 20: 3D spiral drawing after voxelisation process

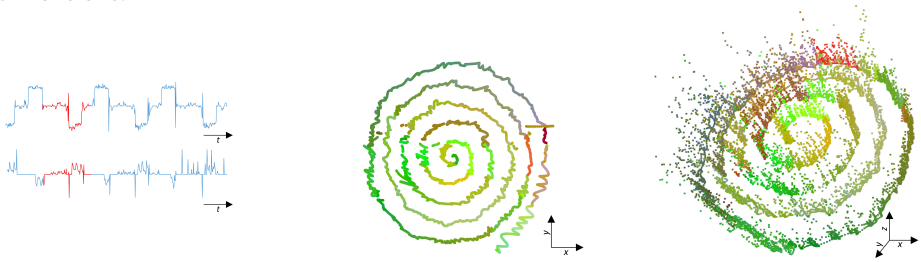
This integration allows for a more robust and detailed analysis, capturing both the static and dynamic aspects of motor symptoms in PD. For example, merging 2D image data with kinematic data from 3D models provides a richer, more nuanced understanding of motor impairments, enhancing the predictive accuracy of diagnostic models.

While Figure 20 provides an illustrative example of how kinematic features like velocity and pressure might be visualised in 3D, the actual feature encoding used in **Publication VI** follows a different methodology, as outlined below. Initially, the raw dataset undergoes preprocessing to standardise feature units through min-max normalisation, ensuring con-

sistency across different dimensional representations. Specifically:

- In the **1D case**, raw dynamic features such as x- and y-coordinates are treated as time series. The timestamp feature is not used directly; instead, it is used to compute velocity, which is then included in the feature set.
- In the **2D case**, the x- and y-coordinates determine pixel positions in RGB images. Azimuth, altitude, and pressure values are encoded as the red, green, and blue channels, respectively, while velocity is represented through line width.
- In the **3D case**, spatial representation includes x- and y-coordinates along with computed velocity, which together define each data point's 3D position. The azimuth, altitude, and pressure features are again encoded as RGB color information in the voxelized space.

Following this, the raw point cloud data is voxelised into a matrix format suitable for CNN analysis, maintaining a fixed grid resolution. It is noteworthy that CNNs exhibit significant feature extraction capabilities; thus, aside from the velocity, no additional hand-crafted features are designed. Figure 21 illustrates the results of data enhancement in various dimensions.



(a) 1D - Time Series *(b) 2D - RGB Image* *(c) 3D - Point Cloud*

Figure 21: Multi-dimensional representation: temporal (a), spatial (b) and volumetric (c) data.

By embedding multi-dimensional data, temporal, spatial, and volumetric, in the process, we capture both static and dynamic aspects of motor function, creating a more enriched dataset for PD diagnosis. Such an integrative approach not only enhances feature representation but also improves the robustness and accuracy of CNN-based predictive models.

2.3.4 Automated segmentation and element analysis in digitised drawing tests

The digitisation of drawing tests enables the capture of precise pen movement parameters, offering insights beyond what is perceptible to the naked eye. To bridge the gap between traditional visual assessment and modern kinematic analysis, this section describes an automated segmentation process introduced in **Publication IV**. By leveraging DL for object detection and classical machine learning techniques for parameter analysis, segmentation identifies individual test elements and evaluates their informativeness at different stages of the task (beginning, middle, and end). This approach enhances the understanding of fine motor performance and contributes to more robust decision support systems for PD diagnosis as part of the data enhancement steps in the ML pipeline.

To further enhance the precision of the analysis, we incorporate automated corner detection within the LAS test. By leveraging the "You Only Look Once" (YOLO) algorithm,

a fast and efficient deep learning-based object detection model developed by Joseph Redmon *et al* [51], we can identify and classify distinct corner types in the drawing task. The segmentation step allows for a more detailed examination of kinematic features associated with specific geometric elements, bridging the gap between visual assessment and data-driven analysis. The detected corners are categorised as follows:

- Upper acute angle corners
- Right angle corners
- Lower acute angle corners

When labelling the corners, a surrounding area is selected rather than just the centre point of the corner. This ensures that sufficient data points are captured, allowing the corners to be analysed separately from the straight and diagonal lines. The process is illustrated in Figure 22, which outlines the overall segmentation workflow. First, the YOLO model is trained on synthetic corner data and subsequently applied to predict corners in real patient data.

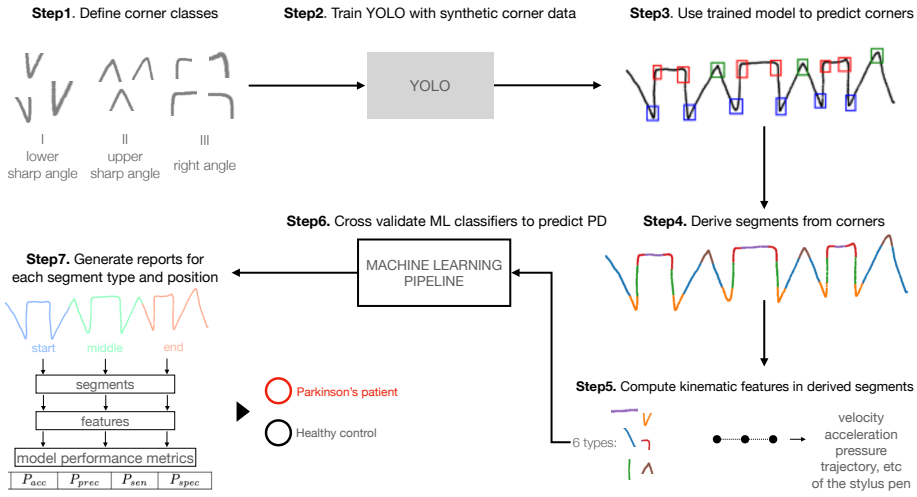


Figure 22: The role of YOLO in the overall workflow. The Step 4 image visualizes the different segment types derived after corner detection. 1: lower acute angle corners (orange), 2: vertical lines (green), 3: right angle corners (red), 4: horizontal lines (purple), 5: diagonal lines (blue), 6: upper acute angle corners (brown).

To enhance the training dataset, three augmentation techniques: horizontal flip, rotation, and shear, were applied, expanding the initial dataset of 90 images to a total of 288 training images. The dataset was split with an 80/20 ratio, resulting in 72 images for training and 18 for validation. YOLO was trained for 300 epochs, with the best performance observed in epoch 237. The model achieved the following metrics:

- **mAP 0.5**: 0.833

- **Precision**: 0.84
- **Recall**: 0.82

The mean average precision (mAP) improved steadily during the first 50 epochs and plateaued after approximately 150 epochs. Precision and recall followed a similar trend, stabilising after around 100 epochs. Although box loss gradually decreased throughout training without overfitting, class loss plateaued for both training and validation with minor improvements in training loss. However, object loss decreased as expected during training but showed noticeable overfitting in the validation set after 150 epochs. Validation in real patient data underscored the importance of selecting an appropriate confidence threshold to mitigate false positives and overlapping bounding boxes. A confidence threshold of 0.5 provided the optimal balance between precision and recall for further tests.

In the subsequent phase of segmenting the $\Pi\Lambda$-tests, the goal is to extract distinct segments from the drawing task. Although YOLO identifies only the corner points, segmenting the entire drawing introduces the challenge of identifying the start and end points of each segment. Relying solely on timestamps is insufficient, as patients can revisit previous sections to correct errors, complicating the segmentation process. To address this, a clustering-based technique is employed, using the x- and y-coordinates of the data points. The methodology operates as follows:

1. A data point is selected and designated as the seed of a new segment.
2. The algorithm identifies the nearest neighbouring point. If its distance from the seed of the segment is within a specified maximum distance threshold, it is added to the segment.
3. The process is iterated until no additional points meet the distance criterion.
4. Then a new data point is chosen as the seed of the next segment and the procedure is repeated.

The described approach effectively simulates the patient's sequential movement during the drawing task, delineating segments that correspond to distinct drawing strokes. An important advantage of this method is its independence from a predefined number of clusters, making it more robust to irregularities or anomalous data points. By grouping spatially close data points, the clustering technique reliably overcomes the primary challenge of segmenting the data of the $\Pi\Lambda$ test while maintaining resilience against drawing inconsistencies. We have established six different segment types based on the results of the YOLO algorithm and additional segmentation analysis. Please refer to Figure 22 Step 4 for a more detailed explanation.

Based on the YOLO results, we can extract the corner coordinates and their associated classes. However, it is crucial to distinguish and classify different types of lines separately. The classification of lines follows these criteria:

- A line connecting two corners at right angles is classified as a horizontal line.
- A line connecting two corners at acute angles is classified as a diagonal line.
- A line connecting one acute angle corner and one right angle corner is classified as a vertical line.

To identify adjacent corners within a segment, we need to find the data points next to each corner. As the remaining segment types consist of lines, we consider the tips of the lines, representing the two points with the maximum distance between them, as the closest points to the corners. To achieve this, we used the convex hull of the data points and derive the maximum distances between them. The next step involves determining the segment types for the transitions from Π to Λ and from Λ to Π. This is accomplished by categorising each lower acute corner as a) Π to Λ, b) Λ to Π. To make this determination, we analyse the preceding segments in the lower acute corner. If the previous segment is a diagonal straight line or an upper acute angle corner, the type is classified as Λ to Π. Conversely, if the previous segment is a horizontal line, a vertical line, or a right angle corner, the type is classified as Π to Λ. Once the segment types for the lower acute corners are determined, all other segment data points are discarded and excluded from further analysis. The subsequent phase involves determining the position of each segment within the test, specifically identifying the start, middle, and end parts. To accomplish this, we locate the minimum and maximum x values from the data set, assigning to each data point its respective position. The test is then divided into three equal length parts on the basis of the x-axis. If the mean x-value of a segment falls within the first part, it is classified as the start segment. If it falls within the second part, it is categorised as a middle segment. In contrast, if the mean value x is located within the third part, it is designated as an end segment. See Figure 22 Step 5 for a visual explanation. The starting segments are coloured blue, the middle segments green, and the end segments red.

This segmentation-based approach directly addresses **RQ3**, which focuses on developing machine learning frameworks that enhance clinical comprehensibility and ensure alignment with real-world workflows. By automatically identifying and classifying distinct segments in digitised drawing tests, clinicians gain a clear, visual mapping of where and how errors occur, making it simpler to link these patterns to specific motor impairments. Consequently, the framework not only maintains accuracy but also promotes an interpretable, step-by-step assessment that can be seamlessly integrated into existing clinical practices.

3 Developed machine learning pipeline for fine motor skill diagnostics

The integration of advanced ML models, ranging from traditional classifiers to state-of-the-art CNNs, exploits the dynamic features of handwriting and drawing tests to identify subtle motor impairments in patients with PD. This section presents the classification models developed in this work, detailing their application, evaluation, and methodological rigor, while emphasising their underlying structures and performance.

Feature selection is a cornerstone of effective model development. The studies presented in this thesis leverage a sophisticated wrapper technique that evaluates subsets of features based on their contribution to classification accuracy [52]. This approach outperforms filter methods by considering interactions between features, capturing attributes that are diagnostically significant only in conjunction with others. To avoid overfitting and ensure unbiased evaluation, a nested cross-validation pipeline is implemented. This approach confines feature selection and model training to the training set, isolating the test set for independent validation. This setup prevents "information leakage", a common issue in traditional cross-validation that can lead to overly optimistic performance evaluations [53]. Traditional ML classifiers are extensively employed due to their robustness and interpretability. Models such as Logistic Regression (LR), Random Forests (RF), Support Vector Machines (SVMs), k-Nearest Neighbours (KNN), Decision Trees (DT) and AdaBoost (AB) are trained using features derived from dynamic drawing parameters like velocity, pressure, and kinematics, employing metrics such as accuracy, precision, sensitivity, specificity, and F1 score to evaluate model performance. These measures help to determine the model's ability to accurately identify both PD patients and healthy controls, which is critical in medical diagnostics. For a detailed breakdown of the sequential steps and methodologies employed in our ML model development, including the critical implementation of nested cross-validation, refer to Figure 23.

The experimental results of the **Publication I** demonstrated significant variability in the performance of classifiers and feature selection strategies (see Tables 7 and 8). For the DraWritePD dataset, the non-nested wrapper-based feature selection combined with the RF classifier achieved the highest accuracy of 92.16%, with sensitivity and specificity values of 90.48% and 93.94%, respectively. Similarly, for the PaHaW dataset, the non-nested wrapper method paired with LR achieved the best performance with an accuracy of 84.86%, a sensitivity of 80.36%, and a specificity of 88.57%. Across both datasets, the wrapper-based methods consistently outperformed filter-based methods, emphasising the importance of feature selection approaches tailored to the classification task. While non-nested feature selection often resulted in higher accuracy, it carried a risk of feature selection bias, highlighting the need for careful validation. Ensemble classifiers, particularly RF and AB, demonstrated strong performance across most configurations, effectively leveraging complex feature interactions. LR also performed robustly in scenarios that included fewer features. Compared to the state-of-the-art methods in Table 9, the proposed framework achieved a higher accuracy in the PaHaW dataset, notably surpassing the benchmarks of [24] and [25] by 20% in some cases of metrics. These findings validate the clinical relevance of the proposed tremor-related features, enriched with angular and differential parameters, and underscore their utility to discriminate between PD patients and HC. The results also highlight the importance of stringent validation protocols, as nested feature selection methods provided more realistic performance estimates compared to their non-nested counterparts. Future work will focus on expanding these findings to additional handwriting tasks and larger, more diverse datasets to ensure generalisability and robustness.

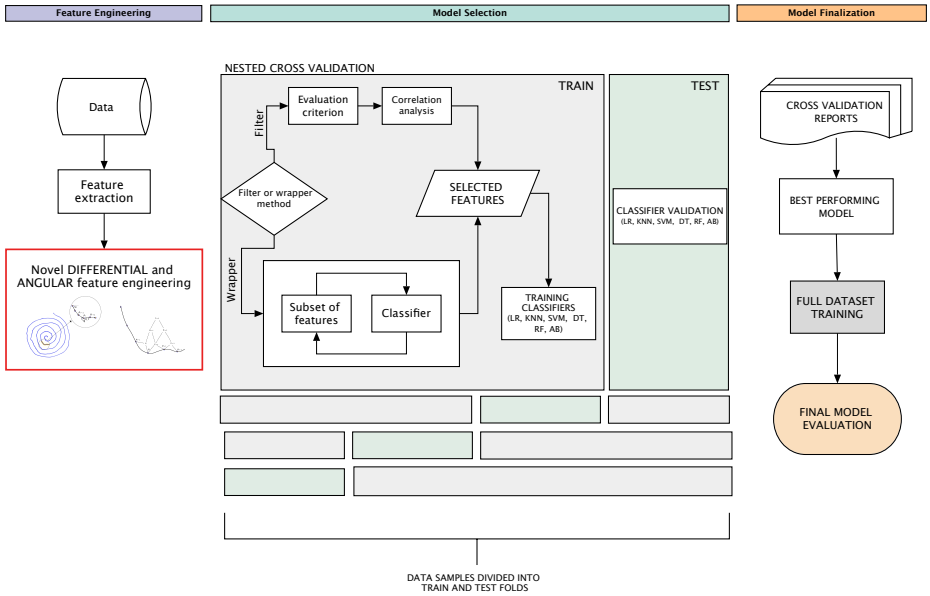


Figure 23: Machine learning pipeline overview with nested cross-validation. In this framework, supervised feature selection strategies are nested within cross-validation iterations to ensure that only the most discriminating features are selected based on the training set, thereby maintaining the integrity of the test set for validation.

Table 7: Classification performance with non-nested and nested feature selection for the DraWritePD dataset. The best scores for each feature selection method are presented in bold.

Feature selection		**Features**	**Classifier**	**Performance metrics**			
				P_{acc}	P_{prec}	P_{sen}	P_{spec}
Wrapper method	non-nested	velocity_median, α_accel_max, pressure_median	LR	90.20%	**95.24%**	80.16%	**96.67%**
			RF	**92.16%**	91.67%	**90.48%**	93.94%
			KNN	80.39%	78.57%	80.16%	81.52%
			SVM	88.24%	90.48%	80.16%	93.94%
			DT	80.39%	76.72%	79.37%	81.52%
			AB	86.27%	86.11%	80.95%	90.61%
	nested	shake_mass, shake_max, snap_mass, crackle_mass, pop_mass	LR	55.33%	25.67%	45.00%	61.67%
			RF	80.33%	80.00%	65.00%	90.00%
			KNN	82.00%	81.67%	70.00%	90.00%
			SVM	84.00%	**86.67%**	**75.00%**	90.00%
			DT	72.67%	65.83%	55.00%	83.33%
			AB	**84.33%**	81.67%	70.00%	**93.33%**

Table 8: Classification performance with non-nested and nested feature selection for the PaHaW dataset. The best scores for each feature selection method are presented in bold.

Feature selection		Features	Classifier	Performance metrics			
				P_{acc}	P_{prec}	P_{sen}	P_{spec}
Wrapper method	non-nested	accel_x_min, α_accel_min, pressure_diff_max, shake_mean, shake_max	LR	**84.86%**	**90.00%**	**80.36%**	**88.57%**
			RF	59.81%	62.61%	52.50%	66.43%
			KNN	68.29%	72.00%	55.00%	80.36%
			SVM	73.62%	78.44%	71.79%	74.29%
			DT	61.24%	63.79%	52.50%	69.64%
			AB	62.86%	63.33%	61.43%	64.29%
	nested	α_velocity_max, α_accel_min, snatch_mean	LR	65.33%	67.33%	60.71%	69.29%
			RF	**73.71%**	**76.62%**	**75.00%**	71.43%
			KNN	63.90%	63.33%	66.79%	60.71%
			SVM	66.76%	70.67%	60.71%	**72.14%**
			DT	66.86%	70.33%	60.36%	71.29%
			AB	64.10%	66.90%	58.57%	69.29%

Table 9: Performance comparison with the state-of-the-art methods based on the Archimedean spiral test from the PaHaW dataset.

	Drotar *et al* (2016)	Impedovo (2019)	Angellilo *et al* (2019)	Present work
non-nested	62.8	97.3	51.3	**84.9**
nested	-	-	53.8	**73.7**

The results of automated segmentation introduced in **Publication IV** and described in Section 2.3.4 are summarised in Tables 10 (a), (b), (c)–11. The segment with the highest performance was the acute angle at the start of the $\Pi\Lambda$-copy task, which achieved an accuracy of 93.8%, precision of 100.0%, sensitivity of 88.3%, and specificity of 100.0% (Table 10 (b)). This result underscores the diagnostic value of acute angles, likely due to their complexity, which may reveal subtle motor impairments in patients with PD. Similarly, vertical lines showed consistently strong predictive power across all tasks. Notably, in the middle of the $\Pi\Lambda$-continue task, vertical lines achieved accuracy of 96.7%, precision of 95.0%, sensitivity of 100.0%, and specificity of 93.3% (Table 10 (c)). These results demonstrate the stability and diagnostic reliability of vertical lines, which consistently outperformed other line types. The $\Pi\Lambda$-copy task stood out with the highest number of informative segments, suggesting that the template's presence may introduce complexity that highlights motor impairments. The $\Pi\Lambda$-trace and $\Pi\Lambda$-continue tasks, while also informative, exhibited slightly lower counts of segments with high predictive power, indicating task-specific nuances. Across all tasks, the performance of different line types varied. Acute angles generally outperformed right angles and diagonals, likely due to their intricate movements, which are more challenging for patients with PD. Horizontal lines, while not as predictive as vertical lines or acute angles, still showed moderate diagnostic potential. Their lower performance might be attributed to their shorter lengths, leading to fewer data points for analysis. These results align with prior findings in [42], which reported an overall accuracy of 91.0%, and with observations in [54], which highlighted the influence of structured templates on motor performance. Interestingly, the position of the segment (start, middle, or end) did not significantly influence the predictive power across tasks. High-performing segments were distributed across positions, suggesting that diagnostic insights are derived more from the type and complexity of the line than its position within the task.

Table 10: Comparison of ΠΛ-trace, ΠΛ-copy, and ΠΛ-continue classification by segment type.

(a) Comparison of ΠΛ-trace classification by segment type

Position	Type	P_{acc}	P_{prec}	P_{sen}	P_{spec}
Start	Vertical	90.5	86.0	100.0	80.0
	Horizontal	77.6	85.0	71.7	86.7
	Acute angle	61.9	65.3	78.3	46.7
	Right angle	77.6	90.0	66.7	93.3
	Diagonal	87.1	91.0	86.7	86.7
Middle	Vertical	81.9	86.7	78.3	86.7
	Horizontal	74.3	88.3	68.3	80.0
	Acute angle	71.4	71.3	83.3	56.7
	Right angle	84.8	82.0	95.0	73.3
	Diagonal	72.4	78.0	71.7	73.3
End	Vertical	93.8	96.0	95.0	90.0
	Horizontal	65.7	76.3	65.0	63.3
	Acute angle	74.8	74.0	90.0	56.7
	Right angle	84.3	93.3	78.3	93.3
	Diagonal	77.6	75.3	86.7	66.7

(b) Comparison of ΠΛ-copy classification by segment type

Position	Type	P_{acc}	P_{prec}	P_{sen}	P_{spec}
Start	Vertical	80.0	79.0	95.0	66.7
	Horizontal	81.0	90.0	78.3	83.3
	Acute angle	93.8	100.0	88.3	100.0
	Right angle	83.8	93.3	81.7	80.0
	Diagonal	87.6	91.0	90.0	83.3
Middle	Vertical	77.1	81.3	81.7	70.0
	Horizontal	80.5	88.3	78.3	86.7
	Acute angle	91.0	91.0	95.0	83.3
	Right angle	84.8	95.0	78.3	93.3
	Diagonal	84.3	85.0	90.0	76.7
End	Vertical	93.8	95.0	95.0	93.3
	Horizontal	82.0	92.0	83.3	80.0
	Acute angle	68.6	75.3	73.3	66.7
	Right angle	78.1	88.3	71.7	86.7
	Diagonal	86.7	86.0	93.3	76.7

(c) Comparison of ΠΛ-continue classification by segment type

Position	Type	P_{acc}	P_{prec}	P_{sen}	P_{spec}
Start	Vertical	62.9	67.0	78.3	46.7
	Horizontal	86.7	90.0	88.3	86.7
	Acute angle	84.0	90.0	86.7	80.0
	Right angle	70.0	74.0	83.3	60.0
	Diagonal	82.0	85.0	88.3	76.7
Middle	Vertical	96.7	95.0	100.0	93.3
	Horizontal	78.0	76.7	86.7	70.0
	Acute angle	78.0	81.7	80.0	76.7
	Right angle	56.0	55.0	73.3	36.7
	Diagonal	82.7	90.0	86.7	80.0
End	Vertical	80.5	76.3	100.0	53.3
	Horizontal	78.7	92.0	76.7	86.7
	Acute angle	66.7	70.0	73.3	56.7
	Right angle	86.7	95.0	81.7	93.3
	Diagonal	78.1	93.3	68.3	93.3

Table 11: ΠΛ-tests transition corners classification

Corner type	Test type	P_{acc}	P_{prec}	P_{sen}	P_{spec}
		Start			
Π to Λ	Trace	75.6	68.0	69.0	81.0
	Copy	68.0	65.6	**80.0**	60.0
	Continue	69.8	66.7	41.7	86.7
Λ to Π	Trace	72.9	65.0	70.0	75.3
	Copy	67.1	63.3	45.0	81.4
	Continue	71.8	68.4	66.7	75.3
		Middle			
Π to Λ	Trace	76.5	77.3	65.0	84.3
	Copy	**86.0**	92.0	75.0	93.3
	Continue	74.4	76.3	63.3	81.3
Λ to Π	Trace	74.0	73.3	55.0	86.7
	Copy	72.7	63.3	65.0	78.6
	Continue	70.6	53.3	36.7	88.7
		End			
Π to Λ	Trace	74.0	79.8	63.0	83.3
	Copy	78.0	84.0	65.0	86.7
	Continue	84.7	**93.3**	66.7	**96.0**
Λ to Π	Trace	81.1	88.3	61.7	92.7
	Copy	74.7	78.0	60.0	84.3
	Continue	71.1	72.7	60.0	76.0

The transition segments, which mark the changes between the Π and Λ configurations, exhibited lower overall performance compared to the other segments (Table 11). These segments showed particularly poor sensitivity, making them less suitable for an accurate diagnosis. For example, in the middle of the ΠΛ-copy task, the Π to Λ transition achieved an accuracy of 86.0% but only a sensitivity of 75.0%. Among the transition segments, the start position was particularly weak, suggesting that transitions alone may not provide reliable diagnostic support.

Figure 24 highlights the most informative segments in tasks, depicted in red. These segments represent areas of highest diagnostic relevance and are error-prone for patients with PD. Visualisation corroborates the quantitative findings, emphasising the importance of acute angles and vertical lines in diagnostic contexts.

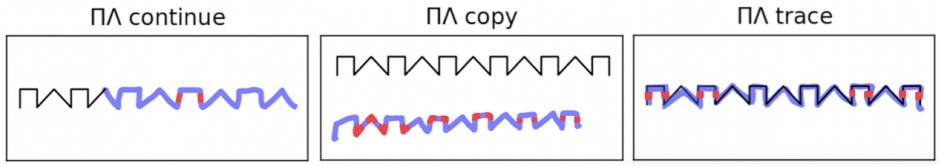


Figure 24: The primary findings highlight the most informative (i.e. error-prone) segments (in red) in Luria's alternating series test for PD.

This analysis highlights the diagnostic potential of specific segments in the ΠΛ tests, particularly vertical lines and acute angles. While transition segments and horizontal lines are less informative, their inclusion can still contribute to a comprehensive diagnostic framework. Future work could explore the interaction between segment type and cognitive factors influencing task performance.

An analysis of movements preceding freezing events in **Publication III** has yielded the following results. Regardless of the number of features, the accuracy ranged from 77.0% to 82.0%, the precision varied between 79.0% and 85.0%, and the sensitivity (recall) was observed in the range of 82.0% to 93.0%. Specificity emerged as the only metric that lag behind, fluctuating between 57.0% and 72.0%. The SVM classifier consistently exhibited the highest performance in all feature sets. The mean velocity values were included in every feature set, followed by the angular velocity mass and the maximum altitude of the stylus, which appeared in three feature sets.

When examining movements after freezing events, the analysis revealed a higher accuracy (80.0% to 86.0%), precision (84.0% to 88.0%), and specificity (68.0% to 79.0%), while sensitivity remained relatively unchanged. In this case, the SVM classifier again demonstrated superior performance among the classifiers tested. However, the features differed significantly. The standard deviation of velocity along the horizontal axis was included in all feature sets alongside the maximum pen altitude. This pair was followed by the mean value of the angle that describes the change in the directional vector. Overall, post-freezing movements appear to be better suited for analysis compared to pre-freezing events.

To visualize these findings, a graph depicting velocity in a one-second window around freezing episodes is provided:

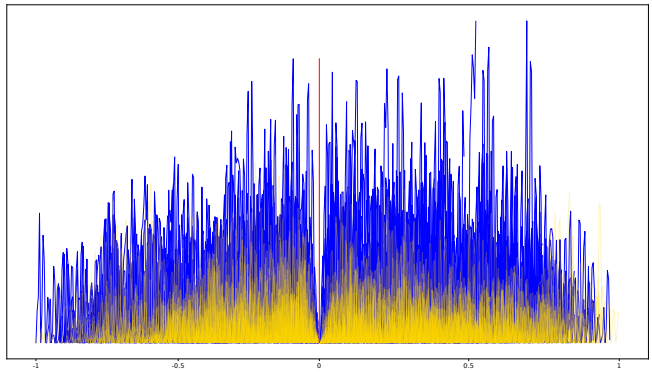


Figure 25: Velocity profiles around freezing episodes for healthy control (HC) subjects (blue) and Parkinson's disease (PD) patients (yellow). The red line indicates the freezing point. The distinct differences in mean values and standard deviations between the groups suggest measurable variations in motion patterns associated with freezing episodes.

In Figure 25, the blue lines represent the velocities observed around freezing episodes in the motions of subjects with HC, while the yellow lines correspond to those observed in patients with PD. The red line marks the freezing point. It is evident that the mean values and standard deviations differ noticeably between these groups.

When the same methodology was applied to the entire sentence, the variability in the metrics that describe model performance increased. The accuracy ranged from 74.0% to 85.0%, the precision from 70.0% to 95.0%, and the sensitivity from 76.0% to 93.0%. However, the specificity rose to 93.0% for certain models with four or five features. Although SVM continued to perform best for models with four features, LR matched its performance with four variables and surpassed it in quality when five features were included. The feature sets in this case more closely resembled those of post-freezing events, with the notable addition of acceleration-based features.

A key challenge in comparing the results of the sentence writing test lies in the variation in languages and sentence lengths. One frequently cited work employing techniques

similar to the present research is [24], where subjects wrote the Czech sentence *Tramvaj dnes už nepojede* (*The tram will not go today*). Unlike the current study, writing was combined with other tasks in a broader testing battery. Nonetheless, the model performance observed in the pre-freezing movement analysis matches the goodness reported by [24], while the analysis of post-freezing movements resulted in models with higher performance.

The study by [55] also investigates sentence writing tests, but its approach focusses on analysing individual letters rather than entire movements. The performance metrics of the model in [55] fall within a similar range: accuracy between 73.0% and 82.0%, precision from 71.0% to 91.0%, and recall from 61.0% to 93.0% (excluding SVM and KNN, which performed poorly). No single model emerged as a definitive winner. Furthermore, [55] identified the mean velocity and the mean acceleration mass of the angular change as the most used features.

While the findings are insufficient to claim that the analysis of freezing episodes is more informative than micrographia or other tests, they demonstrate comparable model performance, making this approach a valuable addition to computer-aided diagnostic support. Furthermore, analysing individual elements of the test proved as informative as analysing the entire test. This conclusion is consistent with [27], which showed that, in drawing tests, specific test segments can be as informative as the full test to diagnose PD.

3.1 Dimensional variants and transfer learning in convolutional neural networks for drawing test classification

CNNs have become a pivotal tool in PD diagnostics, offering unparalleled capacity to analyse raw data and extract hierarchical features. In **Publication VI**, CNN architectures were extended to one, two, and three dimensions to explore their ability to classify digital handwriting tests (Figure 26). Each CNN variant was designed with identical architectures, varying only in the size and dimensionality of the convolution kernels.

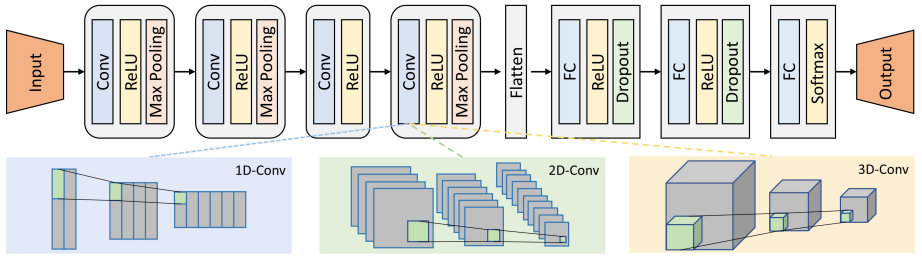


Figure 26: The CNN model employs the same architecture for one-, two-, and three-dimensional convolutional networks, with the only difference being the convolution method used.

The experimental results summarised in Tables 12 and 13 demonstrate the impact of encoding different dynamic features on the diagnostic performance of CNNs across one-, two-, and three-dimensional spaces using the DraWritePD and PaHaW datasets. In general, the addition of more dynamic features improved classification metrics such as precision, sensitivity, specificity, accuracy, and F1 score. In a 1D space, the inclusion of velocity, acceleration, and jerk features progressively improved the performance of the model, with accuracy increasing from 51.67% to 62.56%. However, adding velocity features alone led to minimal improvement, likely due to redundancy with coordinate features. The 2D CNN models showed a marked improvement over the 1D models, achieving a peak accuracy of 80.38% with the inclusion of all dynamic features, including azimuth, altitude, and

pressure. Furthermore, 3D CNNs consistently outperformed lower-dimensional models, achieving the highest diagnostic accuracy of 85.38%, specificity of 87.25%, and F1-score of 85.51% when all dynamic features were included. In particular, 3D CNNs maintained competitive performance even with limited feature sets, highlighting their ability to extract spatial and temporal patterns effectively. These results underscore the advantages of encoding comprehensive dynamic features and using higher-dimensional convolutional operations to improve diagnostic accuracy in distinguishing PD patients from healthy controls.

Table 12: Performance comparison in the DraWritePD dataset.

Dimension	Dynamic features								Metrics (in %)				
	x	y	a	l	p	v	c	j	P_{prec}	P_{sen}	P_{spec}	P_{acc}	P_{f_1}
1D	✓	✓							50.50	53.32	50.25	51.67	52.03
	✓	✓				✓			51.67	61.75	52.67	56.93	54.51
	✓	✓	✓	✓	✓				52.25	63.32	51.67	57.73	56.67
	✓	✓	✓	✓	✓	✓			59.72	62.50	56.25	59.38	61.03
	✓	✓	✓	✓	✓		✓		63.72	67.25	59.67	62.56	65.21
	✓	✓	✓	✓	✓			✓	60.67	65.32	57.75	58.73	63.45
2D	✓	✓							53.25	56.32	68.67	62.56	58.61
	✓	✓				✓			66.67	62.75	75.25	69.38	67.14
	✓	✓	✓	✓	✓				68.72	75.00	78.67	73.67	72.67
	✓	✓	✓	✓	✓	✓			75.00	76.50	80.00	77.73	76.51
	✓	✓	✓	✓	✓		✓		77.50	78.25	81.75	80.38	79.32
	✓	✓	✓	✓	✓			✓	76.67	74.75	78.67	75.93	77.14
3D	✓	✓				✓			72.25	78.00	80.25	76.58	75.21
	✓	✓	✓	✓	✓	✓			77.50	**86.50**	81.75	82.34	81.45
	✓	✓	✓	✓	✓		✓		**82.50**	82.50	**87.25**	**85.38**	**85.51**
	✓	✓	✓	✓	✓			✓	77.25	86.25	80.00	83.34	81.95

Note: The abbreviations x, y denote the x- and y-coordinate features; and a, l and p are the azimuth, altitude and pressure features, respectively; velocity, acceleration, and jerk are represented by v, c, and j, respectively.

Table 13: Performance comparison in the PaHaW dataset.

Dimension	Dynamic features								Metrics (in %)				
	x	y	a	l	p	v	c	j	P_{prec}	P_{sen}	P_{spec}	P_{acc}	P_{f_1}
1D	✓	✓							50.00	57.14	50.00	53.33	53.33
	✓	✓				✓			53.93	57.14	61.75	56.67	57.33
	✓	✓	✓	✓	✓				58.14	58.48	62.50	60.67	59.14
	✓	✓	✓	✓	✓	✓			59.03	71.43	56.25	63.33	64.58
	✓	✓	✓	✓	✓		✓		62.31	75.71	62.50	64.22	65.29
	✓	✓	✓	✓	✓			✓	57.93	68.73	56.25	60.67	63.92
2D	✓	✓							56.93	75.71	57.25	64.33	63.16
	✓	✓				✓			72.31	81.48	75.00	75.33	73.43
	✓	✓	✓	✓	✓				82.33	71.43	82.50	80.00	78.92
	✓	✓	✓	✓	✓	✓			76.25	85.71	75.50	81.33	80.51
	✓	✓	✓	✓	✓		✓		81.03	84.73	78.75	83.67	82.29
	✓	✓	✓	✓	✓			✓	79.61	83.67	76.25	80.33	79.97
3D	✓	✓				✓			62.31	82.73	61.25	68.67	73.29
	✓	✓	✓	✓	✓	✓			75.93	**90.48**	75.00	82.22	82.50
	✓	✓	✓	✓	✓		✓		**83.61**	87.31	**85.50**	**84.67**	**85.71**
	✓	✓	✓	✓	✓			✓	82.71	85.71	80.25	81.73	81.50

Note: The abbreviations x, y denote the x- and y-coordinate features; and a, l and p are the azimuth, altitude, and pressure features, respectively; velocity, acceleration, and jerk are represented by v, c, and j, respectively.

The results in Table 14 highlight the performance of the proposed methodology (**Publication VI**) compared to state-of-the-art approaches for the PaHaW dataset. Drotár *et al* [24] achieved an accuracy of 62.80% using handcrafted features combined with an SVM classifier, serving as a traditional benchmark. Diaz *et al* [23] demonstrated significant advancements using DL techniques; their approach with 1D CNN-extracted features and a

hybrid 1D CNN + BiGRU model achieved the highest accuracy of 93.75%. In contrast, the present work focused on 1D, 2D, and 3D CNNs trained on CNN-extracted features. The 1D CNN achieved an accuracy of 64.22%, which is slightly better than the traditional hand-crafted approach but falls short of advanced hybrid models. However, the accuracy improved significantly with 2D CNNs (83.67%) and 3D CNNs (84.67%), showcasing the benefits of higher-dimensional feature representations. These results underscore the strength of leveraging 3D convolutional architectures to capture spatial and temporal information more effectively, narrowing the gap with the state-of-the-art performance of hybrid models. The findings also demonstrate that higher-dimensional feature extraction is a critical factor in improving classification accuracy in Parkinson's diagnostics.

Table 14: Comparisons with state-of-the-art works.

Author(s)	Dataset	Features	Models	Accuracy (in %)
Drotár et al.	PaHaW	hand-crafted	SVM	62.80
Diaz et al.	PaHaW	1D CNN-extracted	1D CNN + BiGRU	93.75
Diaz et al.	PaHaW	2D CNN-extracted	2D CNN + SVM	75.00
Present work	PaHaW	1D CNN-extracted	1D CNN	64.22
		2D CNN-extracted	2D CNN	83.67
		3D CNN-extracted	3D CNN	**84.67**

Evaluation of GAN-augmented data presented in **Publication II** and Section 2.3.2 reveals an improvement in CNN-based diagnostic performance compared to baseline and traditional augmentation methods. Using CNN architectures such as AlexNet [56], ResNet50 [57], VGG11 [58], Inceptionv3 [59], and Xception [60], pre-trained on ImageNet [61] and adapted through transfer learning, we handle the classification of images derived from digitised drawing tests. ResNet50 and Xception demonstrated the highest sensitivity of 96.6% when trained with ProjectedGAN-generated images, outperforming all other augmentation techniques. These results underscore the strength of ProjectedGAN in generating synthetic data that effectively capture disease-relevant patterns, aligning with its demonstrated superiority in convergence speed and data efficiency, as reported by its authors [47]. Although StyleGAN-based augmentations showed better specificity, particularly with AlexNet and Inception v3, they fell short of the sensitivity improvements achieved by ProjectedGAN. The findings also highlight the limitations of using unaugmented datasets, as none of the CNN architectures achieved optimal scores without augmentation. Despite the computational cost associated with GAN training, the results emphasise the potential of GAN-based augmentation in addressing the challenges posed by limited labelled data, particularly in the context of PD diagnostics. The addition of synthetic GAN-generated images not only improved sensitivity scores by up to 5.7%, but also demonstrated the importance of using advanced enhancement techniques to improve the robustness and reliability of deep learning models in clinical applications.

*Table 15: Test dataset results. Mean scores over five runs. The values in **bold** indicate the best results for a model. Sn - Sensitivity, Sp - Specificity.*

Augmentation method	AlexNet		ResNet50		VGG11		Inception v3		Xception	
	P_{sen}	P_{spec}	P_{sen}	P_{spec}	P_{sen}	P_{spec}	P_{sen}	P_{spec}	P_{sen}	P_{spec}
None	88.0	73.1	94.3	68.0	92.6	66.9	92.0	69.1	90.9	65.7
Traditional	**91.4**	72.0	93.7	**76.0**	89.7	**73.1**	92.6	**76.6**	93.1	**72.6**
StyleGAN2-ADA	85.1	**75.4**	90.9	71.4	94.3	68.0	90.3	**76.6**	93.7	68.0
StyleGAN2-ADA + LeCam	88.6	68.6	95.4	69.1	93.7	66.9	**94.3**	69.1	95.4	63.4
StyleGAN3	88.0	73.1	88.0	73.1	92.0	65.7	93.1	68.0	91.4	65.7
Projected GAN	90.3	68.0	**96.6**	69.7	**95.4**	65.1	92.6	66.3	**96.6**	59.4

The results in Table 15 illustrate the sensitivity and specificity achieved by various CNN

models under different augmentation methods. Without augmentation, ResNet50 exhibited the highest sensitivity (94.3%), while AlexNet showed the best specificity (73.1%). Traditional augmentation improved specificity for most models, with Inception v3 achieving the highest value (76.6%), though sensitivity improvements were modest across architectures. StyleGAN2-ADA demonstrated competitive specificity, particularly with AlexNet and Inception v3, but sensitivity scores lagged behind traditional augmentation. StyleGAN2-ADA + LeCam achieved the highest sensitivity with AlexNet (88.6%) and improved specificity for Inception v3 compared to non-augmented data. StyleGAN3 provided results similar to traditional augmentation but did not outperform it. Projected GAN achieved the best sensitivity for ResNet50 and Xception (96.6%) and also the highest sensitivity overall for VGG11 (95.4%). However, its specificity values were generally lower than those achieved with traditional or StyleGAN-based augmentations. These results indicate that GAN-based augmentation, particularly with Projected GAN, enhances sensitivity for specific models, making it a promising tool for improving CNN performance in PD diagnostics. The following are the most informative findings:

1. The addition of GAN-generated images improved the baseline sensitivity score by 1.7-5.7% for four CNN models (ResNet50, VGG11, Inception v3, Xception).

2. The highest sensitivity score (96.6%) was achieved with the combination of ProjectedGAN generated images and pre-trained CNN models of ResNet50 or Xception.

3. Models trained on the original dataset without any augmentation techniques did not achieve top scores in any experimental settings.

4. It can be seen that the overall specificity scores were lower for the majority of settings. A highly specific test is good at excluding most people who do not have the condition. However, minimising the probability of false negatives is more important in this case.

3.2 Evaluating sensitivity-specificity trade-offs across experimental models for Parkinson's disease diagnostics

This section analyses the performance of various experimental models designed to detect PD. Sensitivity and specificity, two critical metrics for evaluating diagnostic accuracy, are compared in different publications. Sensitivity reflects the model's ability to correctly identify individuals with PD, while specificity measures its effectiveness in correctly identifying those without the disease. By examining the distribution of these metrics through violin plots in Figure 27, we highlight the trade-offs, variability, and overall performance of each model, providing insight into their strengths and limitations.

The violin plots illustrate the distribution of sensitivity (red) and specificity (blue) across different experimental models, represented as **Publications I, II, III, IV and VI**. Each plot captures the density of model performance scores, reflecting the range of results observed under varying configurations or experimental conditions. Each data point within the violin plot represents a single experiment or model configuration's result in terms of sensitivity or specificity.

Publication II stands out as a top performer due to its high median values and consistent results, making it potentially the most reliable model among those tested. This highlights the effectiveness of GAN-based data augmentation combined with CNNs in improving generalisation and robustness in PD classification. The sensitivity values peak between 0.8 and 1.0, demonstrating the model's consistent ability to correctly identify PD

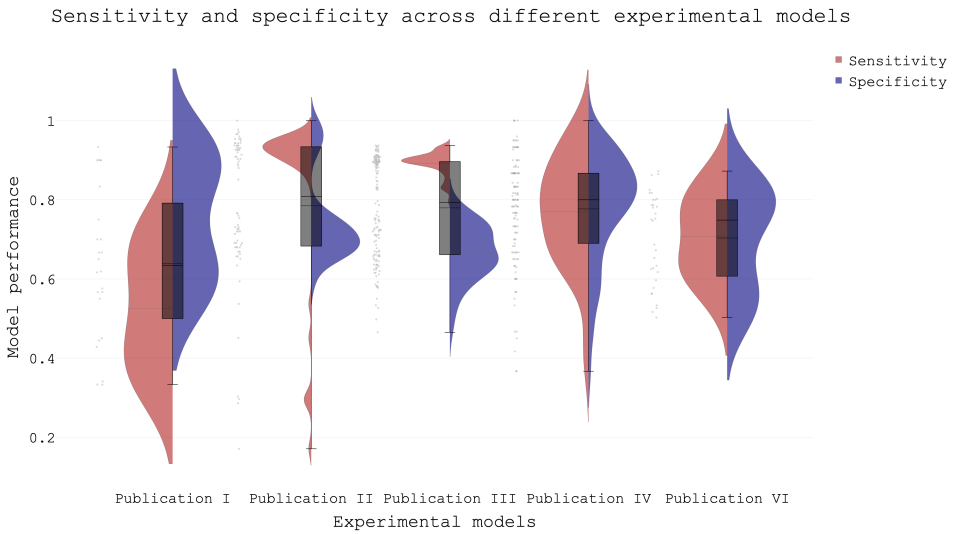


Figure 27: ML models' performance in detecting PD across different experimental settings

cases. The sensitivity violin plot shows a wider spread, indicating a range of sensitivity scores across different configurations. However, the peak of the distribution is between 0.8 and 1, suggesting that many configurations resulted in relatively high sensitivity scores. This concentration towards the higher end of the sensitivity range indicates that the **Publication II** model often achieved strong sensitivity performance. Meanwhile, the specificity plot has its own spread and, although it also shows higher values in some configurations, the peaks are not aligned with those of sensitivity, reinforcing the trade-off between optimising for sensitivity versus specificity. **Publication III**, which focused on microkinematics around freezing episodes, presents a unique distribution pattern. While its sensitivity is somewhat moderate with a broader spread (indicating model instability in some conditions), specificity tends to be higher and more concentrated. This suggests that while the model reliably detects non-PD cases, it may underperform in complex detection scenarios like freezing, possibly due to intra-subject variability in symptom expression. **Publication IV** applies YOLO-based segmentation combined with microkinematic corner feature analysis. The violin plots show tight clustering around high values for both metrics in some configurations, suggesting high precision and discriminative power, particularly within individual handwriting segments. However, there is also variability, indicating sensitivity to segmentation accuracy and feature locality. This model showcases the potential of precise handwriting decomposition when paired with motor descriptors. The combined performance is centered around 0.8, reflecting moderate overall performance with significant variability. **Publication I**, which leveraged high-order derivatives (microkinematics) and nested feature selection, shows more variability in sensitivity than in specificity. The wider spread in sensitivity indicates performance fluctuations across feature subsets, though many configurations still achieved moderate to high accuracy. This highlights the challenge of feature stability, but also underscores the diagnostic potential of microkinematics. In **Publication I**, sensitivity shows a broad distribution with a peak around 0.4 and 0.7, suggesting moderate performance with noticeable variability across configurations. Specificity, however, peaks higher around 0.6 to 0.9, indicating that the models performed better at correctly identifying negatives. This contrast reveals a trade-off where specificity

outperformed sensitivity, pulling the overall performance to a moderate level.

Overall, the results reveal that DL-based models generally outperform classical pipelines, particularly those enhanced by data augmentation **Publication II** and structural decomposition **Publication VI**. The segmentation-based approach in **Publication IV** demonstrates that even classical ML workflows can reach high precision if paired with meaningful spatial decomposition and motor-relevant features. Moreover, the inclusion of microkinematics consistently improves interpretability and diagnostic granularity across models. While GAN-augmented CNNs exhibit the most robust sensitivity and specificity balance, segmentation-based methods suggest promising potential for task-specific refinement. These findings support the hybrid integration of fine-grained motor descriptors with advanced DL frameworks to enhance early PD detection in clinical and real-world environments.

It is important to explore these trade-offs further to determine which metric (sensitivity or specificity) is more critical for the application at hand, and whether you can find a model configuration that offers an acceptable compromise between the two.

3.3 Extending the framework to fatigue detection and analysis

The ML-based methodology presented in this subsection demonstrates the scalability of the PD-focused pipeline, adapting it for the detection and classification of fatigue. Fatigue is a complex and multifactorial condition, characterised by sustained cognitive or physical exertion and commonly observed in both neurological disorders such as PD [62] and broader contexts like transportation, healthcare, and occupational safety. Research estimates that fatigue-related drowsiness contributes to 3.6% of fatal road accidents [63], emphasizing the need for accessible and early detection tools. Despite its widespread impact, fatigue remains underexplored in clinical and technological domains. Its diagnosis is hindered by nonspecific symptoms, scarce labelled data, and the lack of reliable monitoring tools [64, 65]. Fatigue has been associated with reduced physical functioning, diminished quality of life, and poor cognitive performance [66, 67, 68]. These challenges mirror those seen in PD diagnostics and further validate the need for objective and scalable approaches. To address this, we extend our handwriting-based kinematic framework to classify fatigue states using smartphone input. This extension illustrates the versatility of AI-based motor analysis for applications beyond PD, reinforcing the broader feasibility of smartphone-based diagnostics in everyday and clinical contexts.

Accurate labelling and classification of fatigue is particularly challenging, as it integrates subjective self-assessments with measurable behavioral data. Our approach in **Publication V** employs a combination of self-reported questionnaires and motor performance features to define thresholds for fatigue categorisation. As shown in Table 16, fatigue is segmented into binary classes (fatigued, not fatigued) using indicators such as physical or mental exertion levels, sleep duration, and self-assessed tiredness level.

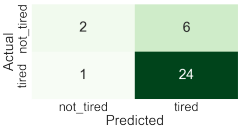
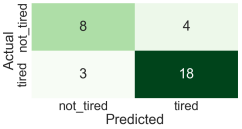
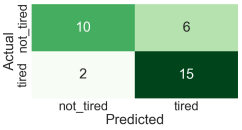
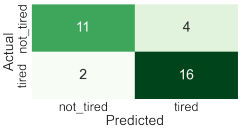
In **Publication VIII**, fatigue categorisation was conducted by distinguishing 'non-fatigued' and 'fatigued' states through sequential mental tasks, where the first session was presumed 'non-fatigued' and the subsequent one 'fatigued.' Self-assessment of fatigue levels underwent iterative refinement, progressively narrowing fatigue class boundaries to enhance ML model performance by increasing decision boundary separation. The features used to train these models, as introduced in Section 2.3, capture intricate patterns of motor activity and behavioural cues. For example, trajectory angles (e.g., ϕ_mass) and micro-accelerations (e.g., crackle_mass) reflect subtle fluctuations in fine motor control that correlate strongly with fatigue states. These features offer valuable insight into the nuanced effects of fatigue on motor performance. The final models, their correspond-

Table 16: Fatigue categories for classification

Fatigue category	Threshold	Label
Physical exertion (PEF)	$= 0$ ≥ 1	Non-fatigued (32) Fatigued (99)
Mental exertion (MEF)	$= 0$ ≥ 1	Non-fatigued (84) Fatigued (47)
Sleep hours (SHF)	≥ 7 ≤ 6	Non-fatigued (62) Fatigued (69)
Self-assessed (SAF)	≤ 3 ≥ 6	Non-fatigued (37) Fatigued (47)

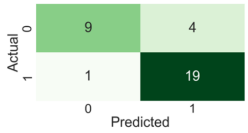
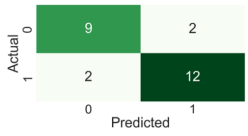
ing features and performance metrics are presented in Tables 17 and 18. The confusion matrices further illustrate their classification accuracy.

Table 17: Best performing machine learning models for fatigue detection using android application data ***Publication V****.*

Fatigue category	Features	Classifier	P_{acc}	P_{sen}	P_{spec}	P_{prec}	P_{f1}	Confusion Matrix
PEF	(ASD) ϕ_mass, crackle_mass	KNN	78.8	96.0	25.0	80.0	87.3	
MEF	(ASD) x_jerk_mass, distance, acceleration	RF	78.8	85.7	66.7	81.8	83.7	
SHF	(RTA) timeFromLastTouch, timeFromFirstCorrectColorRender	RF	75.8	88.2	62.5	71.4	78.9	
SAF	(ALL TESTS) α_velocity_mass, crackle_mass, y_acceleration_mass, α_jerk, γ_acceleration, velocity, y, z, abs, timeFromLastTouch	RF	75.8	83.3	66.7	79.0	84.2	

The six classifiers exhibited comparable performance, with a slight advantage observed for the KNN and RF classifiers. The ASD test, complemented by RTA, emerged as the most informative assessment for detecting fatigue. In addition, for certain fatigue categories determined by self-assessment (SAF), the combination of all tests yielded the most favourable results. Trajectory angles (e.g. ϕ_mass) and micro-changes in acceleration (e.g. crackle_mass) as described in Section 2.3, proved to be highly informative in detecting fatigue. These features capture nuanced fluctuations in fine motor movement, providing valuable insight into the effects of fatigue on motor performance. This high-

Table 18: Best performing ML models for fatigue detection using extended dataset and self-assessed features ***Publication VIII****.*

Fatigue category	Features	Classifier	P_{acc}	P_{sen}	P_{spec}	P_{prec}	P_{f1}	Confusion Matrix
MEF	α_jerk, ϕ_acceleration, yaw_acceleration, jerk, y_l, x_r, physicalWorkScale, effortScale, interestScale, timeFromLastTouch_rts	RF	85.0	86.0	82.0	86.0	84.0	
SAF	α_mass, physicalWorkScale, effortScale, anxietyScale	RF	84.0	86.0	82.0	86.0	86.0	

lights the capability of ML algorithms to discern between these two states based on the study's utilised features, which, although subtle and imperceptible to the naked eye, possess informative value for classification.

The **Publication VIII** achieved a significantly higher accuracy 0f 85.0% compared to previous research 0f 78.8% by incorporating a larger dataset, self-assessed fatigue levels, and hours of mental work as key labels. Features such as anxiety and effort scales, along with angular metrics such as trajectory angles (α_*mass*), were critical contributors to robust model performance. These features captured subtle motor fluctuations related to fatigue, highlighting the potential of ML algorithms to detect nuanced changes in motor performance.

4 Discussion, limitations and prospects for future research

In line with the research gaps outlined in Section 1.1.1, this work specifically addressed: **(A)** the pervasive problem of data scarcity and limited diversity, **(B)** the overlooked importance of robust feature engineering and selection practices, **(C)** the challenges of comprehensibility and clinical adoption of AI-based methods, and **(D)** the limited use of non-invasive and cost-effective data acquisition tools. Three primary research questions guided the investigation:

RQ1: *How can advanced feature engineering techniques and data augmentation methods improve the diagnostic accuracy and robustness of AI models to detect Parkinson's disease?* Many earlier studies on PD diagnostics reported high classification metrics, but these findings were often overoptimistic due to improper validation strategies, including data leakage and non-nested feature selection [26][35]. This thesis addressed these shortcomings by implementing a fully nested pipeline, rigorously separating training and validation to ensure unbiased evaluation. A systematic comparison of feature selection methods further underscored the detrimental impact of non-nested approaches, reinforcing the need for rigorous validation practices in AI-based diagnostics.

Beyond improving validation protocols, this work introduced advanced kinematic features, including high-order derivatives and angular metrics, which outperformed several state-of-the-art methods. While high-order derivatives may appear mathematically abstract, they proved clinically relevant by capturing micro-changes in handwriting kinematics - potential indicators of tremor and fine motor impairment in PD. Multi-dimensional transformations, previously unexplored in PD diagnostics, further enhanced classification performance, demonstrating their utility in handwriting-based assessments.

One of the key contributions of this work was the application of GANs for dataset augmentation. The GAN-augmented dataset not only expanded data diversity but also led to measurable classification improvements, outperforming traditional augmentation techniques. This highlights the potential of synthetic data generation for enhancing model robustness in scenarios with limited real-world samples. Additionally, the generated dataset serves as a valuable resource for further experimentation in PD diagnostics and beyond. However, future research should explore the computational constraints associated with GAN-based augmentation, particularly for deployment in real-time clinical applications.

RQ2: *How can scalable and cost-effective tools, such as smartphone-based applications, transform data collection practices for motor function diagnostics, allowing widespread accessibility and real-world applicability?* Prior efforts to digitise fine motor skill assessment have increasingly explored portable devices, but many still relied on specialised or resource-intensive hardware (e.g., high-resolution graphics tablets or laboratory-grade sensors) [69][63]. Building on existing mobile health and wearable research, this thesis developed smartphone-based data acquisition protocols that capture not only self-reported fatigue levels but also objective fine motor features derived from handwriting and movement patterns, lowering both technical and financial barriers (addressing Gap D). Compared to earlier smartphone-driven solutions [70], which rely primarily on subjective Visual Analogue Scales (VAS) and sleep diaries, the proposed framework incorporates high-order kinematic features to quantify subtle motor fluctuations. By integrating both subjective self-reports and micro-kinematic analysis, this work enhances the precision of fatigue detection. In the domain of fatigue analysis, the proposed framework has shown feasibility for real-world applications, such as monitoring fatigue in critical sectors such as transportation. Future research will investigate medical applications, such as symptom monitoring during chemotherapy or stroke rehabilitation. ML could further enhance fine motor skill tests to track the progression of fatigue or recovery, offering a

valuable tool for patient monitoring.

RQ3: *How can machine learning frameworks be developed to enhance comprehensibility for clinicians, integrate advanced features, and maintain high accuracy, ensuring alignment with clinical workflows and fostering real-world adoption?* A key challenge in AI-assisted diagnostics is ensuring comprehensibility, allowing clinicians to align model predictions with clinically relevant observations. This thesis addresses this issue by leveraging automated segmentation techniques that provide visual and structural mappings between model predictions and specific drawing components (addressing Gap C). The segmentation-based approach introduced in **Publication IV** (detailed in Section 2.3.4) enables the isolation of error-prone drawing segments, allowing for a direct comparison between model-identified difficulty regions and patient-executed trajectories. This method bridges the gap between raw AI predictions and visually interpretable motor impairments, making AI-assisted assessments more accessible for clinical reasoning. The segmentation results, summarised in Tables 10 (a), (b), (c) –11, revealed that specific components of the handwriting, such as acute angles and vertical lines, exhibited the highest predictive power for Parkinson's-related motor impairments. Similarly, vertical lines demonstrated consistently high performance across tasks. These results reinforce the importance of specific kinematic features in distinguishing between healthy and pathological motor function. Furthermore, this work integrates high-order kinematic features that provide a quantitative representation of motor impairments. The ability to link extracted microkinematics to observable motor deficits enhances the clinical utility of AI assessments, making them more interpretable rather than opaque decision systems. By combining segmentation-driven visual explanations with clinically relevant microkinematics, this approach enhances the interpretability of AI-based handwriting assessments. It bridges model predictions with clinically meaningful motor dysfunctions, fostering trust and enabling evidence-based, clinician-guided diagnostics.

While this research has yielded promising results across various areas, several limitations need to be addressed in future studies to enhance the applicability and impact of the findings. One of the primary limitations is the interpretability of DL models, particularly in the context of healthcare. The complexity of models, such as CNNs, often makes it challenging to understand how specific decisions are made. This "black box" nature of AI algorithms poses a significant hurdle in clinical settings where transparency is crucial. To address the interpretability challenge, our future research will focus on developing XAI techniques. We plan to integrate various saliency-based methods, such as Class Activation Mapping (CAM) techniques, to visualise which regions of a drawing or aspects of a kinematic signal are most influential in the decision-making process of a model. Early experiments using these techniques have shown promising results, particularly in PD diagnostics, where critical regions, such as the beginning and end portions of spiral drawings, typically problematic areas for patients, are consistently highlighted. Figure 28 illustrates the comparison between two models used for the diagnosis of PD based on spiral drawing tests. Both models demonstrate high accuracy, with only one misclassification. However, the explainability of their decisions varies significantly, which has important implications for clinical applications. In the first model, CAMs highlight regions extending beyond the spiral drawing, indicating potential overfitting or attention to irrelevant features. While the model performs well with specific test data, its ability to generalise to similar tasks remains uncertain, an important concern in medical diagnostics. In contrast, the second model exhibits activation patterns that closely align with clinical expectations. Its CAM visualisations focus within the spiral, particularly on the end portions, which are known to be challenging for PD patients. This clinically meaningful alignment suggests that the

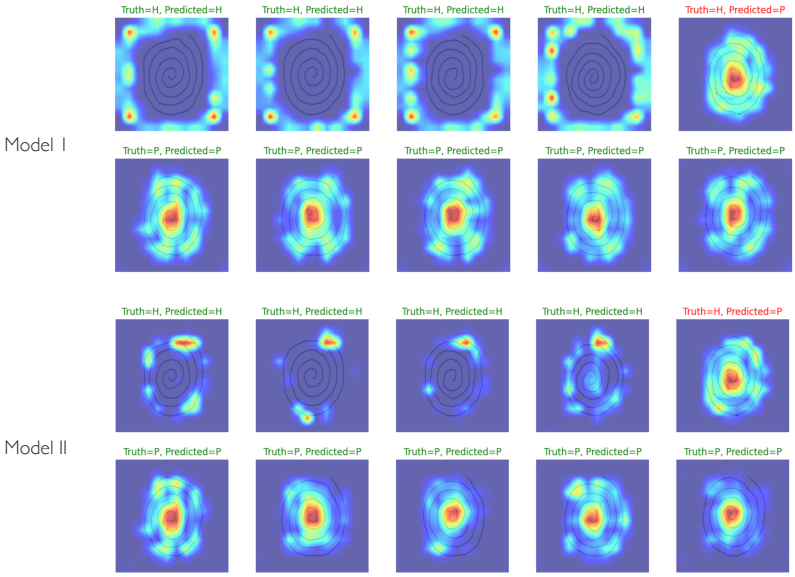


Figure 28: Comparative analysis of model interpretability in PD spiral drawing classification. This image shows the most informative regions highlighted in red by two different models. While both models performed accurately, the second model aligned more closely with neurological expectations, focusing on critical areas of the spiral. The first model, however, showed signs of hallucination, highlighting non-essential regions outside the drawing, indicating potential issues with generalisation.

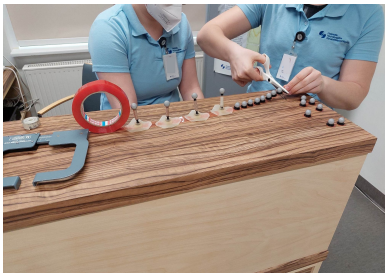
second model is not only accurate but also more interpretable and potentially more reliable for real-world clinical applications. This comparison underscores the importance of using XAI techniques in the development of machine learning models for healthcare care, ensuring that the models are not only accurate, but also transparent and aligned with the evaluations of experts. The alignment between clinical observations and model explanations represents a positive step toward making AI more interpretable and trustworthy in healthcare applications.

The nearly perfect classification performance reported in the literature warrants further scrutiny [23][26]. Clinical observations reveal that patients with early-stage PD under medication may not exhibit significant differences in fine motor skills compared to healthy individuals of similar age. In some cases, they can even outperform elderly individuals without PD. This suggests the absence of a clear categorical distinction between these groups. Future research will test the methodologies described in these studies using the DraWritePD dataset to further explore these findings.

4.1 Exploring gross motor skill assessment as a new research avenue

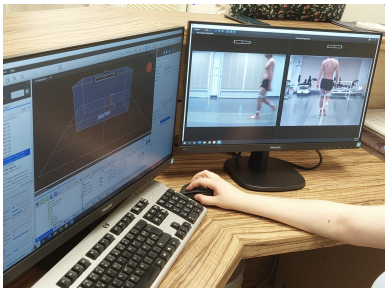
Although fine motor skill assessments such as handwriting and drawing tests offer valuable insight into early detection and monitoring of neurological conditions, a comprehensive evaluation of motor function must also consider gross motor skills. Gross motor assessments, particularly in conditions such as cerebral palsy (CP), provide crucial information on movement and posture abnormalities, which are often overlooked in fine mo-

tor evaluations. This section explores advances in gross motor skill assessments, focussing on gait analysis and how AI-driven solutions are transforming traditional, labour-intensive methods. CP is a group of permanent movement disorders that appear in early childhood, affecting about 2.1 per 1,000 live births [71][72]. CP affects movement and posture due to abnormalities in the developing brain, leading to lifelong disability. Diagnosis and monitoring of CP are challenging due to the variability in symptoms and severity among patients. Traditional diagnostic methods include clinical evaluations and various motor function tests, but these can lack the sensitivity and specificity needed for effective treatment. Marker-based motion capture systems, despite being the gold standard, come with limitations such as high costs, time-intensive setups, and discomfort for patients due to marker attachment [73][74]. Marker placement for gait analysis in patients with CP, which requires 1 to 2.5 hours for setup and additional hours for analysis, demands precision and expertise from clinicians.



(a) Clinicians preparing reflective markers for gait analysis

(b) Utilising a laser for enhanced precision



(c) Comparative analysis of gait dynamics: Video footage and 3D model visualisation

(d) Clinicians attaching reflective markers for gait analysis

Figure 29: Reflective marker setup (a, d), laser alignment for accuracy (b), and 3D model evaluation (c) in HNRC's gait lab.

This process, illustrated in Figure 29, involves extensive measurements and can be uncomfortable for young patients, sometimes prolonging sessions. Additionally, the heat from infrared cameras used in analysis may increase discomfort, adding to the procedure's strain. Acknowledging the challenges of traditional gait analysis, including clinician subjectivity and patient discomfort, researchers are exploring AI-driven, markerless solutions. These advances aim to reduce setup time, costs, and discomfort, enhancing the accuracy and usability of gait analysis for patients with CP.

4.1.1 Advancing markerless gait analysis for cerebral palsy

This section contributes to the broader objective of digitalising motor skill assessments by exploring the use of 3D computer vision for markerless analysis of gait in individuals with cerebral palsy (CP). As detailed in **Publication VII**, this research supports the general objective of the thesis of making data acquisition more accessible and efficient, aligning directly with **RQ2**. The primary goal of this study was to demonstrate how video-based pose estimation techniques can replace traditional markers-dependent methods for gait analysis. CP gait analysis is typically performed with physical markers attached to the patient's body, which can be invasive and uncomfortable. This research introduced a dual-camera setup using advanced pose estimation algorithms such as MediaPipe, OpenPose, Detectron2, HRNet, and Metrabs, which allowed the extraction of key gait parameters, such as joint angles, stride length, and walking speed, directly from video footage.

Creating a comprehensive 3D model of human gait involves a meticulous process that begins with the acquisition of video data from various angles. This is crucial for extracting 2D keypoints using advanced pose estimation algorithms, such as MediaPipe, which capture essential posture details. These keypoints are then triangulated to construct a 3D representation by correlating 2D data points across different camera views (see Figure 30). In order to evaluate the performance of multiple pose estimation frameworks, Figure 31 presents sample outputs from Detectron2, HRNet, OpenPose, Metrabs, and MediaPipe. Each framework applies a distinct approach to detecting and mapping keypoints in human motion, as reflected in the varying densities and precision of the skeletal overlays. For the single-patient context of CP gait analysis, MediaPipe emerged as the most suitable choice, offering robust accuracy, free availability, and near real-time processing without an Nvidia GPU - features that align well with the overarching aim of creating a scalable and cost-effective solution for clinical settings. This work underscores that foot keypoints play a pivotal role in evaluating stride and stance phases, making frameworks like MediaPipe, which provides comprehensive lower-limb coverage, particularly suitable in a clinical context. By enabling robust, near real-time detection without reliance on high-end hardware, MediaPipe aligns with the overarching goal of delivering a cost-effective, scalable approach to markerless gait analysis. To ensure accuracy, the process requires precise camera calibration to determine intrinsic parameters such as focal lengths and optical centres. Calibration typically involves the use of a calibration pattern, such as a chessboard, to fine-tune these parameters, allowing for an accurate overlay of the 2D points into 3D space. This calibration facilitates the calculation of relative positions and orientations of the cameras using geometric transformations based on observed calibration images. Triangulation then synthesises these calibrated measurements into a full 3D model by estimating the spatial relationships and real-world coordinates of the observed keypoints. This methodology not only provides a dynamic visualisation of gait, but also enables detailed kinematic analysis across different planes, sagittal, frontal, and transverse, offering insights into the mechanics of movement during the gait cycle. Such models are invaluable for both clinical evaluations and research on gait abnormalities, providing a rich data set from which to derive quantitative kinematic variables. By removing the need for physical markers, this approach simplifies the data collection process, reducing setup time, and enhancing patient comfort. The digital nature of the system also allows for assessments in more natural environments, further increasing the ecological validity of the results. The study demonstrated the potential for these pose estimation technologies to produce accurate gait metrics comparable to traditional methods, making it a practical solution for clinical settings.

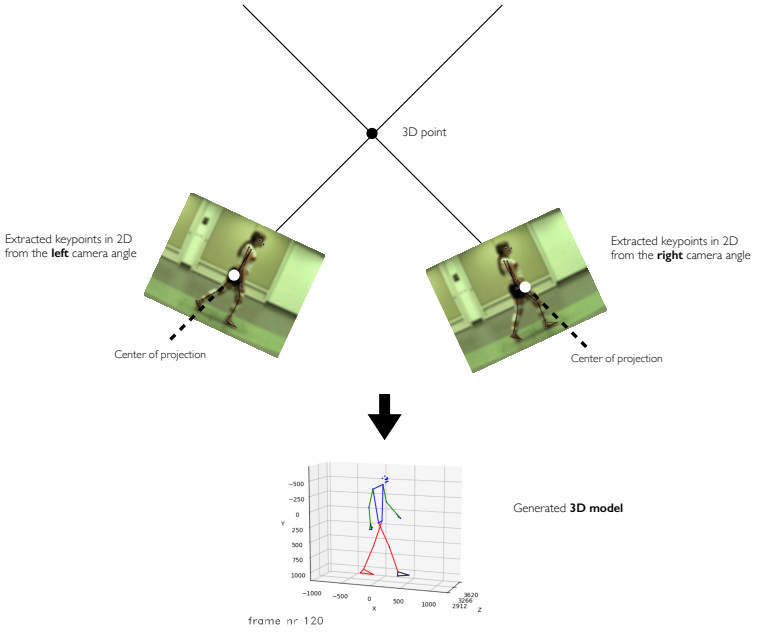


Figure 30: Overview of the 3D gait analysis pipeline, illustrating camera calibration, 2D keypoint detection, and triangulation methods used to build a comprehensive 3D model for accurate kinematic assessments.

Figure 31: Comparison of pose estimation framework outputs: Visualising results from Detectron2, Hrnet, OpenPose, Metrabs, and MediaPipe for evaluating performance and accuracy across different models.

Looking ahead, this research opens the door to future applications of mobile devices in gait analysis. With advances in smartphone camera technology and real-time processing, it is feasible that similar 3D vision-based methods could be integrated into smartphone applications, offering a scalable and cost-effective solution for the continuous monitoring of motor impairments. Such developments would be especially beneficial for conditions like CP, where ongoing monitoring of motor function is critical.

Another promising avenue for future work is the development of a rehabilitation decision support system for CP. Several studies including **Publication VII** have demonstrated success in assessing keypoints and gait dynamics through pose estimation techniques. The next logical step involves leveraging ML for more sophisticated analyses. Ongoing collaborations with Estonian hospitals aim to expand data collection and improve the system. A particular focus will be on leveraging accessible camera systems and smart devices for early diagnosis, which is critical for minimising the long-term impact of CP. Integrating ML into these tools has the potential to significantly enhance neurological assessments.

5 Conclusion

Parkinson's disease (PD) places a heavy burden on healthcare systems due to its lifelong, progressive nature and the absence of a cure. Managing the disease requires continuous medical care, and the wide range of motor and non-motor symptoms makes diagnosis and treatment more challenging. Traditional diagnostic methods, such as clinical evaluations and neuroimaging, often lack sensitivity or are prohibitively expensive, highlighting the need for scalable, objective solutions. In recent years, machine learning based diagnostics have gained significant popularity for their ability to analyse real-time data, detect subtle abnormalities, and potentially streamline patient care. This thesis introduced AI-driven methodologies for fine motor function diagnostics, addressing key challenges in feature engineering, data augmentation, scalable assessments, and clinical adoption. By leveraging high-order kinematic features, GAN-based data expansion, smartphone-integrated motor tests, and machine learning, this work contributes to cost-effective and scalable solutions for neurological assessments.

The methodologies developed in this thesis bridge the gap between laboratory-based diagnostics and scalable digital health solutions, making AI-powered motor assessments more accessible for both clinical and remote applications. The culmination of these efforts has led to several contributions, outlined below:

Contributions

- ✓ Developed novel kinematic and angular feature engineering techniques (e.g., high-order derivatives, angular metrics) to capture subtle motor anomalies indicative of PD. — **Addressed:** Gap B, C | **Publication:** I
- ✓ A framework for GAN-driven augmentation, expanding the variability of handwriting and drawing samples. — **Addressed:** Gap A | **Publication:** II
- ✓ Introduced a smartphone-based dataset for fine motor function assessment. — **Addressed:** Gap A | **Publication:** V, VIII
- ✓ Established a rigorous AI validation pipeline, mitigating common issues like data leakage and non-nested feature selection. — **Addressed:** Gap B | **Publication:** I, III, IV
- ✓ Improved the clinical comprehensibility of handwriting-based diagnostics by automating drawing segmentation, enabling clinicians to visually align AI-identified patterns with symptomatically relevant drawing characteristics. — **Addressed:** Gap C | **Publication:** IV
- ✓ Demonstrated the feasibility of smartphone-based diagnostics for continuous motor function and fatigue assessment, offering a cost-effective, scalable alternative to specialised lab equipment, thereby broadening accessibility in real-world settings. — **Addressed:** Gap D | **Publication:** V

Moving forward, a key priority is enhancing the comprehensibility and real-world utility of AI-driven diagnostics to support evidence-based clinical decisions and personalised therapy. Additionally, ongoing collaborations with healthcare providers will explore pilot implementations of these AI-driven systems across diverse clinical populations. Ensuring generalisability requires rigorous validation beyond controlled datasets, particularly

through multi-institutional studies that assess system performance in real-world clinical settings. Beyond PD diagnostics, this research lays the groundwork for broader applications in motor function assessment. The demonstrated feasibility of smartphone-based monitoring suggests future expansion into continuous, remote tracking of motor impairments. AI-driven gait analysis, particularly using markerless 3D pose estimation, has the potential to transform cerebral palsy rehabilitation and early diagnosis by offering non-invasive, cost-effective alternatives to traditional motion capture. Similarly, fine motor skill assessments may be further adapted for detecting fatigue in high-risk occupations, monitoring rehabilitation progress in stroke recovery, or evaluating neurological decline in aging populations.

Ultimately, this thesis highlights how AI can advance clinical diagnostics, rehabilitation, and long-term patient monitoring. The developed frameworks not only improve diagnostic precision but also emphasise the need for ethically sound, transparent, and interpretable AI models. As AI continues to shape the future of healthcare, its success will depend on aligning with clinical workflows, addressing real-world constraints, and truly serving patient needs.

List of Figures

List of Tables

References

[1] Elli Valla, Sven Nõmm, Kadri Medijainen, Pille Taba, and Aaro Toomela. Tremor-related feature engineering for machine learning based Parkinson's disease diagnostics. *Biomedical Signal Processing and Control*, 75:103551, 2022.

[2] Erik Dzotsenidze, Elli Valla, Sven Nõmm, Kadri Medijainen, Pille Taba, and Aaro Toomela. Generative adversarial networks as a data augmentation tool for CNN-based Parkinson's disease diagnostics. volume 55, pages 108–113. Elsevier, 2022.

[3] Vassili Gorbatsov, Elli Valla, Sven Nõmm, Kadri Medijainen, Pille Taba, and Aaro Toomela. Machine learning based analysis of the upper limb freezing during handwriting in Parkinson's disease patients. volume 55, pages 91–95. Elsevier, 2022.

[4] Elli Valla, Henry Laur, Sven Nõmm, Kadri Medijainen, Pille Taba, and Aaro Toomela. Deep learning based segmentation of Luria's alternating series test to support diagnostics of Parkinson's disease. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1066–1071. IEEE, 2023.

[5] Elli Valla, Ain-Joonas Toose, Sven Nõmm, and Aaro Toomela. Transforming fatigue assessment: Smartphone-based system with digitized motor skill tests. *International journal of medical informatics*, 177:105152, 2023.

[6] Xuechao Wang, Junqing Huang, Marianna Chatzakou, Sven Nõmm, Elli Valla, Kadri Medijainen, Pille Taba, Aaro Toomela, and Michael Ruzhansky. Comparison of one-two-and three-dimensional CNN models for drawing-test-based diagnostics of the Parkinson's disease. *Biomedical Signal Processing and Control*, 87:105436, 2024.

[7] Elli Valla, Gert Kanter, Sven Nõmm, Anton Osvald Kuusk, Peeter Maran, Karl Mihkel Seenmaa, Killu Mägi, and Aaro Toomela. Enhancing cerebral palsy gait analysis with 3D computer vision: A dual-camera approach. In *2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 1352–1357, 2024.

[8] Elli Valla, Lilian Väli, Sven Nõmm, and Aaro Toomela. Smartphone-based microkinematic feature analysis for mental fatigue detection using machine learning. *Computers in Biology and Medicine*, Submitted 2025.

[9] Robert L Nussbaum and Christopher E Ellis. Alzheimer's disease and parkinson's disease. *New england journal of medicine*, 348(14):1356–1364, 2003.

[10] ER Dorsey, A Elbaz, E Nichols, F Abd-Allah, A Abdelalim, JC Adsuar, MG Ansha, C Brayne, JY Choi, D Collado-Mateo, et al. Gbd 2016 Parkinson's disease collaborators. global, regional, and national burden of Parkinson's disease, 1990-2016: a systematic analysis for the global burden of disease study 2016. *Lancet Neurol*, 17(11):939–953, 2018.

[11] J Jankovic. Parkinson's disease: clinical features and diagnosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(4):368–376, 2008.

[12] Runcheng He, Xinxiang Yan, Jifeng Guo, Qian Xu, Beisha Tang, and Qiying Sun. Recent advances in biomarkers for Parkinson's disease. *Frontiers in aging neuroscience*, 10:305, 2018.

[13] Esther Smits, Antti Tolonen, Luc Cluitmans, Mark Gils, Bernard Conway, Rutger C Zietsma, Klaus Leenders, and Natasha Maurits. Standardized Handwriting to Assess Bradykinesia, Micrographia and Tremor in Parkinson's disease. *PloS one*, 9, 05 2014.

[14] Aleksandr Talitckii, Ekaterina Kovalenko, Aleksei Shcherbak, Anna Anikina, Ekaterina Bril, Olga Zimniakova, Maxim Semenov, Dmitry V Dylov, and Andrey Somov. Comparative study of wearable sensors, video, and handwriting to detect Parkinson's disease. *IEEE Transactions on Instrumentation and Measurement*, 71:1–10, 2022.

[15] Zhu Li, Jiayu Yang, Yanwen Wang, Miao Cai, Xiaoli Liu, and Kang Lu. Early diagnosis of parkinson's disease using continuous convolution network: Handwriting recognition based on off-line hand drawing without template. *Journal of biomedical informatics*, 130:104085, 2022.

[16] Zoltan Galaz, Peter Drotar, Jiri Mekyska, Matej Gazda, Jan Mucha, Vojtech Zvoncak, Zdenek Smekal, Marcos Faundez-Zanuy, Reinel Castrillon, Juan Rafael Orozco-Arroyave, et al. Comparison of CNN-learned vs. handcrafted features for detection of parkinson's disease dysgraphia in a multilingual dataset. *Frontiers in Neuroinformatics*, 16:877139, 2022.

[17] Moises Diaz, Momina Moetesum, Imran Siddiqi, and Gennaro Vessio. Sequence-based dynamic handwriting analysis for Parkinson's disease detection with one-dimensional convolutions and bigrus. *Expert Systems with Applications*, 168:114405, 2021.

[18] Ibtissame Aouraghe, Ammour Alae, Khaissidi Ghizlane, Mostafa Mrabti, Ghita Aboulem, and Belahsen Faouzi. A novel approach combining temporal and spectral features of arabic online handwriting for Parkinson's disease prediction. *Journal of Neuroscience Methods*, 339:108727, 2020.

[19] Isabel Sarzo-Wabi, Daniel-Alejandro Galindo-Lazo, and Roberto Rosas-Romero. Feature extraction and classification of static spiral tests to assist the detection of Parkinson's disease. *Multimedia Tools and Applications*, 83(15):45921–45945, 2024.

[20] Elina Kuosmanen, Valerii Kan, Aku Visuri, Simo Hosio, and Denzil Ferreira. Let's draw: Detecting and measuring parkinson's disease on smartphones. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–9, New York, NY, USA, 2020. Association for Computing Machinery.

[21] Julian Varghese, Alexander Brenner, Michael Fujarski, Catharina Marie van Alen, Lucas Plagwitz, and Tobias Warnecke. Machine learning in the Parkinson's disease smartwatch (pads) dataset. *npj Parkinson's Disease*, 10(1):9, 2024.

[22] Iqra Kamran, Saeeda Naz, Imran Razzak, and Muhammad Imran. Handwriting dynamics assessment using deep neural network for early identification of Parkinson's disease. *Future Generation Computer Systems*, 117:234–244, 2021.

[23] Moises Diaz, Momina Moetesum, Imran Siddiqi, and Gennaro Vessio. Sequence-based dynamic handwriting analysis for Parkinson's disease detection with one-dimensional convolutions and bigrus. *Expert Systems with Applications*, 168:114405, 2021.

[24] Peter Drotár, Jiří Mekyska, Irena Rektorová, Lucia Masarová, Zdeněk Smékal, and Marcos Faundez-Zanuy. Evaluation of handwriting kinematics and pressure for differential diagnosis of parkinson's disease. *Artificial intelligence in Medicine*, 67:39–46, 2016.

[25] Maria Teresa Angelillo, Donato Impedovo, Giuseppe Pirlo, and Gennaro Vessio. Performance-driven handwriting task selection for Parkinson's disease classification. In *International Conference of the Italian Association for Artificial Intelligence*, pages 281–293. Springer, 2019.

[26] Donato Impedovo. Velocity-based signal features for the assessment of parkinsonian handwriting. *IEEE Signal Processing Letters*, 26(4):632–636, 2019.

[27] Sven Nomm, Tanel Kossas, Aaro Toomela, Kadri Medijainen, and Pille Taba. Determining necessary length of the alternating series test for parkinson's disease modelling. In *2019 International Conference on Cyberworlds (CW)*, pages 261–266. IEEE, 2019.

[28] Lucas S Bernardo, Angeles Quezada, Roberto Munoz, Fernanda Martins Maia, Clayton R Pereira, Wanqing Wu, and Victor Hugo C de Albuquerque. Handwritten pattern recognition for early Parkinson's disease diagnosis. *Pattern recognition letters*, 125:78–84, 2019.

[29] Antonio Parziale, Rosa Senatore, Antonio Della Cioppa, and Angelo Marcelli. Cartesian genetic programming for diagnosis of parkinson disease through handwriting analysis: Performance vs. interpretability issues. *Artificial intelligence in medicine*, 111:101984, 2021.

[30] Tsung-Lung Yang, Ping-Ju Kan, Chia-Hung Lin, Hsin-Yu Lin, Wei-Ling Chen, and Her-Terng Yau. Using polar expression features and nonlinear machine learning classifier for automated Parkinson's disease screening. *IEEE Sensors Journal*, 20(1):501–514, 2019.

[31] Tsung-Lung Yang, Chia-Hung Lin, Wei-Ling Chen, Hsin-Yu Lin, Chen-San Su, and Chih-Kuang Liang. Hash transformation and machine learning-based decision-making classifier improved the accuracy rate of automated Parkinson's disease screening. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(1):72–82, 2019.

[32] Claudio Loconsole, Gianpaolo Francesco Trotta, Antonio Brunetti, Joseph Trotta, Angelo Schiavone, Sabina Ilaria Tatò, Giacomo Losavio, and Vitoantonio Bevilacqua. Computer vision and emg-based handwriting analysis for classification in parkinson's disease. In De-Shuang Huang, Kang-Hyun Jo, and Juan Carlos Figueroa-García, editors, *Intelligent Computing Theories and Application*, pages 493–503, Cham, 2017. Springer International Publishing.

[33] Claudio Loconsole, Giacomo Donato Cascarano, Antonio Brunetti, Gianpaolo Francesco Trotta, Giacomo Losavio, Vitoantonio Bevilacqua, and Eugenio Di Sciascio. A model-free technique based on computer vision and semg for classification in Parkinson's disease by using computer-assisted handwriting analysis. *Pattern Recognition Letters*, 121:28–36, 2019. Graphonomics for e-citizens: e-health, e-society, e-education.

[34] Donato Impedovo and Giuseppe Pirlo. Dynamic handwriting analysis for the assessment of neurodegenerative diseases: a pattern recognition perspective. *IEEE reviews in biomedical engineering*, 12:209–220, 2018.

[35] Peter Drotar, Jiri Mekyska, Irena Rektorova, Lucia Masarova, Zdeněk Smékal, and Marcos Faundez-Zanuy. Evaluation of handwriting kinematics and pressure for differential diagnosis of parkinson's disease. *Artificial Intelligence in Medicine*, 67:39 – 46, 2016.

[36] Peter Drotár, Jiří Mekyska, Irena Rektorová, Lucia Masarová, Zdenek Smékal, and Marcos Faundez-Zanuy. Analysis of in-air movement in handwriting: A novel marker for parkinson's disease. *Computer methods and programs in biomedicine*, 117(3):405–411, 2014.

[37] Santiago Perez-Lloret, Laurence Negre-Pages, Philippe Damier, Arnaud Delval, Pascal Derkinderen, Alain Destée, Wassilios G Meissner, Ludwig Schelosky, Francois Tison, and Olivier Rascol. Prevalence, determinants, and effect on quality of life of freezing of gait in parkinson disease. *JAMA neurology*, 71(7):884–890, 2014.

[38] Elke Heremans, Evelien Nackaerts, Griet Vervoort, Sarah Vercruysse, Sanne Broeder, Carolien Strouwen, Stephan P. Swinnen, and Alice Nieuwboer. Amplitude manipulation evokes upper limb freezing during handwriting in patients with parkinson's disease with freezing of gait. *PLOS ONE*, 10(11):1–13, 11 2015.

[39] Reza N. Jazar. *Advanced Dynamics. Rigid Body, Multibody, and Aerospace Applications*. John Wiley & Sons, Inc, 2007.

[40] Sven Nõmm and Aaro Toomela. An alternative approach to measure quantity and smoothness of the human limb motions. *Estonian Journal of Engineering*, 19(4):298–308, 2013.

[41] Seth L Pullman. Spiral analysis: a new technique for measuring tremor with a digitizing tablet. *Movement Disorders*, 13(S3):85–89, 1998.

[42] S. Nõmm, K. Bardõš, A. Toomela, K. Medijainen, and P. Taba. Detailed analysis of the Luria's alternating seriestests for parkinson's disease diagnostics. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1347–1352, Dec 2018.

[43] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[44] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data, 2020.

[45] Hung-Yu Tseng, Zhiding Yu Liu, Jia-Bin Huang, and Ming-Hsuan Yang. Regularizing generative adversarial networks under limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7921–7931, 2021.

[46] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021.

[47] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[48] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[49] Mikołaj Binkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018.

[50] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pages 6626–6637, 2017.

[51] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, June 2016.

[52] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422, 2002.

[53] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. *2nd Edition*. Springer Series in Statistics. Springer, 2002.

[54] Lev Semenovich Vygotsky. *The collected works of LS Vygotsky: Problems of the theory and history of psychology*, volume 3. Springer Science & Business Media, 1987.

[55] Aleksei Netšunajev, Sven Nõmm, Aaro Toomela, Kadri Medijainen, and Pille Taba. Parkinson's disease diagnostics based on the analysis of digital sentence writing test. *Vietnam Journal of Computer Science*, 8(04):493–512, 2021.

[56] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[57] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[58] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[59] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015.

[60] François Chollet. Xception: Deep learning with depthwise separable convolutions, 2016.

[61] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[62] J Friedman and H Friedman. Fatigue in parkinson's disease. *Neurology*, 43(10):2016–2016, 1993.

[63] Sanjay Dey, Sami Ahbab Chowdhury, Subrina Sultana, Md Ali Hossain, Monisha Dey, and Sajal K Das. Real time driver fatigue detection based on facial behaviour along with machine learning approaches. In *2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON)*, pages 135–140. IEEE, 2019.

[64] Abhijit Chaudhuri and Peter O Behan. Fatigue in neurological disorders. *The lancet*, 363(9413):978–988, 2004.

[65] Benzi M Kluger, Qing Zhao, Jared J Tanner, Nadine A Schwab, Shellie-Anne Levy, Sarah E Burke, Haiqing Huang, Mingzhou Ding, and Catherine Price. Structural brain correlates of fatigue in older adults with and without parkinson's disease. *NeuroImage: Clinical*, 22:101730, 2019.

[66] Ana M Abrantes, Joseph H Friedman, Richard A Brown, David R Strong, Julie Desaulniers, Eileen Ing, Jennifer Saritelli, and Deborah Riebe. Physical activity and neuropsychiatric symptoms of parkinson disease. *Journal of geriatric psychiatry and neurology*, 25(3):138–145, 2012.

[67] Roy G Elbers, Erwin EH van Wegen, John Verhoef, and Gert Kwakkel. Impact of fatigue on health-related quality of life in patients with Parkinson's disease: a prospective study. *Clinical Rehabilitation*, 28(3):300–311, 2014.

[68] Benzi M Kluger. Fatigue in parkinson's disease. *International review of neurobiology*, 133:743–768, 2017.

[69] M.T. Tarata, W. Wolf, D. Alexandru, D. Georgescu, and M. Serbanescu. P13-9 SEMG derived parameters vs blood oxygen saturation in monitoring neuromuscular fatigue. *Clinical Neurophysiology*, 121:S180, 2010.

[70] Miklos Palotai, Max Wallack, Gergo Kujbus, Adam Dalnoki, and Charles Guttmann. Usability of a mobile app for real-time assessment of fatigue and related symptoms in patients with multiple sclerosis: observational study. *JMIR mHealth and uHealth*, 9(4):e19564, 2021.

[71] Peter Rosenbaum, Nigel Paneth, Alan Leviton, Murray Goldstein, Martin Bax, Diane Damiano, Bernard Dan, Bo Jacobsson, et al. A report: the definition and classification of cerebral palsy april 2006. *Dev Med Child Neurol Suppl*, 109(suppl 109):8–14, 2007.

[72] Karen W Krigger. Cerebral palsy: an overview. *American family physician*, 73(1):91–100, 2006.

[73] Vladimir Medved. *Measurement of human locomotion*. 01 2000.

[74] Matteo Moro, Giorgia Marchesi, Filip Hesse, Francesca Odone, and Maura Casadio. Markerless vs. marker-based gait analysis: A proof of concept study. *Sensors*, 22(5), 2022.

[75] Gurpreet Singh and Sukesha Sharma. Comparative study of various machine learning techniques for parkinson disease detection based on handwriting. In Sandeep Kumar, K. Balachandran, Joong Hoon Kim, and Jagdish Chand Bansal, editors, *Fourth Congress on Intelligent Systems*, pages 1–15, Singapore, 2024. Springer Nature Singapore.

[76] Vincenzo Randazzo, Giansalvo Cirrincione, Annunziata Paviglianiti, Eros Pasero, and Francesco Carlo Morabito. *Neural Feature Extraction for the Analysis of Parkinsonian Patient Handwriting*, pages 243–253. Springer Singapore, Singapore, 2021.

[77] Mahima Sivakumar, A. Hepzibah Christinal, and S. Jebasingh. Parkinson's disease diagnosis using a combined deep learning approach. In *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, pages 81–84, 2021.

[78] S.M. Rhydh Arnab, Mohammad Rakin Uddin, and Atik Jawad. Enhancing Parkinson's disease detection through handwriting analysis: A deep learning approach with segmentation and transfer learning. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*, pages 1–5, 2023.

[79] Sai Vaibhav Polisetti Venkata, Shubhankar Sabat, Chinmay Anand Deshpande, Asiful Arefeen, Daniel Peterson, and Hassan Ghasemzadeh. On-device machine learning for diagnosis of Parkinson's disease from hand drawn artifacts. In *2022 IEEE-EMBS International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 1–4. IEEE, 2022.

[80] Lina Tong, Jiaji He, and Liang Peng. CNN-based pd hand tremor detection using inertial sensors. *IEEE Sensors Letters*, 5(7):1–4, 2021.

[81] Lambert Igene, Anika Alim, Masudul H Imtiaz, and Stephanie Schuckers. A machine learning model for early prediction of parkinson's disease from wearable sensors. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0734–0737, 2023.

[82] Aleksandr Talitckii, Ekaterina Kovalenko, Anna Anikina, Olga Zimniakova, Maksim Semenov, Ekaterina Bril, Aleksei Shcherbak, Dmitry V Dylov, and Andrey Somov. Avoiding misdiagnosis of Parkinson's disease with the use of wearable sensors and artificial intelligence. *IEEE Sensors Journal*, 21(3):3738–3747, 2020.

[83] Dong Jun Park, Jun Woo Lee, Myung Jun Lee, Se Jin Ahn, Jiyoung Kim, Gyu Lee Kim, Young Jin Ra, Yu Na Cho, and Weui Bong Jeong. Evaluation for parkinsonian bradykinesia by deep learning modeling of kinematic parameters. *Journal of Neural Transmission*, 128:181–189, 2021.

[84] Alexandros Papadopoulos, Konstantinos Kyritsis, Lisa Klingelhoefer, Sevasti Bostanjopoulou, K Ray Chaudhuri, and Anastasios Delopoulos. Detecting parkinsonian tremor from imu data collected in-the-wild using deep multiple-instance learning. *IEEE Journal of Biomedical and Health Informatics*, 24(9):2559–2569, 2019.

[85] Clayton Reginaldo Pereira, Silke Anna Theresa Weber, Christian Hook, Gustavo de Rosa, and João Paulo Papa. Deep learning-aided parkinson's disease diagnosis from handwritten dynamics. *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 340–346, 2016.

[86] Claudio Loconsole, Giacomo Donato Cascarano, Antonio Lattarulo, Antonio Brunetti, Gianpaolo Francesco Trotta, Domenico Buongiorno, Ilaria Bortone, Irio De Feudis, Giacomo Losavio, Vitoantonio Bevilacqua, and Eugenio Di Sciascio. A comparison between ann and svm classifiers for Parkinson's disease by using a model-free computer-assisted handwriting analysis based on biometric signals. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2018.

[87] Rosa Senatore, Antonio Della Cioppa, and Angelo Marcelli. Automatic diagnosis of parkinson disease through handwriting analysis: A cartesian genetic programming approach. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 312–317, 2019.

[88] Ishita Gopalakrishnan, Harene Maharajan, Chandrasekar A, and Shanthini S. Utilizing keras model for dynamic drawing analysis in predicting Parkinson's disease through spiral and wave patterns. In *2024 3rd International Conference for Innovation in Technology (INOCON)*, pages 1–6, 2024.

[89] Zhilin Guo, Weiqi Zeng, Taidong Yu, Yan Xu, Yang Xiao, Xuebing Cao, and Zhiguo Cao. Vision-based finger tapping test in patients with Parkinson's disease via spatial-temporal 3D hand pose estimation. *IEEE Journal of Biomedical and Health Informatics*, 26(8):3848–3859, 2022.

[90] Kamila BIAŁEK, Jacek JAKUBOWSKI, Anna POTULSKA-CHROMIK, Monika NOJSZEWSKA, and Anna KOSTERA-PRUSZCZYK. The application of convolutional neural networks in the diagnosis of parkinson's disease on the basis of handwriting samples. *Przeglad Elektrotechniczny*, 2024(4), 2024.

[91] C Kotsavasiloglou, Nikolaos Kostikis, Dimitrios Hristu-Varsakelis, and Marianthi Arnaoutoglou. Machine learning-based classification of simple drawing movements in parkinson's disease. *Biomedical Signal Processing and Control*, 31:174–180, 2017.

[92] Krzysztof Wrobel, Rafal Doroz, Piotr Porwik, Tomasz Orczyk, Agnieszka Betkowska Cavalcante, and Monika Grajzer. Features of hand-drawn spirals for recognition of Parkinson's disease. In *Asian Conference on Intelligent Information and Database Systems*, pages 458–469. Springer, 2022.

[93] Manuel Gil-Martín, Juan Manuel Montero, and Rubén San-Segundo. Parkinson's disease detection from drawing movements using convolutional neural networks. *Electronics*, 8(8):907, 2019.

[94] Eugênio Peixoto Júnior, Italo LD Delmiro, Naercio Magaia, Fernanda M Maia, Mohammad Mehedi Hassan, Victor Hugo C Albuquerque, and Giancarlo Fortino. Intelligent sensory pen for aiding in the diagnosis of Parkinson's disease from dynamic handwriting analysis. *Sensors*, 20(20):5840, 2020.

[95] Vera Miler Jerkovic, Vladimir Kojic, Natasa Dragasevic Miskovic, Tijana Djukic, Vladimir S Kostic, and Mirjana B Popovic. Analysis of on-surface and in-air movement in handwriting of subjects with Parkinson's disease and atypical parkinsonism. *Biomedical Engineering/Biomedizinische Technik*, 64(2):187–194, 2019.

[96] João Paulo Folador, Maria Cecilia Souza Santos, Luiza Maire David Luiz, Luciane Aparecida Pascucci Sande de Souza, Marcus Fraga Vieira, Adriano Alves Pereira, and Adriano de Oliveira Andrade. On the use of histograms of oriented gradients

for tremor detection from sinusoidal and spiral handwritten drawings of people with Parkinson's disease. *Medical & biological engineering & computing*, 59:195–214, 2021.

[97] Cristian D Rios-Urrego, Juan Camilo Vásquez-Correa, Jesús Francisco Vargas-Bonilla, Elmar Nöth, Francisco Lopera, and Juan Rafael Orozco-Arroyave. Analysis and evaluation of handwriting in patients with Parkinson's disease using kinematic, geometrical, and non-linear features. *Computer methods and programs in biomedicine*, 173:43–52, 2019.

[98] Ujjwal Gupta, Hritik Bansal, and Deepak Joshi. An improved sex-specific and age-dependent classification model for parkinson's diagnosis using handwriting measurement. *Computer methods and programs in biomedicine*, 189:105305, 2020.

[99] Jay Chandra, Siva Muthupalaniappan, Zisheng Shang, Richard Deng, Raymond Lin, Irina Tolkova, Dignity Butts, Daniel Sul, Sammer Marzouk, Soham Bose, et al. Screening of Parkinson's disease using geometric features extracted from spiral drawings. *Brain Sciences*, 11(10):1297, 2021.

[100] Shubham Parab, Jerry Boster, Peter Washington, et al. Parkinson disease recognition using a gamified website: machine learning development and usability study. *JMIR Formative Research*, 7(1):e49898, 2023.

[101] Giacomo Donato Cascarano, Claudio Loconsole, Antonio Brunetti, Antonio Lattarulo, Domenico Buongiorno, Giacomo Losavio, Eugenio Di Sciascio, and Vitoantonio Bevilacqua. Biometric handwriting analysis to support Parkinson's disease assessment and grading. *BMC medical informatics and decision making*, 19:1–11, 2019.

[102] Najd Al-Yousef, R Al, R Al, Reem Al-Abdullatif, Felwa Al-Mutairi, and Ouiem Bchir. Parkinson's disease diagnosis using spiral test on digital tablets. *International Journal of Advanced Computer Science and Applications*, 11(5):461–470, 2020.

[103] Aite Zhao, Huimin Wu, Ming Chen, and Nana Wang. A spatio-temporal siamese neural network for multimodal handwriting abnormality screening of Parkinson's disease. *International Journal of Intelligent Systems*, 2023(1):9921809, 2023.

[104] Sergei Zarembo, Sven Nõmm, Kadri Medijainen, Pille Taba, and Aaro Toomela. CNN based analysis of the luria's alternating series test for Parkinson's disease diagnostics. In *Asian Conference on Intelligent Information and Database Systems*, pages 3–13. Springer, 2021.

[105] Sven Nomm, Konstantin Bardos, Aaro Toomela, Kadri Medijainen, and Pille Taba. Detailed analysis of the Luria's alternating seriestests for parkinson's disease diagnostics. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1347–1352, 2018.

[106] 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)*, pages 1–6, Nov 2016.

[107] Fei Teng, Yanjiao Chen, Yushi Cheng, Xiaoyu Ji, Boyang Zhou, and Wenyuan Xu. Pdges: An interpretable detection model for Parkinson's disease using smartphones. *ACM Trans. Sen. Netw.*, 19(4), April 2023.

[108] Decho Surangsrirat, Apichart Intarapanich, Chusak Thanawattano, Roongroj Bhidayasiri, Sitthi Petchrutchatachart, and Chanawat Anan. Tremor assessment using spiral analysis in time-frequency domain. In *2013 Proceedings of IEEE Southeastcon*, pages 1–6. IEEE, 2013.

[109] Quoc Cuong Ngo, Nicole McConnell, Mohammod Abdul Motin, Barbara Polus, Arup Bhattacharya, Sanjay Raghav, and Dinesh Kant Kumar. Neurodiag: Software for automated diagnosis of Parkinson's disease using handwriting. *IEEE Journal of Translational Engineering in Health and Medicine*, 2024.

[110] Georgia Mitsi, Enrique Urrea Mendoza, Benjamin D Wissel, Elena Barbopoulou, Alok K Dwivedi, Ioannis Tsoulos, Athanassios Stavrakoudis, Alberto J Espay, and Spyros Papapetropoulos. Biometric digital health technology for measuring motor function in Parkinson's disease: results from a feasibility and patient satisfaction study. *Frontiers in neurology*, 8:273, 2017.

[111] Nikolaos Kostikis, Dimitris Hristu-Varsakelis, Marianthi Arnaoutoglou, and C Kotsavasiloglou. A smartphone-based tool for assessing parkinsonian hand tremor. *IEEE journal of biomedical and health informatics*, 19(6):1835–1842, 2015.

[112] Dimitrios Iakovakis, K Ray Chaudhuri, Lisa Klingelhoefer, Sevasti Bostantjopoulou, Zoe Katsarou, Dhaval Trivedi, Heinz Reichmann, Stelios Hadjidimitriou, Vasileios Charisis, and Leontios J Hadjileontiadis. Screening of parkinsonian subtle fine-motor impairment from touchscreen typing via deep learning. *Scientific reports*, 10(1):12623, 2020.

[113] Dimitrios Iakovakis, Stelios Hadjidimitriou, Vasileios Charisis, Sevasti Bostantjopoulou, Zoe Katsarou, Lisa Klingelhoefer, Simone Mayer, Heinz Reichmann, Sofia B Dias, José A Diniz, et al. Early Parkinson's disease detection via touchscreen typing analysis using convolutional neural networks. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3535–3538. IEEE, 2019.

[114] Alexandros Papadopoulos, Dimitrios Iakovakis, Lisa Klingelhoefer, Sevasti Bostantjopoulou, K Ray Chaudhuri, Konstantinos Kyritsis, Stelios Hadjidimitriou, Vasileios Charisis, Leontios J Hadjileontiadis, and Anastasios Delopoulos. Unobtrusive detection of Parkinson's disease from multi-modal and in-the-wild sensor data using deep learning techniques. *Scientific reports*, 10(1):21370, 2020.

[115] Luiz CF Ribeiro, Luis CS Afonso, and Joao P Papa. Bag of samplings for computer-assisted parkinson's disease diagnosis based on recurrent neural networks. *Computers in biology and medicine*, 115:103477, 2019.

Acknowledgements

Embarking on this PhD journey four years ago was a leap into the unknown, and as I reflect on this time, I feel immense pride and gratitude for having chosen this path. The topic I have had the privilege to explore—leveraging AI for fine motor skill-based Parkinson's disease diagnostics—has been both intellectually stimulating and deeply meaningful. It is a subject that pushes the boundaries of technology while contributing practical value to the critical healthcare sector. I am sincerely grateful to my supervisors, Prof. Sven Nõmm and Prof. Aaro Toomela, for entrusting me with such an exciting and forward-thinking research topic. Your contributions and guidance are deeply appreciated.

I would also like to extend my heartfelt thanks to my first mentor in AI, Joonatan Samuel, who introduced me to the foundational principles of machine learning and prepared me mentally to take on the challenges this field entails. The knowledge and skills I gained during our work together were pivotal in shaping my career trajectory.

Special thanks are due to my colleague, Alejandro Guerra Manzanares, for generously sharing your expertise and insights, which have been a key part of my growth throughout this PhD. I also want to express my gratitude to my dear friend Heidi Lees - thank you for your insightful feedback and for always being there with your thoughtful perspective.

I owe much to my partner, Tarmo Tamm, who not only sparked my initial interest in machine learning seven years ago but also provided constant encouragement and invaluable guidance throughout this journey. Your foresight on the potential of this field was truly ahead of its time and I am grateful for your unwavering support.

Finally, I would like to thank my parents, both educators, for their endless love and encouragement. Their support has been a solid foundation throughout my life and my academic journey. This work is as much a result of their belief in me as it is of my own efforts.

This research was partially supported by the European Social Fund through the 'ICT programme' project and by the Estonian Science Foundation ETAG through the PRG 2100 research project. Their financial support has been instrumental in making this work possible.

To all the individuals mentioned above, as well as to those whose contributions made a difference in my academic and personal life, I am deeply grateful. This thesis stands as a testament not only to my own perseverance but also to the collective support and inspiration of a truly remarkable community.

Abstract

Leveraging artificial intelligence for microkinematic analysis of fine motor skills in Parkinson's disease detection

This doctoral research explores the transformative potential of artificial intelligence (AI) and digital tools in advancing motor function diagnostics and addressing the challenges of data scarcity, scalability, and clinical adaptability. Central to this work is the integration of advanced feature engineering techniques, machine learning architectures, and smartphone-based applications to improve diagnostic precision for Parkinson's disease (PD) and beyond. Novel tremor-related characterictics were engineered through high-order diffential- and angular-type features to capture micro-movements in handwriting, enhancing early PD detection. Additionally, the thesis tackles the scarcity of diverse, high-quality datasets by employing generative adversarial networks (GANs), which enriched datasets with realistic synthetic variations, improving the robustness and generalisability of diagnostic models. A notable contribution was the development of a comprehensive experimental workflow using smartphone-based tools to assess fine motor skills and fatigue. This feasibility study demonstrated the potential for scalable, cost-effective diagnostics outside clinical settings. By combining structured motor skill tests with self-reported metadata, and applying machine learning techniques, the system achieved high sensitivity in detecting fatigue, highlighting the viability of smartphones as accessible platforms for digital health assessments.

To bridge the gap between AI-driven diagnostics and clinical workflows, interpretable frameworks were developed, including deep learning-based handwriting segmentation and kinematic feature analysis. These tools revealed diagnostically relevant patterns, enabling precise differentiation between PD patients and healthy controls while aligning with clinical expectations.

This research advances AI-driven motor skill diagnostics, providing tangible, clinically relevant solutions. The methodologies developed herein promise to improve early detection, enable continuous patient monitoring, and enhance diagnostic precision for PD and related neurological conditions, ultimately fostering better patient outcomes and supporting widespread, practical adoption of AI in healthcare.

Kokkuvõte

Tehisintellekti rakendamine peenmotoorika mikrokinemaatilises analüüsis Parkinsoni tõve tuvastamiseks

Käesolev doktoritöö käsitleb tehisintellekti (TI) ja digivahendite transformatiivset potentsiaali motoorsete funktsioonide diagnostikas, keskendudes ühtlasi andmenappuse, skaleeritavuse ja kliinilise tõlgendatavuse probleemide lahendamisele. Uurimistöö fookuses on tunnuste inseneerimise tehnikate, masinõppe arhitektuuride ja nutitelefonirakenduste integreerimine Parkinsoni tõve (PD) diagnostilise täpsuse ning praktilise rakendatavuse parandamiseks. Töö raames tutvustati kõrgemat järku diferentsiaalseid ja geomeetrilisi tunnuseid käekirja mikroliigutuste analüüsiks, mis võimaldas täpsemalt tuvastada Parkinsoni tõvele iseloomulikke motoorseid kõrvalekaldeid. Töös käsitleti ka mitmekesiste ja kvaliteetsete andmekogumite nappuse probleemi, rakendades generatiivseid vastanduvaid närvivõrke (GAN), mis täiendasid olemasolevaid andmeid realistlike sünteetiliste variatsioonidega. See parandas diagnostiliste mudelite töökindlust ja üldistamisvõimet erinevates olukordades. Tähtis osa uurimistööst oli ka nutitelefonipõhise hindamisplatvormi loomine, mis võimaldab motoorsete oskuste ja väsimuse skaleeritavat ja kulutõhusat jälgimist. Näiteks töötati välja väsimuse tuvastamise süsteem, mis saavutas kõrge tundlikkuse, kasutades masinõppemudeleid, mis olid treenitud motoorsete testide ja isehinnanguliste metaandmete põhjal tuletatud tunnuste abil.

Töös arendati ka tõlgendatavaid tehisintellektil põhinevaid raamistikke, et ületada lõhe automaatse diagnostika ja kliiniliste töövoogude vahel. Sellised lahendused nagu süvaõppel põhinev käekirja automaatne segmenteerimine ning kinemaatiline tunnuste analüüs võimaldasid tuvastada diagnostiliselt olulisi mustreid, mis parandasid oluliselt Parkinsoni tõve patsientide ja tervete kontrollisikute eristamise täpsust.

Käesolev uurimistöö edendab tehisintellektil põhinevat motoorsete oskuste diagnostikat, pakkudes kliiniliselt asjakohaseid ja tõlgendatavaid lahendusi. Väitekirjas välja töötatud metoodikad parandavad haiguse varajast avastamist, võimaldavad patsientide pidevat jälgimist ning suurendavad Parkinsoni tõve ja teiste neuroloogiliste häirete diagnostilist täpsust. See omakorda toetab paremaid ravitulemusi ning soodustab tehisintellekti laialdasemat ja praktilisemat rakendamist tervishoiusüsteemis.

Appendix 1

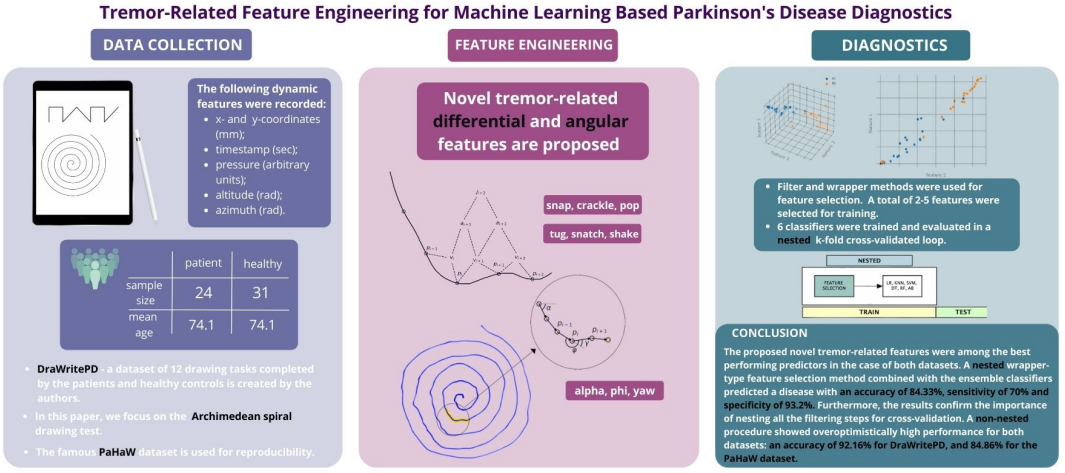
I

Elli Valla, Sven Nõmm, Kadri Medijainen, Pille Taba, and Aaro Toomela. Tremor-related feature engineering for machine learning based Parkinson's disease diagnostics. *Biomedical Signal Processing and Control*, 75:103551, 2022

Graphical Abstract

Tremor-Related Feature Engineering for Machine Learning Based Parkinson's Disease Diagnostics

Elli Valla, Sven Nõmm, Kadri Medijainen, Pille Taba, Aaro Toomela



Biomedical Signal Processing and Control 75 (2022) 103551

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Tremor-related feature engineering for machine learning based Parkinson's disease diagnostics

Elli Valla [a,*], Sven Nõmm [a], Kadri Medijainen [b], Pille Taba [c], Aaro Toomela [d]

[a] Department of Software Science, School of Information Technology, Tallinn University of Technology (TalTech), Akadeemia tee 15a, 12618 Tallinn, Estonia
[b] Institute of Sport Sciences and Physiotherapy, University of Tartu, Ujula 4, Tartu 51014, Estonia
[c] Department of Neurology and Neurosurgery, University of Tartu, Puusepa 8, Tartu 51014, Estonia
[d] School of Natural Sciences and Health, Tallinn University, Narva mnt. 25, 10120 Tallinn, Estonia

ARTICLE INFO

Keywords:
Parkinson's disease
Handwriting database
Machine learning
Decision support system
Tremor

ABSTRACT

Growing research interest has arisen towards the possibility to automatically discriminate between the patients with neurodegenerative disease and healthy controls based on the information extracted from the digital drawing tests.

In this paper, we propose novel higher-order derivative based, angular-type and integral-like features extracted from the Archimedean spiral drawing tests for machine learning based Parkinson's disease diagnostics. The proposed features describe micro-changes in the handwriting trajectory, which are hard or impossible to detect with visual observation. However, they may hold valuable information in terms of tremor-like symptom analysis.

Two datasets are considered in this study: DraWritePD (acquired by the authors) and PaHaW (well known from the literature). A filter (Fisher's score) and wrapper (Recursive Feature Elimination) methods were used for feature selection. Six classifiers were trained and evaluated in a nested cross-validated loop to discriminate between healthy controls and Parkinson's patients.

A nested wrapper-type feature selection method combined with the ensemble classifiers predicted a disease with an accuracy of 84.33%, sensitivity of 70.00% and specificity of 93.20% (DraWritePD), and accuracy of 73.71%, sensitivity of 75.00% and specificity of 71.43% (PaHaW). The non-nested feature selection showed an over-optimistically high performance for both datasets: an accuracy of 92.16% (DraWritePD) and 84.86% (PaHaW).

The proposed novel tremor-related features were among the best performing predictors in the case of both datasets. Furthermore, the results indicate that the nested feature selection procedure plays a significant part in the classification performance.

1. Introduction

Neurodegenerative diseases, such as Alzheimer's and Parkinson's, are a class of neurological disorders where neurons from the central nervous system die or are damaged, causing severe disabilities [1]. Parkinson's disease (PD) is the second most common neurodegenerative disease after Alzheimer's. PD has a prevalence of approximately 0.5 to 1% among persons 65 to 69 years of age, rising to 1 to 3% among persons 80 years of age and older [2]. Several studies have produced evidence that pinpoints neurological disorders as one of the greatest burdens on the healthcare system. The Global Burden of Disease study suggests that 6.2 million patients are diagnosed with PD, and this number will double by 2040, surpassing the growth of Alzheimer's disease [3]. Finding accurate biomarkers for early diagnosis may significantly improve clinical intervention and treatment and can be utilised to monitor the progress of the disease [1,4]. Disorders in motor function performance, such as tremor, bradykinesia (slowness of movement), and rigidity (muscular stiffness), are the cardinal symptoms of PD [1,5]. Tremor occurs in approximately 75% of patients with PD [5–7]. According to the study, [5], early diagnosed PD patients find tremor the second most troublesome symptom, and advanced PD patients (progressed over six years) ranked it as the first motor-related symptom that diminishes the quality

* Corresponding author.
E-mail addresses: elli.valla@taltech.ee (E. Valla), sven.nomm@taltech.ee (S. Nõmm), kadri.medijainen@ut.ee (K. Medijainen), pille.taba@kliinikum.ee (P. Taba), aaro.toomela@tlu.ee (A. Toomela).

https://doi.org/10.1016/j.bspc.2022.103551
Received 18 October 2021; Received in revised form 11 January 2022; Accepted 2 February 2022
Available online 13 February 2022

of their life. Traditionally, medical diagnosis is based on subjective observations from different clinical tests. For example, standardised handwriting tasks can provide quantitative measures for the assessment of tremor [8]. In their classical pen and paper setting, these pure human assessments suffer from several drawbacks: the presence of a subjective component; the limits of human perception, like inability to measure velocity, pressure applied, not to mention derivatives such as jerk, shake, etc.; completion time being the only precisely measurable parameter of the test; and finally, there is no clear definition nor description of errors in these type of assessments. It has been well documented that the digital signals extracted from the handwriting of PD patients are affected and therefore might serve as a diagnostic marker in a computer-aided analysis [9–11]. While existing results may lead highly accurate results [16,19,21] not a lot attention has been paid to the feature selection process, making achieved results less interpretable and attractive for the medical community. While in some papers, features are selected purely based on their discriminating power, many other contributions omit in-depth discussion of the feature selection process. In this paper, we also compare non-nested and nested feature selection processes to confirm the importance of the chosen methods in a classification pipeline (Section 3.3). Based on the digital Archimedean spiral drawing test [22], we demonstrate that the proposed novel tremor-related features possess high discriminating power and provide accurate diagnostics support.

The rest of this paper is organised as follows. Section 1.1 reviews the works related to this problem. Sections 2 and 3 respectively describe the materials and methods used in this research. Sections 4 and 5 report and discuss the experimental results to highlight the effectiveness of the proposed method.

1.1. State of the art

Drawing and writing tests have been used in psychology and neurology for at least a century [23]. The analysis of these handwritten tasks has proven effective in the diagnosis and progression monitoring of PD patients [13,14]. Spiral drawing tests have been frequently used for studying motor control deficits in PD patients [9,12]. It is proved to be a useful tool for assessing tremor-related symptoms [24]. Dynamic methods make use of digital tablets [18], smartpens with axial pressure of ink and tri-axial accelerometers [25], tablet computers [19] and other devices [9]. The main advantage of online acquisition devices is their ability to acquire dynamics of the writing process, which are lost with offline systems. More specifically, dynamic features are the position of the pen (coordinates), pressure (force applied on the writing surface), azimuth (pen orientation), altitude (pen inclination), and timestamp [10]. The most commonly used approaches in the relevant studies can be categorised as follows: numeric methods, where kinematics of the handwriting are analysed [16–19] and deep learning based approach [21,25], where the image or time-series data is extracted and used for classification. In numerical analysis, a typical processing chain involves the pre-processing of raw signals followed by feature extraction and classification. Most proposed methods make use of supervised learning techniques, such as Logistic Regression, AdaBoost, Naive Bayes, SVM, KNN, Random Forests, Decision Trees, LDA. By far the most used classifier in all researched diseases is SVM [9,10,12]. Deep neural networks are not as popular as the more conventional algorithms mentioned before [9,10], but they are starting to gain their popularity [21,25–27].

One of the first contributions describing the results of the digitised drawing test was Marquardt et al. (1994) [15]. More than 30 years of research has resulted in the fact that the original set of four parameters proposed by [15] has grown significantly to hundreds of parameters [16,18,19].

Significant research has been done by Drotar et al. [18], where the various features for the prediction of PD are extracted using digital tablets. Kinematic and pressure parameters were computed from in-air as well as on-surface time intervals. The classification was carried out by applying Support Vector Machine (SVM), and the maximum accuracy of the spiral drawing test was 62.9%. The authors also suggested that classification accuracy depends on the choice of the template. Ensemble of all tests obtained an accuracy of 81.3% on the kinematic and pressure features. The study was conducted on the PaHaW database with a sample size of 75 test subjects.

In the research of Nõmm et al. [28,19] the set of features initially proposed in [29] to measure the quantity and smoothness of the human motions observed during the gross motor activity were adapted for the case of fine motor motions observed during drawing tests. The main distinctive component of [28,19] is the tuple integral-like parameters referred as *motion mass*. On the example of Luria's alternating series tests, it was demonstrated by [19] that motion mass parameters possess higher discriminating power compared to commonly used average values and time duration of the test.

Impedovo (2019) [16] investigates a wide set of velocity-based features for PD patients and healthy controls (HC) discrimination. The extended feature set includes parameters obtained from the Sigma-Lognormal model, the Maxwell–Boltzmann distribution, and the Discrete Fourier Transform applied to the velocity profile of handwriting. The prediction was 97.3% accurate based on the spiral drawing test.

Rios-Urrego et al. (2019) [30] proposed to use geometrical and non-linear dynamic features in addition to kinematic features. It was assumed that these features are able to capture the irregularities of handwriting, which increase as the disease advances. The results showed that the kinematic features were most accurate and that it is possible to discriminate between Parkinson's patients and healthy controls with accuracies up to 83.3%.

Angelillo et al. (2019) [17] investigated the predictive potential of an optimal subset of tasks for an automatised PD diagnosis. First, several features exploiting the dynamics of the handwriting process are extracted from the raw data of different tasks. Then, the predictive potential of each task is evaluated individually. Finally, the best tasks, i.e., those with the highest prediction accuracy, are fed into an ensemble of classifiers whose predictions are obtained via majority voting. Experiments were performed on the PaHaW dataset, as it includes several tasks performed by the same subjects. Non-nested and nested cross-validation performance were compared and analysed. Overall performance degradation was noticeable and therefore concluded the importance of the nested feature selection step in the cross-validation. Poor performance was obtained by the spiral task. The accuracy score of 53.75% was achieved using SVM with RBF kernel, confirming the findings already reported in [18]. Assembling all tasks for classification showed significant improvement in performance. The best performing model with SVM (linear kernel) achieved 88.75% accuracy. It is important to note that it was obtained with non-nested feature selection, as was done in [18,16]. Ensemble of tasks performance with nested feature selection was 66.25% and was obtained with SVM_{RBF}.

Yang et al. (2019) [31] extracted features using polar expressions from the hand-drawn Archimedean spiral and straight-line drawing tests. Features including deviation (cm), accumulation angle (rad), and drawing velocity (cm/s), were engineered and differentiated for normal control groups and groups with Parkinson's disease or essential tremor. The SVM-based classifier showed promising results, with a mean true negative rate (specificity) of 86.79% for identifying normal controls, a mean true positive rate (sensitivity) of 93.72% for identifying PD and ET cases, and a mean hit rate of 90.84% for identifying the correct classes. In [20] hash transformation is used to map polar expression features to a high-dimensional space. The proposed decision-making classifier achieved a higher mean true negative rate (98.96%), mean true positive rate (98.93%), and mean hit rate (98.93%). The important thing to note here is that the difference between the mean age of control subjects and Parkinson's patients who participated in the study is 15.27 (17.57 in [31]) years, and there is no overlap in terms of standard deviation. This is a quite significant difference considering that the performance of

human motor functions declines with age [32–34]. High classification accuracy in their studies may be partly related to the fact that the fine motor skills of Parkinson's patients were impaired both because of the disease and because of age; this fact might have increased the differences between the groups.

For the sake of fair comparison, we included the studies that used statistical machine learning methods, excluding the papers with deep learning algorithms. Other criteria for inclusion are the use of the Archimedean spiral task from the PaHaW dataset and a clear description of the feature selection procedure. For this reason, results of the following studies were chosen for comparison: [16–18].

2. Materials

Two datasets are considered in this research. The first one, here and after referred to as *DraWritePD*, was acquired by the authors. The second dataset, *PaHaW*, is well known from the literature and was kindly provided by the authors of [35,18]. The PaHaW [18] database was used as an additional dataset to test the stability and performance of the proposed method.

2.1. Drawing and handwriting tests for Parkinson's diagnostics, DraWritePD

Data acquisition was performed with an Apple iPad Pro (2016) tablet computer and an Apple Pencil. The tablet has a 26.77 cm (10.5 inches) diagonal. The iPad Pro scans the Apple Pencil's signal with a frequency of 240 points per second. From a software perspective - data was collected using a custom iOS application developed by the research team. The dynamic features (time-sequences) captured by the tablet are as follows: x-coordinate (mm); y-coordinate (mm); timestamp (sec); pressure (arbitrary unit of force applied on the surface: [0,…, 6.0]); altitude (rad); azimuth (rad). Total of 24 PD patients (mean age 74.1 $\pm$ 6.7) and 34 age- and gender-matched healthy control subjects (mean age 74.1 $\pm$ 9.1)) participated in the creation of the database. The overall task was to complete a testing battery consisting of 12 different drawing and writing tests. In this paper, we focus only on the Archimedean spiral test. Few sample images of healthy and Parkinson's patient's spiral drawings are depicted in Fig. 1. The data acquisition process was conducted with the strict guidance of privacy law. The Research Ethics Committee of the University of Tartu (No. 1275T −9) approved the study.

2.2. Parkinson's disease handwriting (PaHaW) database

Data acquisition of the *PaHaW* dataset is described in detail in [18,35]. For the sake of self-sufficiency main properties of this dataset important for the present studies are described in this section. Age and gender distribution of the *PaHaW* dataset is similar to those of *DraWritePD*. The data set consists of 37 PD patients and 38 age- and gender-matched healthy controls (HC). HC subjects have a mean age of 62.4 years (standard deviation 11.3), whereas PD patients have a mean age of 69.3 years (standard deviation 10.9) [18]. During the acquisition of *PaHaW* dataset, each subject was asked to complete a handwriting task according to the prepared pre-filled template at a comfortable speed. Subjects were allowed to repeat the task in case of some error or mistake during handwriting [18]. The handwriting signals were recorded using a Wacom Intuos digitising tablet overlaid with a blank sheet of paper, the sampling rate was set to 100 samples per second. The tablet captured the following dynamic features: x-coordinate; y-coordinate; timestamp; button status; pressure; altitude, and azimuth. All features were converted to the same units as in DraWritePD. The battery of the tasks presented in *PaHaW* dataset differs much from the one employed in *DraWritePD*. However, the Archimedean spiral drawing test is present in both datasets and was thus used in this research.

3. Methods

The methodology used in the present study consists of three main stages: feature engineering, feature selection, and classification. Details of each stage are presented in the following subsections and depicted in Fig. 2.

3.1. Feature engineering

Raw time-series described in Section 2.1: pen position (x- and y-coordinates), timestamp, pen pressure, pen inclination (altitude), and pen orientation (azimuth) can be used to compute an infinite number of features. Tables 1, 2 represent feature classes selected for the present research. Kinematic (displacement, velocity, acceleration, etc.), spatial–temporal (duration, distance), geometric (altitude, azimuth, yaw, etc.), and pressure features were derived. Feature extraction resulted in either a single-valued feature or a vector feature. For all resulting vector features, the following statistical measures were calculated: mean, median, standard deviation, maximum and minimum value. In addition, horizontal and vertical components of the kinematic features were computed. In this research, each subject was instructed to draw a spiral in one stroke; therefore, stroke-related and on/off-screen time values, that are explored in [35,18], are omitted.

Our contribution to this feature engineering is the addition of the higher-order derivatives with respect to time. For instance, given a respective timestamp we can calculate the velocity of the position vector $\overrightarrow{r} = [p_i, p_{i+1}]$. In other words, velocity is the rate at which displacement of the position vector changes with respect to time. Similarly, acceleration was computed as the rate of change in velocity and jerk as the rate of change in acceleration with respect to time. This approach is visualised in Fig. 3. Following the sequence, we considered up to the sixth time derivative of the position vector. There are no universally accepted names for the fourth and higher time derivatives of the displacement. However, the terms snap, crackle, and pop are used in literature for the fourth, fifth, and sixth time derivatives of displacement [37]. The same

(a) HC (b) PD (c) PD

Fig. 1. Sample drawings of an Archimedean spiral performed by a healthy control subject (a) and the PD patients (b, c) from DraWritePD dataset.

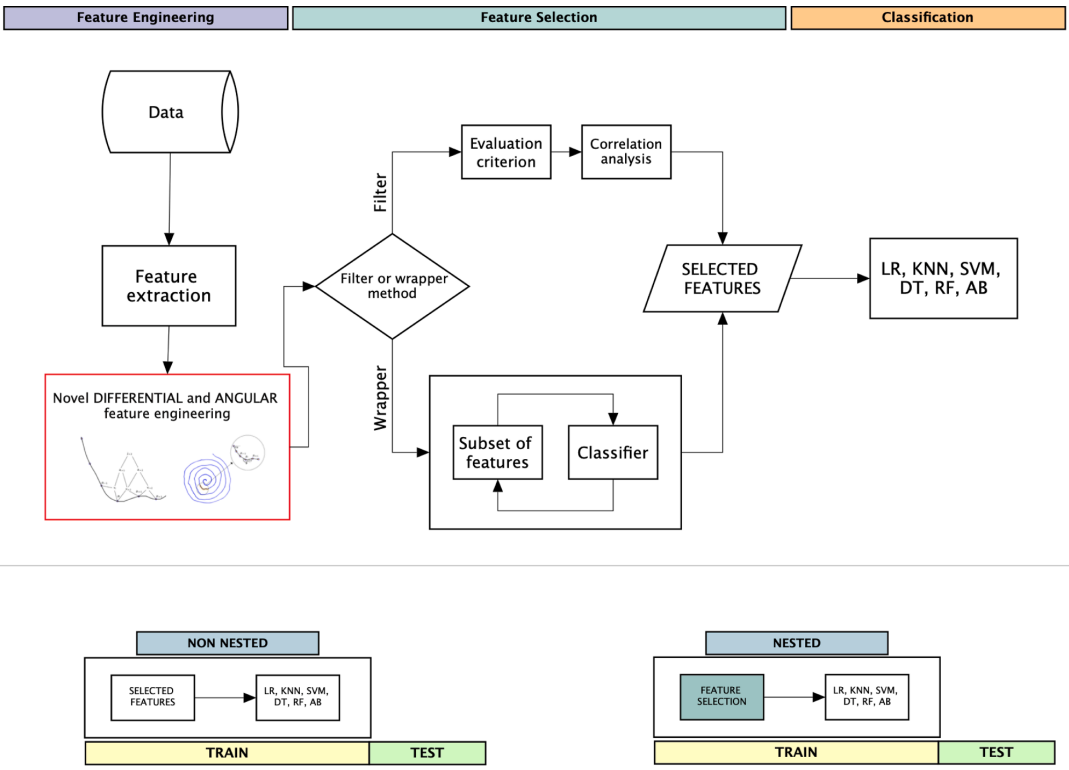


Fig. 2. General methodology. The novel features engineered are described in Section 3.1. Non-nested and nested feature selection visualisation depicts the difference between the two approaches. In the case of the non-nested method, the features are selected based on all the samples and then used for training. The correct way (nested) would be to perform feature selection inside the cross-validation fold using only the samples from the train set, keeping the test set "unseen". [36].

approach was taken for calculating additional pressure parameters. Up to fourth-time derivatives of pressure (force applied on the surface) were computed. These derivatives are referred to as yank, tug, snatch, and shake, respectively [37].

It is worth noting that pen inclination and orientation angles are not widely exploited in related studies. In addition to altitude (pen inclination) and azimuth (pen orientation), we are considering three angles of the position vectors. Pen's trajectory can be observed in Fig. 3 (note that in Fig. 3 magnified part was downsampled for illustration purposes). Given the slope k of the position vector, we can extract angle α. Let N be the number of observation points and (x_i, y_i) are the coordinates of the point p_i, where $i \in \{1, 2, \ldots, N\}$, then the slope (k) and respective angle are represented as follows:

$$k = \frac{y_i - y_{i-1}}{x_i - x_{i-1}}, \tag{1}$$

$$\alpha = \arctan k, \tag{2}$$

Fig. 3 depicts other two angles that are considered in the current research: a rotational angle ϕ (3) and a yaw angle γ (4):

$$\phi_i = \pi + \alpha_{i-1} - \alpha_i \tag{3}$$

$$\gamma_i = \alpha_i - \alpha_{i-1} \tag{4}$$

Yaw is described as the change in direction in which the point vector is pointing. The angular feature set was also enriched with up to third respective time derivatives.

These micro-changes in changes or differential features alongside the angular features of the drawing trajectory are hard or even impossible to detect with the visual observation; however, they may hold valuable information in terms of tremor-like symptom analysis. Tremor associated with the PD is usually reflected by the less smooth motions, which in turn means greater accelerations, more directional changes, and increases in other similar parameters. This symptom is often described as an involuntary quivering, shaky movement and rhythmical declination of trajectory, which gives the basis to relate proposed differential and angular features with the disease. The results indicate that proposed features can be successfully exploited in addition to kinematic and pressure information to enrich feature representation further.

A study conducted by [19] showed that the tuple of integral-like features computed based on kinematic parameters and pressure possess sufficiently high discriminating power to distinguish PD patients from HC control subjects. In [38] it was also demonstrated that these features might allow machine learning techniques to detect mental fatigue; therefore, these features are included in the present research. For the sake of self-sufficiency, the computational procedure of the motion parameters is described in the following subsection.

3.1.1. Motion mass parameters

Motion mass parameters were introduced by [29] to describe the amount and smoothness of motion of a limb or some other group of joints. For each kinematic, geometric, and pressure parameter that changes during the test sum of the absolute values at each observation point may be computed. Let N be the number of observation points in the test (or a part of the test). Denote v_k the velocity along the directional vector of the stylus movement at observation point k where $k \in \{1, \ldots, N\}$

Table 1
Sample subset of vector features.

Feature set	Feature	Description
Spatial-temporal features	displacement	$d_i = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}$
Kinematic features	velocity	Rate of change in displacement with respect to time. First time derivative of the displacement.
	acceleration	Rate of change in velocity with respect to time. Second time derivative of the displacement.
	jerk	Rate of change in acceleration with respect to time. Third time derivative of the displacement.
	snap	Rate of change in jerk with respect to time. Fourth time derivative of the displacement.
	crackle	Rate of change in snap with respect to time. Fifth time derivative of the displacement.
	pop	Rate of change in crackle with respect to time. Sixth time derivative of the displacement.
Pressure features	pressure_diff	Change in pressure between points $[p_i, p_{i+1}]$
	yank	Rate of change in pressure. First time derivative of the force applied on the surface.
	tug	Rate of change in yank. Second time derivative of the force applied on the surface.
	snatch	Rate of change in tug. Third time derivative of the force applied on the surface.
	shake	Rate of change in snatch. Fourth time derivative of the force applied on the surface.
Geometric features	altitude_diff	Change in altitude angle between points $[p_i, p_{i+1}]$
	azimuth_diff	Change in the azimuth angle between points $[p_i, p_{i+1}]$
	alphas_diff	Change in alpha angle between points $[p_i, p_{i+1}]$
	phi_angle_diff	Change in phi angle between points $[p_i, p_{i+1}]$
	yaw_diff	Change in yaw angle between points $[p_i, p_{i+1}]$

then *velocity mass* is defined by equation

$$V_N = \sum k = 1^N |v_k| \tag{5}$$

In the same way, mass parameters are defined for the acceleration - A_N, jerk - J_N, yank Y_N, tug - T_N, snatch - Sn_N and shake - Sh_N. The same logic applies to the pressure (force applied on the surface) and angular parameters describing changes in the stylus direction. In [19] trajectory length and time duration were combined with velocity, acceleration, jerk, and pressure masses into the tuple referred as *motion mass*. The present research adds mass parameters for the higher derivatives of the accelerations and treats them as the features to be used by machine learning methods.

Table 2
Sample subset of single-value features.

Feature set	Feature	Description
Spatial-temporal features	duration	Time interval between first and last time stamp signal
Kinematic features	velocity_mass	Velocity mass of the point vector $[p_1, p_2, \ldots, p_k, \ldots, p_N]$
	acceleration_x_mass	Mass of the x-directional rate of change in velocity
	jerk_median	Median value of the rate of change in acceleration
	snap_mass	Mass of the fifth time derivative of displacement.
Pressure features	shake_median	Median value of the fifth time derivative of pressure (force applied on the surface).
	pressure_diff_min	Minimum difference of pressure between points $[p_i, p_j]$
	tug_mass	Mass of the second time derivative of pressure (force applied on the surface)
Geometric features	ϕ_mass	Mass of the angle ϕ (in radians), see Fig. 3
	α_accel_min	Minimum acceleration of the angle α, see Fig. 3
	yaw_std	Standard deviation of the yaw, see Fig. 3

3.2. Feature selection

A total of 202 features were extracted from the raw signals. Most discriminative predictors were selected to reduce dimensionality and increase interpretability using filter- and wrapper-type feature selection procedures.

Filter methods use independent statistical techniques to evaluate the relationship between a feature predictor and a target variable. They are widely used among bioinformatics researchers due to their straightforward and computationally inexpensive implementation [39]. Among filter models suggested by [40], the Fisher score is the most natural choice due to the numerical representation of the features. For each feature, the Fisher score is computed by the following Eq. (6). Large values of the Fisher's score indicate higher discriminating power and, therefore, better suitability for machine learning classifiers. The algorithm used in this paper returns the ranks of the variables based on Fisher's score in descending order.

$$F = \frac{\sum_{j=1}^{k} p_j (\mu_j - \mu)^2}{\sum_{j=1}^{k} p_j \sigma_j^2} \tag{6}$$

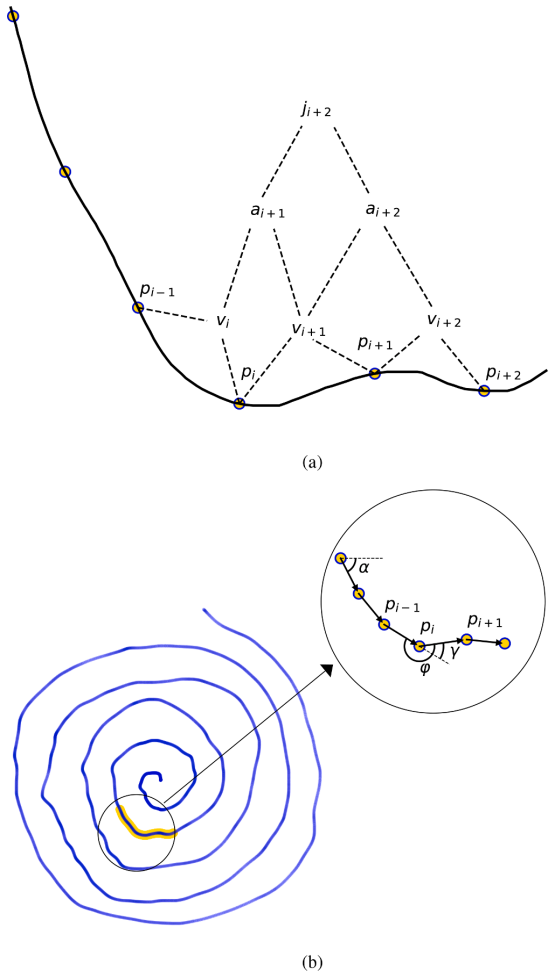


Fig. 3. Visual representation of the differential-type (a) and angular-type (b) features.

Pair-wise Pearson correlations among the sorted features were then computed and explored. A set of features with a Pearson correlation below the threshold of 0.7 was constructed.

Although filter methods are fast and scalable, they have the disadvantage of ignoring the interaction with the classifier. For this reason, wrapper methods are commonly utilised in the feature selection process. In this case, the evaluation is done by training and testing a specific classification model that estimates the relevance of a given subset of features. Wrapper methods are proven to be more efficient but also more computationally expensive [41]. In this paper, we consider the SVM recursive feature elimination (SVM-RFE) wrapper method proposed by [41].

3.3. Statistical classification

In the proposed framework, supervised feature selection strategies are nested within the cross-validation iterations so that the most discriminating features are chosen based only on the training set, while the test set is kept only for validation. The problem with the a priori or so-called non-nested feature selection is that the predictors have an unfair advantage, as they were chosen on the basis of all of the samples. This procedure does not correctly mimic the application of the classifier to a completely independent test set since these predictors "leak the information" from train to test set. [17,36] In other words, a non-nested feature selection may introduce a bias that may lead to overfitting a model and, therefore, to an over-optimistic performance. In this paper, we report results for both nested and non-nested feature selection to analyse this effect. Six machine learning classifiers Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), AdaBoost (AB), were trained and cross-validated in a k-fold loop, $k \in [3,5,10]$. Training and validation of the classifiers were performed using the scikit-learn library [42] for Python.

The performance of the classifiers was reported according to the following metrics: accuracy, precision, sensitivity, and specificity. Accuracy alone does not provide insights into the rate of true positive and true negative predictions by ignoring per-class performance evaluation. Statistical measures: specificity (true negative rate) - refers to the ability of the test to identify subjects without the disease correctly; sensitivity (recall or true positive rate) - refers to the ability of the test to identify those patients with the disease correctly; are therefore widely used in medical diagnostic settings.

4. Results

In this section, we report the results of a series of experiments aimed at evaluating the performance of the proposed workflow.

4.1. Numerical results

In the following Tables 3, 4, the mean accuracy, precision, sensitivity, and specificity values are reported, averaged over all the iterations of a k-fold cross-validation scheme. Feature set remained the same from one fold of cross-validation to another. Low variability is an indication of a stable feature selection algorithm.

The related state-of-the-art results obtained on the PaHaW dataset are depicted in Table 5.

4.2. Discussion

The degradation of the overall performance in the case of the nested feature selection indicates that the model built from the pre-selected features may have been overoptimistic. In addition, in Tables 3 and 4 it can be seen that the wrapper method outperformed the filter-type feature selection algorithm. Proposed tremor-related features were present for all feature selection procedures. For the DraWritePD dataset, the differential features shake, snap, crackle, and pop with their respective *motion mass* parameters selected by a nested wrapper-type feature selection combined with an ensemble classifier, demonstrated a high performance with an accuracy score of 84.33%, sensitivity 70.00%, and specificity of 93.33%. The same methodology performed on the PaHaW dataset showed an accuracy of 73.71%, a sensitivity of 75.00%, and specificity of 71.43%; the selected predictors contained a combination of the differential and angular features with their respective statistical measurements. Although Logistic Regression was among the best performing classifiers in a majority of the settings, in the case of nested wrapper-type feature selection, ensemble algorithms (AdaBoost and Random Forest) showed greater performance.

It is important to note, that Impedovo (2019) [16] and Drotar et al. (2016) [18] used the non-nested validation scheme. This means that, while the results reported by [16] (see Table 5) are remarkable, they still suffer feature selection bias described in Section 3.3.

The nearly perfect classification performance presented in the literature [16,20,21,26,43] is another topic for discussion. Clinical observations show that early-stage patients with Parkinson's disease under medication do not necessarily differ significantly from healthy

Table 3
Classification performance with non-nested and nested feature selection for the DraWritePD dataset. The best scores for each feature selection method are presented in bold.

		Features	Classifier	P_{acc}	P_{prec}	P_{sen}	P_{spec}
Filter method	non-nested	ϕ_mass, duration, pressure_median	LR	84.18%	88.33%	75.00%	90.00%
			RF	**88.18%**	**92.00%**	**80.00%**	**93.81%**
			KNN	78.36%	78.33%	65.00%	87.14%
			SVM	80.36%	88.33%	65.00%	87.14%
			DT	80.18%	72.00%	75.00%	83.81%
			AB	78.36%	71.00%	70.00%	83.81%
	nested	slopes_min, altitude_median, duration	LR	**68.63%**	**55.56%**	34.13%	**90.00%**
			RF	56.86%	33.99%	**61.90%**	56.67%
			KNN	49.02%	25.10%	42.86%	56.67%
			SVM	52.94%	27.06%	44.44%	57.58%
			DT	54.90%	13.73%	33.33%	66.67%
			AB	54.90%	13.73%	33.33%	66.67%
Wrapper method	non-nested	velocity_median, α_accel_max, pressure_median	LR	90.20%	**95.24%**	80.16%	**96.67%**
			RF	**92.16%**	91.67%	**90.48%**	93.94%
			KNN	80.39%	78.57%	80.16%	81.52%
			SVM	88.24%	90.48%	80.16%	93.94%
			DT	80.39%	76.72%	79.37%	81.52%
			AB	86.27%	86.11%	80.95%	90.61%
	nested	shake_mass, shake_max, snap_mass, crackle_mass, pop_mass	LR	55.33%	25.67%	45.00%	61.67%
			RF	80.33%	80.00%	65.00%	90.00%
			KNN	82.00%	81.67%	70.00%	90.00%
			SVM	84.00%	**86.67%**	**75.00%**	90.00%
			DT	72.67%	65.83%	55.00%	83.33%
			AB	**84.33%**	81.67%	70.00%	**93.33%**

Table 4
Classification performance with non-nested and nested feature selection for the PaHaW dataset. The best scores for each feature selection method are presented in bold.

		Features	**Classifier**	P_{acc}	P_{prec}	P_{sen}	P_{spec}
Filter method	non-nested	shake_mean, accel_x_mass, α_accel_max, shake_median, α_accel_min	LR	**79.17%**	**80.67%**	**77.78%**	**80.56%**
			RF	69.44%	72.50%	66.67%	72.22%
			KNN	61.11%	58.63%	75.00%	47.22%
			SVM	63.89%	62.16%	80.56%	47.22%
			DT	62.50%	62.63%	66.67%	58.33%
			AB	66.67%	67.64%	69.44%	63.89%
	nested	pressure_diff_min, shake_mean, shake_median	LR	**65.28**%	55.00%	67.50%	**65.00%**
			RF	58.33%	**61.64%**	63.89%	52.78%
			KNN	56.94%	42.59%	52.78%	61.11%
			SVM	50.00%	33.33%	66.67%	33.33%
			DT	55.56%	58.28%	**75.00%**	36.11%
			AB	52.78%	55.93%	69.44%	36.11%
Wrapper method	non-nested	accel_x_min, α_accel_min, pressure_diff_max, shake_mean, shake_max	LR	**84.86**%	**90.00%**	**80.36%**	**88.57%**
			RF	59.81%	62.61%	52.50%	66.43%
			KNN	68.29%	72.00%	55.00%	80.36%
			SVM	73.62%	78.44%	71.79%	74.29%
			DT	61.24%	63.79%	52.50%	69.64%
			AB	62.86%	63.33%	61.43%	64.29%
	nested	α_velocity_max, α_accel_min, snatch_mean	LR	65.33%	67.33%	60.71%	69.29%
			RF	**73.71%**	**76.62%**	**75.00%**	71.43%
			KNN	63.90%	63.33%	66.79%	60.71%
			SVM	66.76%	70.67%	60.71%	**72.14%**
			DT	66.86%	70.33%	60.36%	71.29%
			AB	64.10%	66.90%	58.57%	69.29%

Table 5
Performance comparison with the state-of-the-art methods based on the Archimedean spiral test from the PaHaW dataset.

	Drotar et al. (2016)	**Impedovo (2019)**	**Angellilo *et al.* (2019)**	**Present work**
non-nested	62.8	97.3	51.3	**84.9**
nested	–	–	53.8	**73.7**

individuals of comparable age and may perform even better than the fine motor skills of elderly individuals without Parkinson's disease. In other words, there is no clear-cut categorical difference between fine-motor skills of individuals with and without Parkinson's disease. Future work would involve testing methodologies introduced in these studies on our own (DraWritePD) dataset to investigate this conclusion further.

5. Conclusion

The handwriting- and drawing-based computer-aided analysis have the potential to serve as the decision support tool for clinicians in neurodegenerative disease diagnostics. Its successful implementation would play a significant role in reducing the burden on the public health system. Our study proposes a set of tremor-related features to discriminate Parkinson's disease patients (PD) and healthy controls (HC) based on the Archimedean spiral drawing test. More specifically, the set of variables was enriched with angular, differential, and integral-like parameters resulting in a database with over 200 features. To reduce the dimensionality and maximise the model's performance, we applied filter- and wrapper-type feature selection algorithms. As indicated in Table 3, the proposed tremor-related features were among the best performing predictors in the PD and HC classification task. It was also demonstrated that a non-nested feature selection method might lead to over-optimistic results and, therefore, should be avoided in the classification pipeline. These findings were reproduced on the PaHaW dataset, see Table 4. The classification accuracy obtained in the present research is 20% higher compared to those reported in the literature (case of a nested feature selection), which confirms that the novel tremor-related features possess greater discriminating power for the diagnostics of Parkinson's disease.

A couple of remarks have to be recognised when interpreting the results. The proposed framework was conducted based only on the Archimedean spiral test. In our future work, we plan to include other handwriting and drawing tasks. Repeated measurements from the same subjects should be obtained and analysed to confirm the possible relation of the proposed features with tremor-based motor dysfunctions and assess the severity of the symptom. To demonstrate the reliability and generalisability of the proposed framework, we need to perform testing on more extensive and more diverse groups of patients. A small sample size is a significant limitation of the present study. In this particular problem domain, developing a large labelled dataset for automatic machine learning based analysis is an ongoing issue. Our future work strives to overcome this obstacle by providing additional databases and new methods for data enhancement and augmentation. Despite these limitations, the reported performance values are indeed very promising. Concept-wise and from the soft- and hardware perspective, the proposed framework is ready for clinical use.

CRediT authorship contribution statement

Elli Valla: Software, Formal analysis, Investigation, Visualization, Writing - original draft, Writing - review & editing. **Sven Nõmm:** Methodology, Validation, Writing - review & editing. **Kadri Medijainen:** Data curation. **Pille Taba:** Supervision. **Aaro Toomela:** Conceptualization, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work in the project "ICT programme" was supported by the European Union through European Social Fund. The authors would also like to thank the BDALab team for generously providing the PaHaW dataset to help with our research endeavours.

References

[1] L.M. De Lau, M.M. Breteler, Epidemiology of parkinson's disease, Lancet Neurol. 5 (6) (2006) 525–535.
[2] R.L. Nussbaum, C.E. Ellis, Alzheimer's disease and parkinson's disease, New England J. Med. 348 (14) (2003) 1356–1364.
[3] E. Dorsey, A. Elbaz, E. Nichols, F. Abd-Allah, A. Abdelalim, J. Adsuar, M. Ansha, C. Brayne, J. Choi, D. Collado-Mateo, et al., Gbd 2016 parkinson's disease collaborators. global, regional, and national burden of parkinson's disease, 1990–2016: a systematic analysis for the global burden of disease study 2016, Lancet Neurol. 17 (11) (2018) 939–953.
[4] R. He, X. Yan, J. Guo, Q. Xu, B. Tang, Q. Sun, Recent advances in biomarkers for parkinson's disease, Front. Aging Neurosci. 10 (2018) 305.
[5] M. Politis, K. Wu, S. Molloy, P.G. Bain, K.R. Chaudhuri, P. Piccini, Parkinson's disease symptoms: the patient's perspective, Mov. Disord. 25 (11) (2010) 1646–1651.
[6] A. Gironell, B. Pascual-Sedano, I. Aracil, J. Marín-Lahoz, J. Pagonabarraga, J. Kulisevsky, Tremor types in parkinson disease: a descriptive study using a new classification, Parkinson's Disease 2018 (2018).
[7] A. Barbeau, et al., Parkinson's disease: clinical features and etiopathology, Handbook of clinical neurology 5 (49) (1986) 87–152.
[8] E. Smits, A. Tolonen, L. Cluitmans, M. Gils, B. Conway, R.C Zietsma, K. Leenders, N. Maurits, Standardized Handwriting to Assess Bradykinesia, Micrographia and Tremor in Parkinson's disease, PloS one 9 (05 2014). doi:10.1371/journal.pone.0097614.
[9] C. De Stefano, F. Fontanella, D. Impedovo, G. Pirlo, A.S. di Freca, Handwriting analysis to support neurodegenerative diseases diagnosis: A review, Pattern Recogn. Lett. 121 (2019) 37–45.
[10] D. Impedovo, G. Pirlo, Dynamic handwriting analysis for the assessment of neurodegenerative diseases: a pattern recognition perspective, IEEE Rev. Biomed. Eng. 12 (2018) 209–220.
[11] S. Rosenblum, M. Samuel, S. Zlotnik, I. Erikh, I. Schlesinger, Handwriting as an objective tool for parkinson's disease diagnosis, J. Neurol. 260 (9) (2013) 2357–2361.
[12] G. Vessio, Dynamic handwriting analysis for neurodegenerative disease assessment: A literary review, Appl. Sci. 9 (21) (2019) 4666.
[13] T. Eichhorn, T. Gasser, N. Mai, C. Marquardt, G. Arnold, J. Schwarz, W. Oertel, Computational analysis of open loop handwriting movements in parkinson's disease: a rapid method to detect dopamimetic effects, Movement Disorders 11 (3) (1996) 289–297.
[14] J. Phillips, G.E. Stelmach, N. Teasdale, What can indices of handwriting quality tell us about parkinsonian handwriting? Hum. Mov. Sci. 10 (2–3) (1991) 301–314.
[15] C. Marquardt, N. Mai, A computational procedure for movement analysis in handwriting, J. Neurosci. Methods 52 (1) (1994) 39–45, https://doi.org/10.1016/0165-0270(94)90053-1.
[16] D. Impedovo, Velocity-based signal features for the assessment of parkinsonian handwriting, IEEE Signal Process. Lett. 26 (4) (2019) 632–636.
[17] M.T. Angelillo, D. Impedovo, G. Pirlo, G. Vessio, Performance-driven handwriting task selection for parkinson's disease classification, in: International Conference of the Italian Association for Artificial Intelligence, Springer, 2019, pp. 281–293.
[18] P. Drotar, J. Mekyska, I. Rektorova, L. Masarova, Z. Smékal, M. Faundez-Zanuy, Evaluation of handwriting kinematics and pressure for differential diagnosis of parkinson's disease, Artif. Intell. Med. 67 (2016) 39–46, https://doi.org/10.1016/j.artmed.2016.01.004.
[19] S. Nõmm, K. Bardõš, A. Toomela, K. Medijainen, P. Taba, Detailed analysis of the luria's alternating seriestests for parkinson's disease diagnostics, in: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018, pp. 1347–1352, https://doi.org/10.1109/ICMLA.2018.00219.
[20] T.-L. Yang, C.-H. Lin, W.-L. Chen, H.-Y. Lin, C.-S. Su, C.-K. Liang, Hash transformation and machine learning-based decision-making classifier improved the accuracy rate of automated parkinson's disease screening, IEEE Trans. Neural Syst. Rehabil. Eng. 28 (1) (2019) 72–82.
[21] M. Diaz, M. Moetesum, I. Siddiqi, G. Vessio, Sequence-based dynamic handwriting analysis for parkinson's disease detection with one-dimensional convolutions and bigrus, Expert Syst. Appl. 168 (2021), 114405.
[22] J. Westin, S. Ghiamati, M. Memedi, D. Nyholm, A. Johansson, M. Dougherty, T. Groth, A new computer method for assessing drawing impairment in parkinson's disease, J. Neurosci. Methods 190 (1) (2010) 143–148, https://doi.org/10.1016/j.jneumeth.2010.04.027.
[23] E. Hazan, F. Frankenburg, M. Brenkel, K. Shulman, The test of time: a history of clock drawing, Int. J. Geriatric Psychiatry 33 (1) (2018) e22–e30.
[24] S.L. Pullman, Spiral analysis: a new technique for measuring tremor with a digitizing tablet, Mov. Disord. 13 (S3) (1998) 85–89.
[25] C.R. Pereira, S.A. Weber, C. Hook, G.H. Rosa, J.P. Papa, Deep learning-aided parkinson's disease diagnosis from handwritten dynamics, in: 2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI) Ieee, 2016, pp. 340–346.
[26] I. Kamran, S. Naz, I. Razzak, M. Imran, Handwriting dynamics assessment using deep neural network for early identification of parkinson's disease, Future Gener. Comput. Syst. 117 (2021) 234–244.

[27] S. Zarembo, S. Nõmm, K. Medijainen, P. Taba, A. Toomela, Cnn based analysis of the luria's alternating series test for parkinson's disease diagnostics, in: Asian Conference on Intelligent Information and Database Systems Springer, 2021, pp. 3–13.
[28] S. Nõmm, A. Toomela, J. Kozhenkina, T. Toomsoo, Quantitative analysis in the digital luria's alternating series tests, in: 2016 14th International Conference on Control, Automation, Robotics and Vision (ICARCV), 2016, pp. 1–6, https://doi.org/10.1109/ICARCV.2016.7838746.
[29] S. Nõmm, A. Toomela, An alternative approach to measure quantity and smoothness of the human limb motions, Est. J. Eng. 19 (4) (2013) 298–308.
[30] C.D. Rios-Urrego, J.C. Vásquez-Correa, J.F. Vargas-Bonilla, E. Nöth, F. Lopera, J. R. Orozco-Arroyave, Analysis and evaluation of handwriting in patients with parkinson's disease using kinematic, geometrical, and non-linear features, Comput. Methods Programs Biomed. 173 (2019) 43–52.
[31] T.-L. Yang, P.-J. Kan, C.-H. Lin, H.-Y. Lin, W.-L. Chen, H.-T. Yau, Using polar expression features and nonlinear machine learning classifier for automated parkinson's disease screening, IEEE Sens. J. 20 (1) (2019) 501–514.
[32] N.S. Frolov, E.N. Pitsik, V.A. Maksimenko, V.V. Grubov, A.R. Kiselev, Z. Wang, A. E. Hramov, Age-related slowing down in the motor initiation in elderly adults, Plos one 15 (9) (2020), e0233942.
[33] T. Stöckel, K. Wunsch, C.M. Hughes, Age-related decline in anticipatory motor planning and its relation to cognitive and motor skill proficiency, Front. Aging Neurosci. 9 (2017) 283.
[34] Y.Y. Hoogendam, F. van der Lijn, M.W. Vernooij, A. Hofman, W.J. Niessen, A. van der Lugt, M.A. Ikram, J.N. van der Geest, Older age relates to worsening of fine motor skills: a population-based study of middle-aged and elderly persons, Front. Aging Neurosci. 6 (2014) 259.
[35] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, M. Faundez-Zanuy, Analysis of in-air movement in handwriting: A novel marker for parkinson's disease, Comput. Methods Programs Biomed. 117 (3) (2014) 405–411.
[36] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning. 2nd Edition, Springer Series in Statistics, Springer, 2002.
[37] R.N. Jazar, Advanced Dynamics. Rigid Body, Multibody, and Aerospace Applications, John Wiley & Sons Inc, 2007.
[38] O. Senkiv, S. Nõmm, A. Toomela, Applicability of spiral drawing test for mental fatigue modelling, IFAC-PapersOnLine 51 (34) (2019) 190–195, 2nd IFAC Conference on Cyber-Physical and Human Systems CPHS 2018. doi: 10.1016/j.ifacol.2019.01.064.http://www.sciencedirect.com/science/article/pii/S2405896319300679.
[39] K. Tadist, S. Najah, N.S. Nikolov, F. Mrabti, A. Zahi, Feature selection methods and genomic big data: a systematic review, J. Big Data 6 (1) (2019) 1–24.
[40] C. Aggarwal, Data Mining, Springer, 2015.
[41] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1) (2002) 389–422.
[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[43] L.S. Bernardo, A. Quezada, R. Munoz, F.M. Maia, C.R. Pereira, W. Wu, V.H.C. de Albuquerque, Handwritten pattern recognition for early parkinson's disease diagnosis, Pattern Recogn. Lett. 125 (2019) 78–84.

Appendix 2

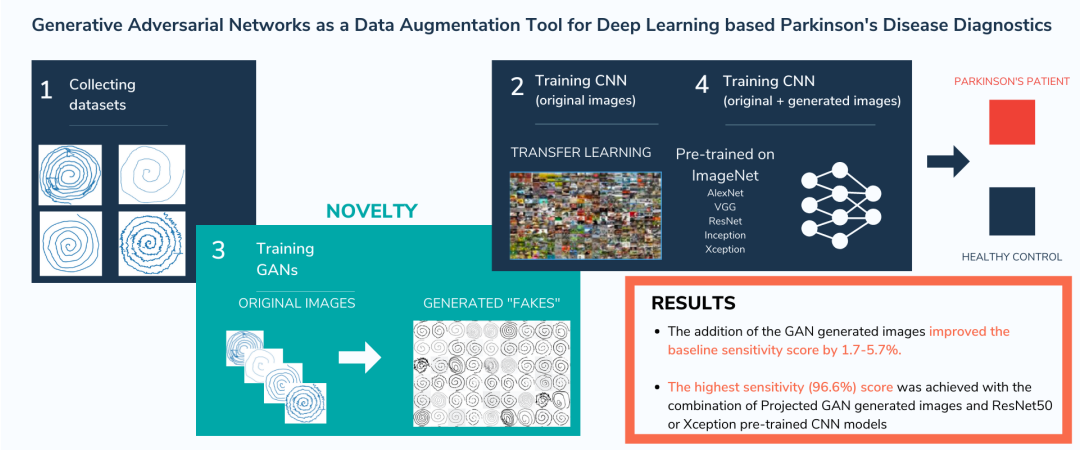
II

Erik Dzotsenidze, Elli Valla, Sven Nõmm, Kadri Medijainen, Pille Taba, and Aaro Toomela. Generative adversarial networks as a data augmentation tool for CNN-based Parkinson's disease diagnostics. volume 55, pages 108–113. Elsevier, 2022

Graphical Abstract

Generative Adversarial Networks as a Data Augmentation Tool for CNN-Based Parkinson's Disease Diagnostics

Erik Dzotsenidze, Elli Valla, Sven Nõmm, Kadri Medijainen, Pille Taba, Aaro Toomela



Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 55-29 (2022) 108–113

IFAC Papers Online CONFERENCE PAPER ARCHIVE

Generative Adversarial Networks as a Data Augmentation Tool for CNN-Based Parkinson's Disease Diagnostics

Erik Dzotsenidze * **Elli Valla** * **Sven Nõmm** *
Kadri Medijainen ** **Pille Taba** ***,**** **Aaro Toomela** †

* *Department of Software Science, School of Information Technology, Tallinn University of Technology (TalTech), Akadeemia tee 15a, 12618, Tallinn, Estonia (e-mail: {erik.dzotsenidze, elli.valla, sven.nomm}@taltech.ee)*
** *Institute of Sport Sciences and Physiotherapy, University of Tartu, Puusepa 8, Tartu 51014, Estonia (e-mail: kadri.medijainen@ut.ee)*
*** *Department of Neurology and Neurosurgery, University of Tartu, Puusepa 8, Tartu 51014, Estonia (e-mail: pille.taba@kliinikum.ee)*
**** *Neurology Clinic, Tartu University Hospital, Puusepa 8, Tartu 51014, Estonia*
† *School of Natural Sciences and Health Tallinn University, Narva mnt. 25, 10120 (e-mail: aaro.toomela@tlu.ee)*

Abstract: Growing research interest has arisen towards automated neurodegenerative disease diagnostics based on the information extracted from the digital drawing tests. Since the performance of modern modelling techniques (machine learning, deep learning) relies heavily on the size of training data available, data scarcity is one of the most significant problems in computer-aided diagnostics. This paper proposes using Generative Adversarial Networks to synthesise digital drawing tests acquired from Parkinson's patients and healthy controls. Four different architectures (StyleGAN2-ADA, StyleGAN2-ADA + LeCam, StyleGAN3 and ProjectedGAN) are evaluated and compared with the traditional data augmentation methods. Convolutional neural networks are utilised for Parkinson's disease diagnostics. Our results indicate that GAN-generated images' addition outperforms the standard augmentation methods in classifying Parkinson's disease in some settings. Therefore, the proposed framework could serve as a potential decision support tool for clinicians in computer-aided fine-motor analysis for neurodegenerative disease diagnostics.

Keywords: Decision making and cognitive processes, assistive technology and rehabilitation engineering, cognitive system engineering, deep learning, generative adversarial networks

1. INTRODUCTION

The Global Burden of Disease study suggests that 6.2 million patients are diagnosed with Parkinson's disease (PD), and this number will double by 2040, surpassing the growth of Alzheimer's disease Dorsey et al. (2018). Finding accurate biomarkers for early diagnosis may significantly improve clinical intervention and treatment and can be utilised to monitor the progress of the disease De Lau and Breteler (2006) He et al. (2018). It has been well documented that the digital signals extracted from the handwriting of PD patients are affected and therefore might serve as a diagnostic marker in a computer-aided analysis Rosenblum et al. (2013), Vessio (2019)Eichhorn et al. (1996), Phillips et al. (1991), Marquardt and Mai (1994), Impedovo (2019), Angelillo et al. (2019), Drotar et al. (2016), Nõmm et al. (2018), Yang et al. (2019), Valla et al. (2022). The most commonly used approaches in the relevant studies can be categorised as follows: numeric methods, where kinematics of the handwriting are analysed Impedovo (2019), Angelillo et al. (2019), Drotar et al. (2016), Nõmm et al. (2018), Valla et al. (2022) and deep learning based approach Diaz et al. (2021), Pereira et al. (2016b), Kamran et al. (2021a) where the digital image is used as an input for a convolutional neural network (CNN) based classification. CNNs have achieved high performance as a decision support tools for Parkinson's disease diagnostics Kamran et al. (2021b), Diaz et al. (2021). However, modern deep learning models rely heavily on the size of training data available. Labelling medical image data is an expensive, time- and resource-consuming task.

Generative neural networks like Generative Adversarial Networks (GAN) Goodfellow et al. (2014) have become popular, primarily because of their ability to generate never before seen images out of training data. In 2014, Goodfellow et al. Goodfellow et al. (2014) designed a new generative neural network framework called Generative Adversarial Network. This new framework used an adversarial process to train the model, which meant

10.1016/j.ifacol.2022.10.240

simultaneously training two models (generator and discriminator), similar to a two-player minimax game. The adversarial training allows GANs to create much better quality outputs than other models, which often produce blurry results. Furthermore, the training process for GANs is unsupervised, which removes the need for data labelling in the training process, making it more straightforward from a data collecting view. However, training two models simultaneously made the training process more unstable. Additionally, the first GAN models needed a lot of real input training data and produced low-resolution images. The work by Karras et al. Karras et al. (2020c) allowed for better control of the image generation process and better quality images. Their GAN model named StyleGAN2 uses a style transfer based approach and a new generative model to improve the state-of-the-art performance and allow for more variety in the generated images. However, the amount of training data needed and the computational resources remained an issue. Karras et al. (2020a) proposed a new method to train GANs called adaptive discriminator augmentation (ADA) to decrease the amount of data needed for training. Their approach applies data augmentation during training to both the generated and real data. This helped stabilise the training process in cases where limited data is available. By using ADA on StyleGAN2, they improved the baseline Fréchet inception distance (FID) and kernel inception distance (KID) scores significantly. Unfortunately, this performance does not linearly scale when using 100% of the data, resulting in only a slight increase in performance. Tseng et al. Tseng et al. (2021a) proposed to add a regularisation scheme to the discriminator under limited data training. They showed that this regulariser limits the LeCam-divergence, which is more prevalent when training with limited data and, as a result, managed to improve the StyleGAN2-ADA baseline FID score by a few points.

With the help of these architectures, it might be possible to generate new Parkinson's handwriting and drawing images and use the synthetic images to train the convolutional neural network (CNN) model. GANs have not been applied to synthesise Parkinson's patients' digital test images for computer-aided disease diagnostics. Therefore, it would be essential to study if those methods could help classify Parkinson's disease. The developed framework could potentially help to reduce the resources and time needed to diagnose a patient and serve as a decision support tool for medical practitioners.

2. PROBLEM STATEMENT

In this work, we evaluate the use of generative adversarial networks (GANs) to increase the robustness and overall classification performance of five pre-trained CNN models.

The problem is divided to the following sub-questions:

- Can generative neural networks be used to generate meaningful Parkinson's patients' handwriting and drawing image training data?
- How does the addition of GAN-generated data affect CNN model classification performance for Parkinson's disease diagnostics?

3. EXPERIMENTAL SETTING

3.1 Data acquisition

Four Parkinson's patients' drawing datasets consisting of digital and hand-drawn Archimedean spiral tests were used in this study.

DraWritePD Our research group performed the data acquisition with an Apple iPad Pro (2016) tablet computer and an Apple Pencil. The tablet has a 26.77cm (10.5 inches) diagonal. The iPad Pro scans the Apple Pencil's signal with a frequency of 240 points per second. From a software perspective — data was collected using a custom iOS application developed by the research team. The dynamic features (time-sequences) captured by the tablet are as follows: x-coordinate (mm); y-coordinate (mm); timestamp (sec); pressure (arbitrary unit of force applied on the surface: [0,..., 6.0]); altitude (rad); azimuth (rad). Total of 24 PD patients (mean age 74.1 $\pm$ 6.7) and 34 age- and gender-matched healthy control subjects (mean age 74.1 $\pm$ 9.1)) participated in the creation of the database. The overall task was to complete a testing battery consisting of 12 different drawing and writing tests. In this paper, we focus only on the Archimedean spiral test. The data acquisition process was conducted with the strict guidance of privacy law. The Research Ethics Committee of the University of Tartu (No. $1275T - 9$) approved the study.

Additionally we used Archimedean spiral drawing tests from *ParkinsonHW* (Sakar et al. (2013); Isenkul et al. (2014)), *HandPD* (Pereira et al. (2016a)), *NewHandPD* (Pereira et al. (2016b)) and Parkinson's drawings from (Zham et al. (2017)).

In total, we collected 930 images, of which 312 were healthy controls (HC), and 618 were Parkinson's patients (PD). As the NewHandPD dataset is an extension of the HandPD dataset, these datasets contain intra-dataset duplicate images. Furthermore, we noted that the NewHandPD dataset contains inter-dataset duplicate images and healthy spiral images with bad image quality (between H16-H37). The duplicates and images with bad quality were removed from our dataset splits. This left us with 702 images (210 Controls and 492 Parkinson's). The images were divided in an 80%/10%/10% train/validation/test split, where the validation and test set were balanced, as seen in Table 2. Every split contained images from each dataset proportionally to the size of each dataset.

For augmenting the images, we used the Albumentations library Buslaev et al. (2020). The augmentation pipeline is described in Table 3.

3.2 Preprocessing

As the images used by the classification and GAN models come from several different datasets, the style and image quality vary significantly between datasets. Before training, each image was passed through preprocessing step, where the background noise and template were removed and the images were turned into grayscale. For HandPD and NewHandPD datasets, we followed the image preprocessing steps from Pereira et al. (2015), which consisted

Table 1. Configurations for GAN training

GAN	Batch size	R1 regularization	Base config	LeCam lambda
StyleGAN2-ADA	64	0.2048	auto	-
StyleGAN2-ADA + LeCam	64	0.2048	auto	$3 * 10^{-7}$
StyleGAN3	64	0.2048	StyleGAN3-T	-
Projected GAN	64	-	FastGAN-lite	-

Table 2. Dataset split. HC - Healthy control, PD - Parkinson's disease

Label	Train	Validation	Test
HC	146	35	35
PD	422	35	35

Table 3. Image augmentation pipeline

Augmentation pipeline
HorizontalFlip()
VerticalFlip()
RandomRotate90()
GridDistortion(border_mode=cv2.BORDER_CONSTANT, value=255)
RandomScale()
RandomBrightnessContrast()

of blurring the image and thresholding. We used a similar preprocessing pipeline with Parkinson's drawing dataset, as these images are grayscale; instead of thresholding based on the difference in each colour channel, we thresholded based on the intensity of the black values.

3.3 GAN

For image generation we used four different GAN architectures: StyleGAN2-ADA (Karras et al. (2020b)), StyleGAN2-ADA + LeCam (Tseng et al. (2021b)), StyleGAN3 (Karras et al. (2021)), Projected GAN (Sauer et al. (2021)). These models were selected because they were specifically created with limited training data in mind and have shown to generate good quality images under these conditions (Karras et al. (2020b); Tseng et al. (2021b); Karras et al. (2021); Sauer et al. (2021)). The overall workflow was implemented using PyTorch framework Paszke et al. (2019).

With each GAN architecture, we trained two unconditional GAN models, one which generates spiral images of healthy controls and the other that generates spiral images of Parkinson's patients.

Each of the GAN models was trained using transfer learning from a model trained on the FFHQ dataset (Karras et al. (2019)). We used only the train split, which was amplified with horizontal, vertical flips and horizontal + vertical flips, during GAN training. The generated images are of size 256x256 pixels. Each model was trained on a single NVIDIA A100 GPU for three days. Configuration for each of the GAN architectures can be seen in Table 1. R1 regularisation value was selected based on the formula found in Section D in Karras et al. (2020b).

For model evaluation, we used Kernel Inception Distance (KID), which measures the dissimilarity between probability distributions and is unbiased when used with small datasets, unlike Fréchet Inception Distance (FID), which is more commonly used as a GAN quality metric (Bińkowski et al. (2018)). We calculated KID every 50 steps and selected the model checkpoint with the lowest score as the best.

3.4 Image classification

For image classification we used PyTorch implementations of AlexNet (Krizhevsky et al. (2012)), ResNet50 (He et al. (2015)), VGG11 (Simonyan and Zisserman (2015)), Inception_v3 (Szegedy et al. (2015)), and Xception (Chollet (2017)). The models were trained using transfer learning with the base models trained on the ImageNet dataset (Deng et al. (2009)). The configuration used for training image classification model can be seen in Table 4.

Table 4. Configuration for classificator training

Option	Value
Epochs	30
Batch size	64
Initial learning rate	0.0001
Learning rate decay	Exponential
Learning rate decay gamma	0.9
Optimizer	Adam
Optimizer weight decay	0.0005
Loss	Cross Entropy

To measure the effectiveness of GAN generated images in the classification task, we created multiple training sets. For each of the GAN architectures, we extended the original training set with images generated from the GAN models. Furthermore, we extended one training set with traditionally augmented images to compare the GAN-augmented training sets. The extended training sets contain nearly the same amount of images of healthy controls and Parkinson's patients. The sizes of the training sets can be seen in Table 5.

Table 5. Training sets

Training set	Size	HC	PD
Original	568	146	422
Augmented	1998	996	1002
GAN-augmented	1998	996	1002

4. MAIN RESULTS

4.1 GAN evaluation

The best KID score for each of the trained GAN architectures can be seen in Table 6. Out of the StyleGAN based architectures, StyleGAN2-ADA gave the best results for both HC and PD. StyleGAN2-ADA + LeCam and StyleGAN3 both gave worse results for both classes. The best KID is achieved by Projected GAN, which has an order of magnitude lower KID. The subset of GAN-generated synthetic images is depicted in Fig 1.

We performed 50 experiments to evaluate the effect the classification model architecture and augmentation type have on the performance of the models. The results are reported as the average of 5 runs using the test set and are seen in Table 7.

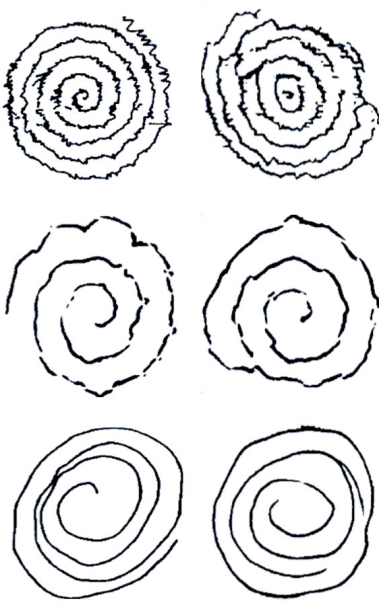


Fig. 1. The comparison of the original (left) and the GAN-generated synthesised (right) digital spirals.

4.2 Baseline evaluation

First, we used the original training set and measured how the CNN architectures perform with no augmented data. We saw that the best results were produced by ResNet, which managed to achieve a sensitivity of 94.3% and specificity of 68.0%. The other architectures produced similar but slightly worse results. AlexNet had the highest specificity with 73.1%.

4.3 Augmentation evaluation

We found that using traditional augmented data achieved better results with every architecture both in terms of specificity with the exception of AlexNet, where specificity fell by 1.1%. The specificity of the other model increased by 6-7%. Sensitivity was more of a mixed bag. Three of the five models saw an increase, with the most significant being AlexNet 3.4% and Xception 2.2%. Inception v3 saw a marginal increase of 0.6%. The sensitivity of VGG decreased the most by 2.9%, and ResNet decreased by 0.6%.

4.4 GAN-augmentation evaluation

Projected GAN augmented data achieved the best sensitivity when compared to the other GAN-augmented training sets. Furthermore, it produced the best sensitivity scores with three of the five CNN models and the highest overall sensitivity of 96.6% with ResNet and Xception. Projected GAN also surpassed the sensitivity of traditional augmentation, with only one exception (AlexNet) and the original training set. StyleGAN based GANs managed to outperform Projected GAN's sensitivity only once.

StyleGAN based augmentation seems to favour specificity more than the Projected GAN. StyleGAN2-ADA matched the highest specificity score of traditional augmentation 76.6% (Inception v3) and got the best specificity out of all the augmentation methods when used with AlexNet. Using StyleGAN2-ADA + LeCam, StyleGAN3, or Projected GAN generated data results in poorer specificity than traditional augmentation with every CNN architecture.

Table 6. Best KID scores of each trained GAN model

GAN	KID (↓)	
	HC	PD
StyleGAN2-ADA	0.01416	0.01054
StyleGAN2-ADA + LeCam	0.01826	0.02517
StyleGAN3	0.02148	0.02113
Projected GAN	0.001264	0.0009285

5. DISCUSSION

The limited amount of labelled data available is a major hurdle for adopting deep learning methods in clinical imaging. To overcome this issue, we evaluated the use of synthetic images derived from the digital handwriting and drawings of Parkinson's patients. Following are the most informative findings:

(1) The addition of the GAN generated images improved the baseline sensitivity score (by 1.7-5.7%) in the case of four CNN models (ResNet50, VGG11, Inception v3, Xception). The shallowest CNN architecture, AlexNet, showed better sensitivity combined with the standard augmentation methods than GAN-based augmentation.
(2) The highest sensitivity (96.6%) score was achieved with the combination of Projected GAN generated images and ResNet50 or Xception pre-trained CNN models. The authors of Projected GAN Sauer et al. (2021) showed its superiority in terms of convergence speed and data efficiency. Their research with Projected-GAN achieved the lowest FID (Fréchet Inception Distance) compared with the state-of-the-art models. Our result, combined with the claims made by Sauer et al. (2021), makes the Projected GAN a suitable choice for this particular problem domain.
(3) The models trained on the original dataset without any augmentation techniques didn't achieve top scores in any experimental settings. This finding concludes the importance of exploring additional augmentations methods to improve the deep learning based diagnostics performance.
(4) It can be seen that the overall specificity scores were lower for the majority of the settings. A highly specific test is good at excluding most people who do not have the condition. However, minimising the probability of false negatives is more important in this case. The more sensitive a test, the less likely an individual with a negative test, will have the disease.

Our results showed that with certain settings, the addition of synthetic GAN-generated images performed better than the standard augmentation method. The only drawback of the proposed methodology is the higher computational cost of training GAN models.

Table 7. Test dataset results. Mean scores over five runs. The values in **bold** indicate the best results for a model. Sn - Sensitivity, Sp - Specificity.

Augmentation method	AlexNet		ResNet50		VGG11		Inception v3		Xception	
	Sn (%)	Sp (%)	Sn (%)	Sp (%)	Sn (%)	Sp (%)	Sn (%)	Sp (%)	Sn (%)	Sp (%)
None	88.0	73.1	94.3	68.0	92.6	66.9	92.0	69.1	90.9	65.7
Traditional	**91.4**	72.0	93.7	**76.0**	89.7	**73.1**	92.6	**76.6**	93.1	**72.6**
StyleGAN2-ADA	85.1	**75.4**	90.9	71.4	94.3	68.0	90.3	**76.6**	93.7	68.0
StyleGAN2-ADA + LeCam	88.6	68.6	95.4	69.1	93.7	66.9	**94.3**	69.1	95.4	63.4
StyleGAN3	88.0	73.1	88.0	73.1	92.0	65.7	93.1	68.0	91.4	65.7
Projected GAN	90.3	68.0	**96.6**	69.7	**95.4**	65.1	92.6	66.3	**96.6**	59.4

6. CONCLUSION

The findings presented in this paper conclude that generative adversarial networks have a strong potential as an augmentation tool for CNN based Parkinson's disease diagnostics. Our future work will extend the experimental settings with more GAN architectures and other neurodegenerative disease oriented handwriting and drawing datasets. In addition to the novel GAN-based Parkinson's digital handwriting and drawing image data augmentation framework, an extensive database of original and synthesised images for Parkinson's disease diagnostics is an outcome of the current research. To minimise the scarcity and high cost of labelled data, all means should be used to make more efficient use of the available data.

ACKNOWLEDGEMENTS

This study was supported by the Grant PRG 957 of the Estonian Research Council. This work in the project "ICT programme" was supported by the European Union through European Social Fund.

REFERENCES

Angelillo, M.T., Impedovo, D., Pirlo, G., and Vessio, G. (2019). Performance-driven handwriting task selection for parkinson's disease classification. In *International Conference of the Italian Association for Artificial Intelligence*, 281–293. Springer.

Bińkowski, M., Sutherland, D.J., Arbel, M., and Gretton, A. (2018). Demystifying mmd gans. doi:10.48550/ARXIV.1801.01401. URL `https://arxiv.org/abs/1801.01401`.

Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., and Kalinin, A.A. (2020). Albumentations: Fast and flexible image augmentations. *Information*, 11(2). doi:10.3390/info11020125. URL `https://www.mdpi.com/2078-2489/11/2/125`.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1800–1807. doi:10.1109/CVPR.2017.195.

De Lau, L.M. and Breteler, M.M. (2006). Epidemiology of parkinson's disease. *The Lancet Neurology*, 5(6), 525–535.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Diaz, M., Moetesum, M., Siddiqi, I., and Vessio, G. (2021). Sequence-based dynamic handwriting analysis for parkinson's disease detection with one-dimensional convolutions and bigrus. *Expert Systems with Applications*, 168, 114405.

Dorsey, E., Elbaz, A., Nichols, E., Abd-Allah, F., Abdelalim, A., Adsuar, J., Ansha, M., Brayne, C., Choi, J., Collado-Mateo, D., et al. (2018). Gbd 2016 parkinson's disease collaborators. global, regional, and national burden of parkinson's disease, 1990-2016: a systematic analysis for the global burden of disease study 2016. *Lancet Neurol*, 17(11), 939–953.

Drotar, P., Mekyska, J., Rektorova, I., Masarova, L., Smékal, Z., and Faundez-Zanuy, M. (2016). Evaluation of handwriting kinematics and pressure for differential diagnosis of parkinson's disease. *Artificial Intelligence in Medicine*, 67, 39 – 46. doi:https://doi.org/10.1016/j.artmed.2016.01.004.

Eichhorn, T., Gasser, T., Mai, N., Marquardt, C., Arnold, G., Schwarz, J., and Oertel, W. (1996). Computational analysis of open loop handwriting movements in parkinson's disease: a rapid method to detect dopamimetic effects. *Movement disorders: official journal of the Movement Disorder Society*, 11(3), 289–297.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

He, R., Yan, X., Guo, J., Xu, Q., Tang, B., and Sun, Q. (2018). Recent advances in biomarkers for parkinson's disease. *Frontiers in aging neuroscience*, 10, 305.

Impedovo, D. (2019). Velocity-based signal features for the assessment of parkinsonian handwriting. *IEEE Signal Processing Letters*, 26(4), 632–636.

Isenkul, M., Sakar, B., and Kursun, O. (2014). Improved spiral test using digitized graphics tablet for monitoring parkinson's disease. In *The 2nd International Conference on E-Health and TeleMedicine*, 171–175. doi:10.13140/RG.2.1.1898.6005.

Kamran, I., Naz, S., Razzak, I., and Imran, M. (2021a). Handwriting dynamics assessment using deep neural network for early identification of parkinson's disease. *Future Generation Computer Systems*, 117, 234–244. doi:https://doi.org/10.1016/j.future.2020.11.020. URL `https://www.sciencedirect.com/science/article/pii/S0167739X20330442`.

Kamran, I., Naz, S., Razzak, I., and Imran, M. (2021b). Handwriting dynamics assessment using deep neural network for early identification of parkinson's disease. *Future Generation Computer Systems*, 117, 234–244.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. (2020a). Training generative adversarial networks with limited data.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. (2020b). Training generative adversarial networks with limited data. In *Proc. NeurIPS*.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., and Aila, T. (2021). Alias-free generative adversarial networks. In *Proc. NeurIPS*.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. doi:10.1109/cvpr.2019.00453. URL `https://doi.org/10.1109/cvpr.2019.00453`.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020c). Analyzing and improving the image quality of stylegan.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc. URL `https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`.

Marquardt, C. and Mai, N. (1994). A computational procedure for movement analysis in handwriting. *Journal of Neuroscience Methods*, 52(1), 39 – 45. doi:http://dx.doi.org/10.1016/0165-0270(94)90053-1.

Nõmm, S., Bardõš, K., Toomela, A., Medijainen, K., and Taba, P. (2018). Detailed analysis of the luria's alternating seriestests for parkinson's disease diagnostics. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1347–1352. doi:10.1109/ICMLA.2018.00219.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E.Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703. URL `http://arxiv.org/abs/1912.01703`.

Pereira, C.R., Pereira, D.R., Silva, F.A., Masieiro, J.P., Weber, S.A.T., Hook, C., and Papa, J.P. (2016a). A new computer vision-based approach to aid the diagnosis of parkinson's disease. *Computer Methods and Programs in Biomedicine*, 136, 79–88.

Pereira, C.R., Pereira, D.R., Da Silva, F.A., Hook, C., Weber, S.A., Pereira, L.A., and Papa, J.P. (2015). A step towards the automated diagnosis of parkinson's disease: Analyzing handwriting movements. In *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, 171–176. doi:10.1109/CBMS.2015.34.

Pereira, C.R., Weber, S.A., Hook, C., Rosa, G.H., and Papa, J.P. (2016b). Deep learning-aided parkinson's disease diagnosis from handwritten dynamics. In *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 340–346. Ieee.

Phillips, J., Stelmach, G.E., and Teasdale, N. (1991). What can indices of handwriting quality tell us about parkinsonian handwriting? *Human Movement Science*, 10(2-3), 301–314.

Rosenblum, S., Samuel, M., Zlotnik, S., Erikh, I., and Schlesinger, I. (2013). Handwriting as an objective tool for parkinson's disease diagnosis. *Journal of neurology*, 260(9), 2357–2361.

Sakar, B., Isenkul, M., Sakar, C.O., Sertbaş, A., Gurgen, F., Delil, S., Apaydin, H., and Kursun, O. (2013). Collection and analysis of a parkinson speech dataset with multiple types of sound recordings. *Biomedical and Health Informatics, IEEE Journal of*, 17, 828–834. doi:10.1109/JBHI.2013.2245674.

Sauer, A., Chitta, K., Müller, J., and Geiger, A. (2021). Projected gans converge faster. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567. URL `http://arxiv.org/abs/1512.00567`.

Tseng, H.Y., Jiang, L., Liu, C., Yang, M.H., and Yang, W. (2021a). Regularizing generative adversarial networks under limited data. doi:10.48550/ARXIV.2104.03310. URL `https://arxiv.org/abs/2104.03310`.

Tseng, H.Y., Jiang, L., Liu, C., Yang, M.H., and Yang, W. (2021b). Regularizing generative adversarial networks under limited data. doi:10.48550/ARXIV.2104.03310. URL `https://arxiv.org/abs/2104.03310`.

Valla, E., Nõmm, S., Medijainen, K., Taba, P., and Toomela, A. (2022). Tremor-related feature engineering for machine learning based parkinson's disease diagnostics. *Biomedical Signal Processing and Control*, 75, 103551. doi:https://doi.org/10.1016/j.bspc.2022.103551.

Vessio, G. (2019). Dynamic handwriting analysis for neurodegenerative disease assessment: A literary review. *Applied Sciences*, 9(21), 4666.

Yang, T.L., Lin, C.H., Chen, W.L., Lin, H.Y., Su, C.S., and Liang, C.K. (2019). Hash transformation and machine learning-based decision-making classifier improved the accuracy rate of automated parkinson's disease screening. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(1), 72–82.

Zham, P., Kumar, D.K., Dabnichki, P., Arjunan, S.P., and Raghav, S. (2017). Distinguishing different stages of parkinson's disease using composite index of speed and pen-pressure of sketching a spiral. *Frontiers in Neurology*, 8.

Appendix 3

III

Vassili Gorbatsov, Elli Valla, Sven Nõmm, Kadri Medijainen, Pille Taba, and Aaro Toomela. Machine learning based analysis of the upper limb freezing during handwriting in Parkinson's disease patients. volume 55, pages 91–95. Elsevier, 2022

Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 55-29 (2022) 91–95

Machine Learning Based Analysis of the Upper Limb Freezing During Handwriting in Parkinson's Disease Patients

Vassili Gorbatsov * **Elli Valla** * **Sven Nõmm** *
Kadri Medijainen ** **Pille Taba** ***,**** **Aaro Toomela** †

* *Department of Software Science, School of Information Technologies, Tallinn University of Technology, Akadeemia tee 15a, 12618, Tallinn, Estonia (e-mail: {sven.nomm, elli.valla, vagorb }@ ttu.ee).*
** *Institute of Sport Sciences and Physiotherapy, University of Tartu, Puusepa 8, Tartu 51014, Estonia (e-mail: kadri.medijainen@ut.ee)*
*** *Department of Neurology and Neurosurgery, University of Tartu, Puusepa 8, Tartu 51014, Estonia (e-mail: pille.taba@kliinikum.ee)*
**** *Neurology Clinic, Tartu University Hospital, Puusepa 8, Tartu 51014, Estonia*
† *School of Natural Sciences and Health Tallinn University, Narva mnt. 25, 10120 (e-mail: aaro.toomela@tlu.ee)*

Abstract: Freezing of the upper limb in Parkinson's disease patients occurring during writing tests constitutes the research subject of the present paper. Digitisation of the writing and drawing tests coupled with artificial intelligence techniques have demonstrated accurate results in supporting the diagnostics of Parkinson's disease. In the digital setting, the analysis of freezing episodes did not get much attention. The main goal of the present paper is to determine if the neighbourhood of the point where freezing occurred possesses sufficient discriminating power to distinguish between the Parkinson's disease patients and healthy control individuals. For each freezing episode, time intervals of one second before and after are considered. These intervals are described by the hand movement's kinematic and pressure parameters. These parameters are used as features for the standard machine learning workflow that applies a nested cross-validation loop. The paper's main findings have demonstrated that analysis of the freezing neighbourhoods allows distinguishing Parkinson's disease patients from age matched healthy controls. The best results were achieved based on the movements occurring one second after the freezing episode. Kinematic and pressure-based features describing these movements have allowed training classifiers whose accuracy, precision, and recall have reached the values of 0.86, 0.86 and 0.93, respectively. Furthermore, the achieved results are comparable to those available in the literature.

Keywords: Computer aided diagnosis, Parkinson's disease, machine learning, drawing test, hand freezing.

1. INTRODUCTION

Parkinson's disease (PD) is a neurodegenerative disorder whose symptoms include unintended motions, rigidity, tremors and freezing Kalia and Lang (2015) which may severely affect the life quality of the patient Louis and Machado (2015). Unfortunately, there is no cure for PD. Nevertheless, the quality of the patient's life may be improved by the proper treatment, which relies on the early diagnosis. The results of Smits et al. (2014) demonstrate that PD frequently affect fine-motor motions. The latest does writing and drawing tests the valuable tools to support PD diagnostics Moustafa et al. (2016), Nackaerts et al. (2017). Digitisation of the writing and drawing tests has started with the seminal paper Marquardt and Mai (1994). Although the digital tables dominate as the primary medium of data acquisition Drotár et al. (2016), some recent results demonstrate the applicability of Tablet PC devices Nõmm et al. (2018), Valla et al. (2022) for the same tasks. In the area of sentence writing, the analysis is either based on the kinematic and pressure description of the entire test or its elements; words and individual letters Netšunajev et al. (2021). The present research turns its attention to the *freezing* phenomenon observed during the writing tests Heremans et al. (2015) which alongside with phenomenon of micrographia Wagle Shukla et al. (2012), Van Gemmert et al. (2003) considered to be important to support diagnostics of PD. The working hypothesis of the present research is that the pressure and kinematic description of the hand movements immediately before and after the freezing possesses sufficient discriminating power to distinguish between PD patients and healthy control (HC) subjects. To confirm or reject this hypothesis machine-learning (ML) approach is applied. The workflow of the present research consists of the following steps.

10.1016/j.ifacol.2022.10.237

Recognition of freezing episodes from the test recordings. Feature extraction and selection, application of the ML classifiers, and comparison with the results available from the literature. The paper is organised as follows. Research hypothesis and problem statement constitute the section 2. Software and hardware setting alongside the data acquisition procedure presented in the section 3. Section 4 presents employed techniques. Achieved results are resented in section 5. Discussion of the achieved results and comparison to those available from the literature constitutes the section 6. Conclusions are drawn in the last section

2. PROBLEM STATEMENT

The working hypothesis of the present research assumes that the kinematic and pressure parameters of the hand movements observed in certain time-neighbourhood around the freezing episode allow us to distinguish if the PD patient or HC subject performs the test. Supporting this hypothesis requires the handwriting dataset, a technique to recognise freezing, a method to extract kinematic and pressure-related features, and machine learning classifiers' application.

3. EXPERIMENTAL SETTING

The current research belongs to the large project aiming to support diagnostics of the PD using the analysis of the gross- and fine- motions. Experimental setting for the present research has been explicitly described in Nõmm et al. (2018), Netšunajev et al. (2021) and Valla et al. (2022) to make this paper self-sufficient main facts describing hardware and software settings alongside of testing process are listed below.

3.1 Hardware and software

Apple iPad pro (2016) with a 9.7-inch screen and Apple pen was used as the medium to perform writing tests. The team members developed special software to provide an interface and acquire motions of the tip of the Apple Pen (stylus). The coordinates of the stylus tip and its pressure applied to the screen are collected to the matrix. Rows of the matrix correspond to the observation points acquired up to two hundred times per second, and columns contain information about x and y coordinates, pressure applied to the screen, the orientation of the stylus and time stamp. In addition, the latest allows computing numerous kinematic parameters describing the motion. Upon the completion of each test, this information is saved for future processing in JSON files.

3.2 Handwriting data acquisition

The group of volunteers consisting of 24 PD patients, with confirmed diagnosis, (mean age 74.1 ± 6.7) and 30 gender-matched HC subjects (mean age 74.1 ± 9.1)) took part in the testing process. Participants were asked to complete the batter consisting of 12 tests. In the present paper, only a sentence writing test is considered. The subjects whose native language is *Estonian* were asked to handwrite the sentence *Kui Arno isaga kooli-majja jõudsid, olid tunnid juba alanud*, which means *When Arno with his father arrived at the school lessons had already started*. This sentence is taken from the book learned in all the schools in Estonian by the school children of age between 7 and 9. Figure 1 depicts the sentence written on the screen of a tablet. The research was performed in a strict accordance to the data protection law and was approved by the Research Ethics Committee of the University of Tartu (No. $1275T - 9$).

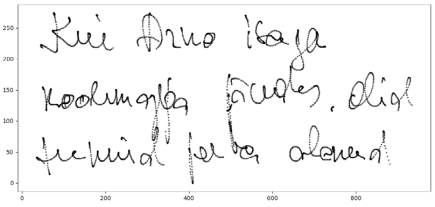


Fig. 1. Sentence written on the screen of a tablet.

4. METHODS AND RESULTS

The data processing was performed offline, employing two standalone applications. The first one was used to recognise freezing and extract the neighbourhoods corresponding to the 0.5 second time intervals around it. Then the second application was used to perform feature extraction and nested cross-validation with respect to the six most commonly used machine learning classifiers Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), AdaBoost (AB). Nested cross-validation allows to avoid bias which may lead to overfitting and, in turn to the over-optimistic model goodness Hastie et al. (2002), Angelillo et al. (2019).

4.1 Feature extraction

The feature extraction step follows the idea proposed by Nõmm et al. (2016) and later extended by Valla et al. (2022). The first part is similar to many other results from Marquardt and Mai (1994) to Drotár et al. (2016) and requires one to compute velocity, acceleration and higher derivatives; jerk - J_N, yank Y_N, tug - T_N, snatch - Sn_N and shake - Sh_N, describing the kinematic parameters of motion Valla et al. (2022). Initial set of parameters is complemented by the parameters describing angle between the directional vectors in two consequent observation points. Alongside with measures of descriptive statistics (maximum, mean, variance etc.) initial feature set is complemented by the integral-like parameters referred as *motion mass* Nõmm et al. (2016). Let us illustrate the way these parameters are computed on the example of velocity and its corresponding mass parameter. Adopt the following notations; N the number of observation points of interest, v_i the velocity (computed on the basis of two neighbouring points and along the directional vector of the stylus tip movement) i where $i \in \{1, \dots, N\}$ then *velocity mass* is defined by equation 4.1

$$V_N = \sum_{i=1}^{N} |v_i| \quad (1)$$

The same logic applies to computing mass values of the other kinematic and pressure parameters.

4.2 Freezing episode

The freezing of the hand during the writing is referred as *freezing episode*. According to Perez-Lloret et al. (2014) the freezing episode is defined as *a sudden, variable, and often unpredictable transient break in movement*. This definition needs to be formalised to allow automatic detection. The authors of Heremans et al. (2015) have adapted the definition for the case of hand movement during writing in the following way: *handwriting freezing was defined as an involuntary stop or clear absence of effective writing movements during at least* 1 *second*. The latest allows being implemented in the form of programming code. Of course, one has to consider that while the hand may freeze, small jigging in the coordinates may occur. The proper setting of threshold values may easily solve such problems. Once freezing episodes are detected, timestamps of the points where they begin and end provide the information about the ending and beginning points of the intervals to extract. In this research, the length of such interval was found experimental to be 1 of a second. Figure 2 depicts freezing episodes, numbered and marked by yellow lines with green arrows.

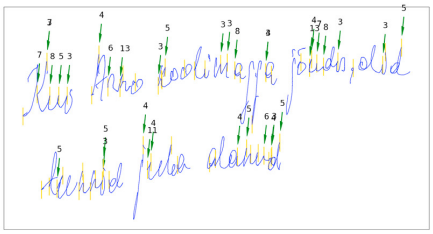


Fig. 2. Freezing episodes.

It is important to note that the number of freezing episodes may vary between the PD patients and HC subjects. To avoid problems caused by unbalanced data sets, proper sampling was applied to guarantee an equal number of freezing episodes and proper proportions of the episodes from the different parts of the sentence. Then feature extraction procedure described in the previous subsection was applied to these intervals. After each test, all the freezing episodes were described by the tuple of kinematic, pressure and motion mass parameters. Each tuple inherited the label of the test it had been computed from, consequently forming the dataset to be used for ML analysis.

4.3 Feature selection and classification

According to Aggarwal (2015) there are four primary feature selection techniques; filter models (sometimes referred to as filter methods), wrapper methods, embedded techniques and probabilistic techniques. In this research, filter models and wrapper techniques are considered. The two remaining techniques require larger data sets and are therefore left for future studies. Kinematic, pressure and motion mass parameters describing the movements of the stylus tip are numeric, which makes Fisher's score 2 the natural choice for feature selection. Aggarwal (2015) Fisher's score belongs to the filter-model feature selection techniques and assigns to each feature a numeric value, allowing to order the features with respect to their discriminating power.

$$F = \frac{\sum_{i=1}^{N} p_j(\mu - \mu_i)^2}{\sum_{i=1}^{N} p_i \sigma_i^2} \quad (2)$$

In (2) N is the number of classes, μ and μ_i are the mean value for the entire set along the given feature and mean value of the class i, respectively; p_i proportion of the class i and σ_i is the standard deviation of the class i. Larger values of the Fisher's score indicate higher discriminating power. In parallel the feature selection was performed using a wrapper method. The evaluation is done by training and testing a specific classification model that estimates the relevance of a given subset of features. Although filter methods are fast and scalable, they have the disadvantage of ignoring the interaction with the classifier. Wrapper methods are proven to be more efficient but also more computationally expensive Guyon et al. (2002). In Valla et al. (2022) authors found that the features selected with a wrapper method possessed greater discriminating power compared to the filter method based selection. In this paper, we consider the SVM recursive feature elimination (SVM-RFE) wrapper method proposed by Guyon et al. (2002). Unlike the non-nested feature selection, when features are selected based on an entire training set, and then cross-validation procedure is applied, nested cross-validation requires one to perform feature selection for each fold Hastie et al. (2002). The difference is that in the first case model sees the testing set implicitly, which may lead to overoptimistic results, whereas the second case resembles real-life scenarios more precisely Angelillo et al. (2019).

5. MAIN RESULTS

The workflow described in the previous section has been applied to the three following cases:

(1) Analysis of the movements immediately before the freezing episodes.
(2) Analysis of the movements immediately after the freezing episodes.
(3) Analysis of the movements of entire sentences.

The last case was performed to ease the comparison with the other results in the area. For each classifier, evaluation was performed with 2, 3, 4 and 5 features. Experimenting with a larger number of features did not make sense, considering the size of the data set available for the research.

5.1 Fisher's score based feature selection

In the case of filter model (Fisher's score) feature selection, the analysis of the movements occurring before the freezing

episodes has led to the following results. The mean of the velocity, its vertical projection and standard deviation alongside the mean and standard deviation of the acceleration appeared to possess the highest discriminating power. The number of features did not affect the metrics describing the goodness of the models. The accuracy have varied between 0.73 and 0.78, precision 0.76 and 0.78, recall 0.81 and 0.93. While goodness metrics of different classifiers did not vary much, SVM has demonstrated the highest performance, closely followed by logistic regression. For the movements occurring after the freezing episodes, acceleration mean and standard deviation topped the list of most discriminating features, followed by their projections on the vertical axis and velocity mean. At the same time, their Fisher's score values were, on average, ten per cent higher compared to those computed based on movements before the freezing episodes. This difference is reflected by the goodness of the classifiers trained on this data. The accuracy have varied between 0.78 and 0.71, precision 0.82 and 0.85, recall 0.85 and 0.91. In this case, logistic regression has demonstrated the strongest performance, closely followed by the SVM classifier. For the analysis of the entire sentence, Fisher's score has indicated that acceleration and jerk based features possess the highest discriminating power. Model goodness metrics have varied as follows: the accuracy has varied between 0.74 and 0.8, precision 0.79 and 0.81, recall 0.86 and 0.91. The SVM demonstrated the best performance.

5.2 Wrapper method based feature selection

Application of the wrapper method has allowed to reach higher goodness for some classifiers. Also it has demonstrated high variations in performance metrics. Analysis of the movements occurring before the freezing events has led to the following results. Although irrespective of the number of features, the accuracy has varied between 0.77 and 0.82, precision 0.79 and 0.85, recall 0.82 and 0.93. These results are comparable with those reported by Drotár et al. (2016) and slightly below model goodness achieved by Valla et al. (2022). Specificity appears to be the only goodness criteria lacking behind the competition ranging between 0.57 and 0.72. SVM classifier has demonstrated the most substantial performance for any number of variables. The mean of the velocity values has been presented in all the feature sets, followed by the angular velocity mass and a maximum altitude of the stylus; the both were presented in three feature sets. Analysis of the movements occurring after the freezing events has led to higher accuracy (0.8 - 0.86), precision (0.84 - 0.88) and specificity (0.68 - 0.79), whereas recall (sensitivity) did not change significantly. For this case SVM was the strongest among all the tested classifiers. However, the feature sets have changed dramatically. The standard deviation of the velocity along the horizontal axis was presented in all the feature sets alongside the maximum value of pen altitude. This pair was followed by the mean value of the angle describing the change of the directional vector. Overall, one can expect that movements after the freezing are better suited than those before the freezing event. To illustrate the results one can plot the graph of the velocity in one second neighbourhood around the freezing episode. In Figure 3 blue lines represent the velocities observed around freezing episodes in the motions of HC subjects,

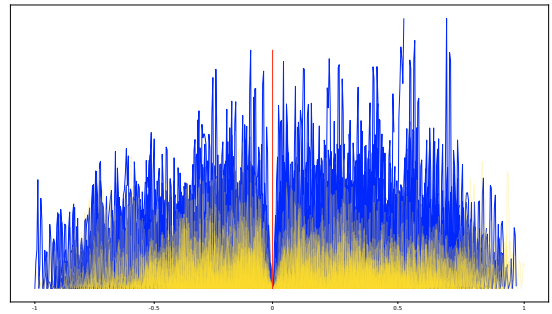


Fig. 3. Freezing episodes.

and yellow lines represent the velocities observed around freezing episodes in the motions of PD patients. Red line represents the freezing point. One may see that the mean values and standard deviations would be distinguishable between those groups. Finally, when the same procedure was applied to the entire sentence variance of the metrics describing model goodness increased. The accuracy has varied between 0.74 and 0.85, precision 0.70 and 0.95, recall 0.76 and 0.93. At the same time, for some models of four and five features, specificity has risen to 0.93. While SVM has remained the top performer for the models with four features, LogReg has demonstrated the same performance with four variables with better quality for the case of five features. The feature sets for this case more closely resemble the case of the movements after the freezing event. The only significant change is the appearance of the features based on acceleration.

6. COMPARISON AND DISCUSSION

The main problem in comparison to the results of sentence writing tests is that they perform in different languages and frequently have different lengths. One of the most cited works in this area whose techniques are similar to those used in the present research is Drotár et al. (2016), where the tested subjects were asked to write the sentence Chezch *Tramvaj dnes už nepojede* which means *The tram will not go today.* Unlike the present research, the writing was combined with the other tasks of the testing battery. Nevertheless, comparing achieved model goodness, one may conclude that the analysis of movements observed before the freezing led to the similar goodness as reported by Drotár et al. (2016). The Movements observed after the freezing episodes lead to models of higher goodness. The work of Netšunajev et al. (2021) also analyses sentence writing tests. Unlike the present research, the research is based on finding and analysing individual letters. Nevertheless, Netšunajev et al. (2021) was achieved on the same dataset as the present research, allowing a more detailed comparison. Metrics of model goodness in Netšunajev et al. (2021) vary in the same range as the current research's; the accuracy has varied between 0.73 and 0.82, precision 0.71 and 0.91, recall 0.61 and 0.93 (excluding SVM and KNN which have demonstrated extremely poor performance). Neither of the models can be declared the ultimate winner. Finally, in Netšunajev et al. (2021) most frequently used features are the mean of the velocity and the mean of acceleration mass of the angular change. While

achieved results are not enough to claim that the analysis of freezing episodes is more informative than micrographia or other tests, it provides the results of comparable goodness and therefore is a valuable addition to computer-aided diagnostics support. Another essential point to note is that analysis of the elements of the test is no less informative in comparison to the analysis of the entire test. The latest, in some sense, coincides with the results of Nomm et al. (2019), which demonstrates that in drawing tests, part of the test may be the same informative as an entire test in diagnostics of PD.

7. CONCLUSIONS

The attention of the present research has been focused on the stylus tip movements observed one second before and one second after the hand freezing episodes. It was demonstrated that movement occurring after the freezing episode might be as informative as the entire test. Also, goodness metrics of the trained ML classifiers are comparable and sometimes even better than those reported in the literature.

ACKNOWLEDGEMENTS

This study was supported by the Grant PRG957 of the Estonian Research Council. This work in the project "ICT programme" was supported by the European Union through European Social Fund.

REFERENCES

Aggarwal, C. (2015). *Data Mining*. Springer.

Angelillo, M.T., Impedovo, D., Pirlo, G., and Vessio, G. (2019). Performance-driven handwriting task selection for parkinson's disease classification. In *International Conference of the Italian Association for Artificial Intelligence*, 281–293. Springer.

Drotár, P., Mekyska, J., Rektorová, I., Masarová, L., Smékal, Z., and Faundez-Zanuy, M. (2016). Evaluation of handwriting kinematics and pressure for differential diagnosis of parkinson's disease. *Artificial Intelligence in Medicine*, 67, 39 – 46. doi: https://doi.org/10.1016/j.artmed.2016.01.004.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1), 389–422.

Hastie, T., Tibshirani, R., and Friedman, J. (2002). *The Elements of Statistical Learning. 2nd Edition*. Springer Series in Statistics. Springer.

Heremans, E., Nackaerts, E., Vervoort, G., Vercruysse, S., Broeder, S., Strouwen, C., Swinnen, S.P., and Nieuwboer, A. (2015). Amplitude manipulation evokes upper limb freezing during handwriting in patients with parkinson's disease with freezing of gait. *PLOS ONE*, 10(11), 1–13. doi:10.1371/journal.pone.0142874. URL `https://doi.org/10.1371/journal.pone.0142874`.

Kalia, L.V. and Lang, A.E. (2015). Parkinson's disease. *The Lancet*, 386(9996), 896 – 912.

Louis, E.D. and Machado, D.G. (2015). Tremor-related quality of life: a comparison of essential tremor vs. parkinson's disease patients. *Parkinsonism & related disorders*, 21(7), 729–735.

Marquardt, C. and Mai, N. (1994). A computational procedure for movement analysis in handwriting. *Journal of Neuroscience Methods*, 52(1), 39 – 45. doi: http://dx.doi.org/10.1016/0165-0270(94)90053-1.

Moustafa, A.A., Chakravarthy, S., Phillips, J.R., Gupta, A., Keri, S., Polner, B., Frank, M.J., and Jahanshahi, M. (2016). Motor symptoms in parkinson's disease: A unified framework. *Neuroscience and Biobehavioral Reviews*, 68, 727 – 740. doi: https://doi.org/10.1016/j.neubiorev.2016.07.010.

Nackaerts, E., Heremans, E., Smits-Engelsman, B.C., Broeder, S., Vandenberghe, W., Bergmans, B., and Nieuwboer, A. (2017). Validity and reliability of a new tool to evaluate handwriting difficulties in parkinson's disease. *PloS one*, 12(3), e0173157.

Netšunajev, A., Nõmm, S., Toomela, A., Medijainen, K., and Taba, P. (2021). Parkinson's disease diagnostics based on the analysis of digital sentence writing test. *Vietnam Journal of Computer Science*, 8(04), 493–512.

Nõmm, S., Bardõš, K., Mašarov, I., Kozhenkina, J., Toomela, A., and Toomsoo, T. (2016). Recognition and analysis of the contours drawn during the poppelreuter's test. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 170–175. doi:10.1109/ICMLA.2016.0036.

Nõmm, S., Bardõš, K., Toomela, A., Medijainen, K., and Taba, P. (2018). Detailed analysis of the luria's alternating seriestests for parkinson's disease diagnostics. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1347–1352. doi: 10.1109/ICMLA.2018.00219.

Nomm, S., Kossas, T., Toomela, A., Medijainen, K., and Taba, P. (2019). Determining necessary length of the alternating series test for parkinson's disease modelling. In *2019 International Conference on Cyberworlds (CW)*, 261–266. IEEE.

Perez-Lloret, S., Negre-Pages, L., Damier, P., Delval, A., Derkinderen, P., Destée, A., Meissner, W.G., Schelosky, L., Tison, F., and Rascol, O. (2014). Prevalence, determinants, and effect on quality of life of freezing of gait in parkinson disease. *JAMA neurology*, 71(7), 884–890.

Smits, E., Tolonen, A., Cluitmans, L., Gils, M., Conway, B., C Zietsma, R., Leenders, K., and Maurits, N. (2014). Standardized Handwriting to Assess Bradykinesia, Micrographia and Tremor in Parkinson's disease. *PloS one*, 9. doi:10.1371/journal.pone.0097614.

Valla, E., Nõmm, S., Medijainen, K., Taba, P., and Toomela, A. (2022). Tremor-related feature engineering for machine learning based parkinson's disease diagnostics. *Biomedical Signal Processing and Control*, 75, 103551. doi: https://doi.org/10.1016/j.bspc.2022.103551.

Van Gemmert, A., Adler, C.H., and Stelmach, G. (2003). Parkinson's disease patients undershoot target size in handwriting and similar tasks. *Journal of Neurology, Neurosurgery & Psychiatry*, 74(11), 1502–1508.

Wagle Shukla, A., Ounpraseuth, S., Okun, M., Gray, V., and Schwankhaus, J. (2012). Micrographia and related deficits in parkinson's disease: a cross-sectional study. *BMJ open*, 2(3). doi:https://doi.org/10.1136/bmjopen-2011-000628.

Appendix 4

IV
Elli Valla, Henry Laur, Sven Nõmm, Kadri Medijainen, Pille Taba, and Aaro Toomela. Deep learning based segmentation of Luria's alternating series test to support diagnostics of Parkinson's disease. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1066–1071. IEEE, 2023

Deep Learning Based Segmentation of Luria's Alternating Series Test to Support Diagnostics of Parkinson's Disease

E. Valla, H. Laur, S. Nõmm, K. Medijainen, P. Taba, A. Toomela
Tallinn University of Technology, Tartu University, Tallinn University

ABSTRACT

Focus: Analysis of segments in Luria's alternating series test for Parkinson's disease diagnostics using digitization and machine learning.
Gap: Disparity between human-assisted assessments and machine learning algorithms.
Approach: Using deep learning for object detection and classical machine learning for kinematic feature analysis.
Contributions: Evaluating informativeness of test elements and establishing an automatic segmentation framework.

INTRODUCTION

Overview: The study aims to analyze the digital Luria's alternating series tests (Figure 1) for computer-aided diagnostics of Parkinson's Disease (PD).
Relevance: Digitization reveals characteristics not visible to the naked eye, aiding in accurate diagnosis.
Challenge: Bridging the gap between human practitioners' assessments and machine learning methods.

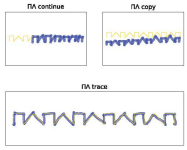


Figure 1. Luria's alternating series patterns. Example of digitized patient's drawings.

METHODS

Data Acquisition: Utilized a custom iOS application, involving 24 Parkinson's disease patients and 34 healthy control subjects.
Segmentation Algorithm: Employed the YOLO algorithm for corner detection in ПА tests, followed by segmentation and feature extraction. Segmented parts of the test were then used in machine learning-based classification (Figure 2).
Machine Learning Classifiers: Six machine learning classifiers were used for the analysis: Logistic Regression, RandomForest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Adaboost.

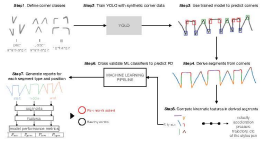


Figure 2. General methodology

Feature Analysis: Focused on kinematic features extracted from the segmented parts, assessing their relevance in distinguishing between Parkinson's disease patients and healthy controls (Figure 3).

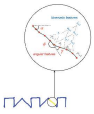


Figure 3. Angular and kinematic features extracted from the digital signal

RESULTS

Best-Performing Segments: The **acute angle** at the beginning of the ПА-copy task showed remarkable performance with an accuracy of **93.8%**, precision of 100.0%, sensitivity of 88.3%, and specificity of 100.0%. In the ПА-continue task, **vertical lines** maintained stable performance with an accuracy of **96.7%**, precision of 95.0%, sensitivity of 100.0%, and specificity of 93.3% (Figure 4).
Segment Distribution in ПА-Copy Task: All segment types in the ПА-copy task were informative and well-distributed, with this task having the largest count of informative segments (Table 1).
Lower Performance of Transition Segments: Transition segments exhibited lower performance metrics, especially in sensitivity, making them less suitable for accurate diagnosis. The start position of these segments was particularly weak in terms of predictive power (Table 2).

Table 1.

Position	Type	F_{acc}	F_{prec}	F_{sens}	F_{spec}
Start	Vertical	0.875	0.875	0.875	0.875
	Diagonal	0.810	0.900	0.781	0.821
	Acute angle	0.938	1.000	0.883	1.000
	Right angle	0.875	0.875	0.875	0.875
	Diagonal	0.876	0.918	0.900	0.850
Middle	Vertical	0.75	0.833	0.625	0.875
	Vertical	0.875	0.875	0.875	0.875
	Acute angle	0.938	0.938	0.949	0.933
	Right angle	0.875	0.875	0.875	0.875
End	Vertical	0.941	0.889	0.980	0.900
	Diagonal	0.875	0.875	0.875	0.875
	Acute angle	0.828	0.828	0.855	0.800
	Acute angle	0.688	0.733	0.733	0.667
	Right angle	0.875	0.875	0.875	0.875
	Diagonal	0.967	0.966	0.951	0.967

Table 2.

Position	Type	F_{acc}	F_{prec}	F_{sens}	F_{spec}
Start	Tr to L	0.776	0.885	0.698	0.810
	L to H	0.729	0.690	0.500	0.793
Middle	H to L	0.765	0.773	0.630	0.843
	L to H	0.748	0.733	0.550	0.865
End	H to L	0.765	0.750	0.630	0.833
	L to H	0.811	0.863	0.617	0.927

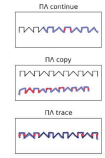


Fig 3. The primary findings highlight the most informative (i.e. error-prone) segments (in red) in Luria's alternating series test for PD.

CONCLUSION

Implications: The study's method can significantly aid clinicians in diagnosing neurodegenerative diseases, reducing public health system burden.
Future Directions: Expanding the methodology to other handwriting tests and exploring deep learning-based analysis.

Deep Learning Based Segmentation of Luria's Alternating Series Test to Support Diagnostics of Parkinson's Disease

1st Elli Valla
Department of Software Science
Tallinn University of Technology
Tallinn, Estonia
elli.valla@taltech.ee

2nd Henry Laur
Department of Software Science
Tallinn University of Technology
Tallinn, Estonia
helaur@taltech.ee

3rd Sven Nõmm
Department of Software Science
Tallinn University of Technology
Tallinn, Estonia
sven.nomm@taltech.ee

4th Kadri Medijainen
Institute of Sport Sciences and Physiotherapy
University of Tartu
Tartu, Estonia
kadri.medijainen@ut.ee

5th Pille Taba
Department of Neurology
University of Tartu
Tartu, Estonia
pille.taba@kliinikum.ee

6th Aaro Toomela
School of Natural Sciences and Health
Tallinn University
Tallinn, Estonia
aaro.toomela@tlu.ee

2023 International Conference on Machine Learning and Applications (ICMLA) | 979-8-3503-4534-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICMLA58977.2023.00158

***Abstract*—This research paper focuses on the analysis of various segments of Luria's alternating series drawing test as a diagnostic support for Parkinson's disease. Digitization of drawing tests has allowed capturing pen movement parameters imperceptible to the naked eye, enabling precise neurological disorder diagnosis. However, this analysis of parameters presents a disparity between the pre-digital era's human-assisted assessment and the machine learning algorithm employed today. While human practitioners primarily emphasized overall performance and subject errors, the machine learning method relies on kinematic and pressure features to characterize pen tip movements. The paper aims to bridge this gap by utilizing the deep learning object detection algorithm to identify test elements and classical machine learning techniques to analyze kinematic and pressure parameters associated with drawing these elements. The main research contribution centers around two key aspects: 1) evaluating the individual informativeness of test elements at different stages, i.e., beginning, middle, and final parts of the test, and 2) establishing an efficient automatic segmentation framework aimed at enhancing decision support systems in the context of Parkinson's disease diagnosis.**

***Index Terms*—deep learning, yolo, Luria's alternating series, Parkinson's disease, handwriting dataset**

I. Introduction

Through the digitization of drawing tests, it has become feasible to describe characteristics that were previously invisible to the naked eye. By combining these descriptions with machine learning techniques, highly accurate models have been developed to support the diagnosis of neurological disorders, including Parkinson's disease (PD). The latest is a neurodegenerative disorder [1] that can severely affect motor functions in a patient. PD symptoms encompass non-purposeful motions, tremor [2], and freezing [3], adversely affecting the patient's everyday life quality. Despite the absence of a cure for PD, early diagnosis and appropriate treatment can alleviate symptoms and help maintain a fulfilling daily life. The evaluation procedure for symptoms in Parkinson's disease, as suggested by [4], involved drawing and writing tests. These tests were originally conducted using paper and pencil, relying on visual assessment by the practitioner. This approach inherently carried subjectivity in the assessment process. However, the introduction of digital tablets and tablet PCs has spurred research into digitizing these drawing and writing tests. From the pioneering findings of [5] until recently, the primary approach involved constructing features based on the kinematic and pressure parameters of pen tip movements during the test. Machine learning classifiers [6] were then applied to these features for analysis. The most commonly used approaches in the relevant studies can be categorised as follows: numeric methods, where kinematics of the handwriting are analysed [7] [8] [9] [10] [11] [12] [13] and deep learning based approach [14] [15] [16] [17] where the digital image is used as an input for a convolutional neural network (CNN) based classification. CNN's have achieved high performance as a decision support tool for Parkinson's disease diagnostics [14] [18]. However, modern deep learning models rely heavily on the size of training data available. Labelling medical image data is an expensive, time- and resource-consuming task. Observing the model performance achieved by [7] and [11], alongside other relevant studies, prompts the question of whether more research should be directed towards studying new tests. In this study, Luria's alternating series test was utilized (see Fig. 1) [19]. Several important arguments warrant consideration. Firstly, Luria's alternating series tests enable the determination of whether the disease has primarily affected the motion planning function or the motion execution function. Secondly, these tests facilitate the identification of errors caused by the persevering phenomenon, which involves the inability to

1946-0759/23/$31.00 ©2023 IEEE
DOI 10.1109/ICMLA58977.2023.00158

switch between consequent motor actions. Lastly, the question of which elements of the test hold greater informativeness to support PD diagnostics has received insufficient attention. This last question has value from both academic and practical perspectives. The practical perspective is concerned with the ability to perform tests on devices with a small screen. To address these questions, deep learning object detection models are employed to recognize individual elements of the alternating series tests. Subsequently, a machine learning-based analysis is conducted on the various types of test elements. The structure of this paper is as follows: Section II outlines the research hypotheses. Section III provides a comprehensive discussion of the research methods, experimental settings, and the hardware and software utilized. The main research findings are presented in Section IV, followed by an in-depth discussion of the attained results in Section V. Finally, the paper concludes in the last section.

II. Research Hypothesis

There are two main hypotheses that span current research. The first hypothesis suggests that different parts of the test (repeating patterns observed in the beginning, middle, and end) possess varying discriminating power. This hypothesis draws inspiration from the results of [20]. The second hypothesis proposes that the kinematic and pressure parameters describing the movements of the stylus tip in perseveration-representing parts of the test are more informative than other segments in distinguishing between PD patients and healthy control (HC) subjects.

The following subproblems must be solved to support or reject the stated hypotheses.

1) Train corner detection model for automatic segmentation.
2) Segment the drawing tests into previously defined parts.
3) Calculate kinematic- and pressure-related features as described in [11].
4) Analyse discriminating power of the different segments:
   a) the beginning, middle, and end part of the test;
   b) the perseveration describing switching points.

III. Research Methods and Workflow

This chapter presents all the methods utilized to extract the various segments from the ΠΛ-tests (see Fig. 1).

A. Data acquisition

The research team utilized a custom iOS application to gather the data. The tablet recorded the following dynamic features in the form of time-sequences: x-coordinate (mm); y-coordinate (mm); timestamp (sec); pressure (arbitrary unit of force applied on the surface: [0,..., 6.0]); altitude (rad); azimuth (rad). Total of 24 PD patients (mean age 74.1 ± 6.7) and 34 age- and gender-matched healthy control subjects (mean age 74.1 ± 9.1)) participated in the creation of the database. The overall task was to complete a testing battery consisting of 12 different drawing and writing tests. In this paper, we focus only on the Luria's alternating series test. The data acquisition process was conducted with the strict guidance of privacy law. The Research Ethics Committee of the University of Tartu (No. $1275T - 9$) approved the study.

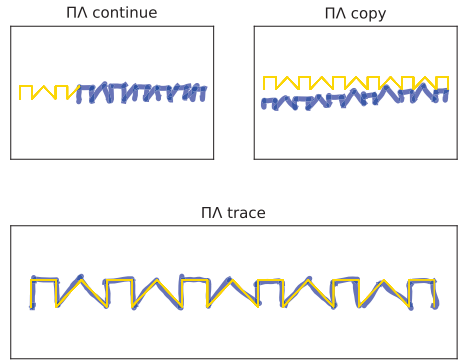


Fig. 1. Luria alternating series patterns. The subject was assigned three distinct tasks. The initial pattern is depicted in yellow, while the blue line represents the trajectory of the subject's drawings.

B. Corner detection

The "You Only Look Once" algorithm, or YOLO, is used to detect corners in the ΠΛ test. It is a series of end-to-end deep learning models designed for fast object detection, developed by Joseph Redmon *et al.* and first described in [21].

The corners are divided into the following categories:

- Right angle corners
- Upper acute angle corners
- Lower acute angle corners

When labeling the corners, an area around the corner is selected to get sufficient data points and analyze the corners separately from the straight and diagonal lines. It is not just the center point of the corner; it encompasses an area around it. Fig. 2 represents the overall segmentation workflow. Firstly, YOLO is trained on synthetic corners and then used to predict corners on the real patients' data. Step 2 of Fig. 2 is described in the subsequent section.

1) YOLO training process: Three augmentation methods, namely horizontal flip, rotation, and shear, were applied to enhance the training dataset, resulting in a total of 288 images for training. The initial dataset consisted of 90 images, with an 80/20 split between training and validation, yielding 18 validation images. YOLO was trained for 300 epochs, and the best epoch (237) achieved the following metrics: mean average precision (mAP) 0.5:0.95 of 0.295, mAP 0.5 of 0.833, precision of 0.84, and recall of 0.82. The mAP scores

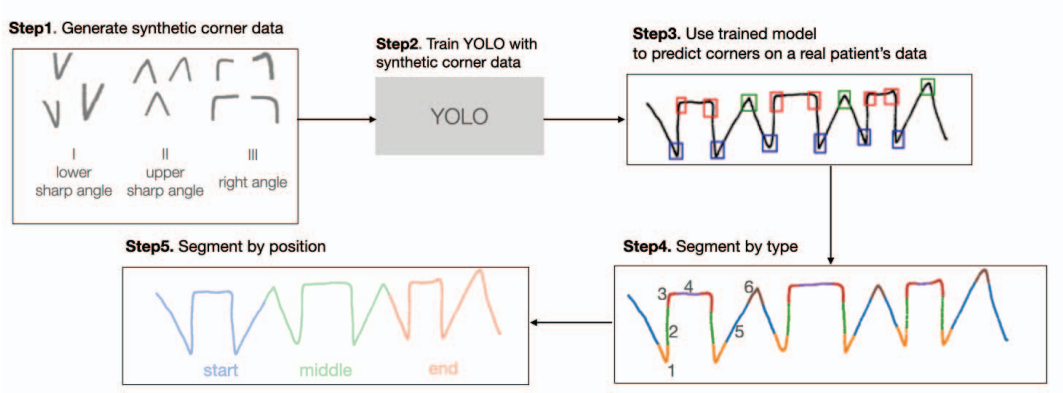


Fig. 2. The role of YOLO in the overall workflow. The Step 4 image visualizes the different segment types derived after corner detection. 1: lower acute angle corners (orange), 2: vertical lines (green), 3: right angle corners (red), 4: horizontal lines (purple), 5: diagonal lines (blue), 6: upper acute angle corners (brown).

improved initially for 50 epochs, plateauing after 150 epochs. Precision and recall showed similar patterns, plateauing after about 100 epochs. The box loss decreased gradually during training without overfitting. Class loss reached a plateau for both training and validation, with slight improvements in training loss. Object loss decreased as expected for training but showed significant overfitting in the validation after approximately 150 epochs. Real patient data validation highlighted the importance of an appropriate confidence threshold to control false positives and overlapping bounding boxes. A confidence threshold of 0.5 offered the best balance of precision and recall for future tests on real patient data.

C. Test segmentation process

The subsequent phase in segmenting the ΠΛ-tests involves the extraction of distinct segments. While YOLO provides solely the corner points of the ΠΛ-test, our objective is to partition the entire dataset, presenting the challenge of discerning the initiation and termination points of individual segments. Employing the timestamps associated with the data points alone does not offer a viable solution, as it is plausible that the patient may have attempted to rectify an earlier mistake made during the drawing procedure. The methodology employs a clustering technique wherein all data points falling within a specified distance are grouped together as a singular segment. Initially, a data point is chosen and assigned as the seed of a new segment. Subsequently, the nearest neighboring point is examined, and if its distance from the segment seed is within the maximum permissible distance threshold, it is incorporated into the segment. This process is iterated until no points satisfying the maximum distance criterion remain. Following this, a new data point is selected as a fresh seed, establishing a new segment, and the procedure is repeated. The rationale behind this approach is to simulate the sequential movement of the patient during the drawing process, thereby delineating segments that correspond to distinct drawing strokes. The clustering technique employed effectively tackles the primary challenge of segmentation by exclusively utilizing the x- and y-coordinates of the data points to demarcate segments. An advantageous aspect of this method is its independence from a predefined number of clusters, rendering it more resilient to irregularities or anomalous data points.

1) Segmentation by type: We have established six distinct segment types based on the outcomes of the YOLO algorithm and additional segmentation analysis. Please refer to Fig. 2 Step 4 for a more detailed explanation.

Based on the YOLO results, we can obtain the corner coordinates and their respective classes. However, it is essential to distinguish and classify different line types separately. The classification of lines is based on the following criteria:

- A line connecting two corners at right angles is classified as a horizontal line.
- A line connecting two corners at acute angles is classified as a diagonal line.
- A line connecting one acute angle corner and one right angle corner is classified as a vertical line.

To identify adjacent corners within a segment, we need to find the data points next to each corner. As the remaining segment types consist of lines, we consider the tips of the lines, representing the two points with the maximum distance between them, as the closest points to the corners. To achieve this, we utilize the convex hull of the data points and derive the maximum distances between them.

The next step involves determining the segment types for transitions from Π to Λ and from Λ to Π. This is accomplished by categorizing each lower acute corner as either: a) Π to Λ, b) Λ to Π. To make this determination, we analyze the preceding segments to the lower acute corner. If the previous segment is a diagonal straight line or an upper acute angle corner, the type is classified as Λ to Π. Conversely, if the previous segment

is a horizontal line, a vertical line, or a right angle corner, the type is classified as Π to Λ. Once the segment types for the lower acute corners are determined, all other segment data points are discarded and excluded from further analysis.

2) Segmentation by position: The subsequent phase involves determining the position of each segment within the test, specifically identifying the start, middle, and end parts. To accomplish this, we locate the minimum and maximum x-values from the dataset, assigning each data point its respective position. The test is then divided into three equal-length parts based on the x-axis. If the mean x-value of a segment falls within the first part, it is classified as a start segment. If it falls within the second part, it is categorized as a middle segment. Conversely, if the mean x-value is situated within the third part, it is designated as an end segment. See Fig. 2 Step 5 for visual explanation. The start segments are denoted by the color blue, the middle segments by green, and the end segments by red.

In subsequent machine learning-based analyses, we will consider both the segment types and positions as relevant factors.

D. Feature engineering and classification

Raw time-series data, including pen position (x- and y-coordinates), timestamp, pen pressure, pen inclination (altitude), and pen orientation (azimuth), can be utilized to compute an infinite number of features. In this study, we derived kinematic features (displacement, velocity, acceleration, etc.), spatial-temporal features (duration, distance), geometric features (altitude, azimuth, yaw, etc.), and pressure features.

Specifically, velocity represents the rate at which the displacement of the position vector changes with respect to time. Similarly, we computed acceleration as the rate of change in velocity and jerk as the rate of change in acceleration with respect to time. To comprehensively analyze the data, we considered up to the sixth time derivative of the position vector.

By examining the slope of the position vector, we were able to extract angle α, and yaw (γ) was defined as the change in the direction in which the point vector is pointing. Additionally, the ϕ angle was derived from the aforementioned angles. Enriching the angular feature set, we included up to a third of their respective time derivatives. Fig. 3 visually illustrates the calculations of the features mentioned above. Feature extraction resulted in either a single-valued feature or a vector feature. For all resulting vector features, the following statistical measures were calculated: mean, median, standard deviation, maximum and minimum value. In addition, horizontal and vertical components of the kinematic features were computed. A total of 202 features were extracted from the raw signals of the fully segmented ΠΛ tests. An SVM recursive feature elimination wrapper method was used to decrease the dimensionality of the data. Six machine learning classifiers are trained, Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Random Forest, and AdaBoost. Each test is divided by position and segment type.

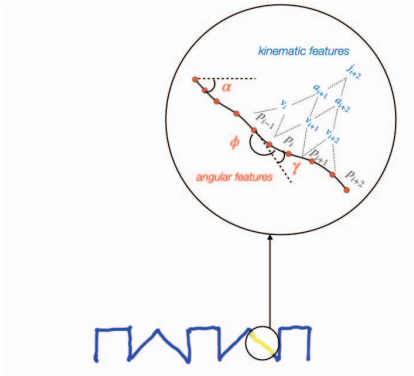


Fig. 3. Angular and kinematic features extracted from the digital signals.

TABLE I
ΠΛ-TRACE CLASSIFICATION BY SEGMENT TYPE

Position	Type	P_{acc}	P_{prec}	P_{sen}	P_{spec}
Start	Vertical	0.905	0.860	1.000	0.800
	Horizontal	0.776	0.850	0.717	0.867
	Acute angle	0.619	0.653	0.783	0.467
	Right angle	0.776	0.900	0.667	0.933
	Diagonal	0.871	0.910	0.867	0.867
Middle	Vertical	0.819	0.867	0.783	0.867
	Horizontal	0.743	0.883	0.683	0.800
	Acute angle	0.714	0.713	0.833	0.567
	Right angle	0.848	0.820	0.950	0.733
	Diagonal	0.724	0.780	0.717	0.733
End	Vertical	0.938	0.960	0.950	0.900
	Horizontal	0.657	0.763	0.650	0.633
	Acute angle	0.748	0.740	0.900	0.567
	Right angle	0.843	0.933	0.783	0.933
	Diagonal	0.776	0.753	0.867	0.667

The best results averaged over a 5-fold nested [22] cross-validated loop are presented in Section IV.

IV. Results

The following section describes highest scores averaged over 5-fold nested cross-validated runs of the segmented ΠΛ tests. These scores are visualized in Tables I, II, III. The best-performing segment with an accuracy of 93.8%, precision of 100.0%, sensitivity of 88.3%, and specificity of 100.0% is the acute angle at the beginning of the ΠΛ-copy task. The vertical lines showed a stable performance across all three ΠΛ tests, achieving high accuracy of 96.7%, precision 95.0%, sensitivity 100.0%, and specificity 93.3% in the middle of the ΠΛ-continue task. The vertical lines provided the most considerable predictive power of all the lines. In the ΠΛ-copy task, all the segment types were informative and distributed across the task. The ΠΛ-copy task also had the largest count of informative segments.

The transition segments exhibited lower performance compared to other segments (see Tables IV, V, VI), particularly showing poor sensitivity, which makes them less suitable for accurately diagnosing patients. Among the transition segments,

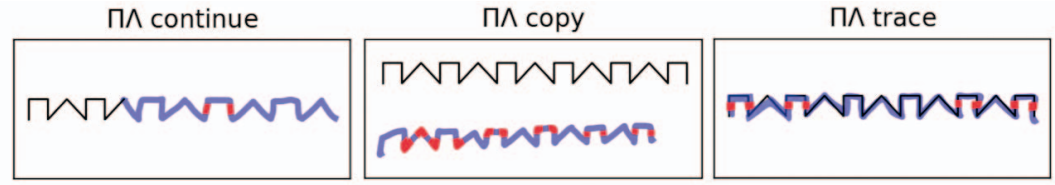


Fig. 4. The primary findings highlight the most informative (i.e. error-prone) segments (in red) in Luria's alternating series test for PD.

TABLE II
ΠΛ-COPY CLASSIFICATION BY SEGMENT TYPE

Position	Type	P_{acc}	P_{prec}	P_{sen}	P_{spec}
Start	Vertical	0.800	0.790	0.950	0.667
	Horizontal	0.810	0.900	0.783	0.833
	Acute angle	0.938	1.000	0.883	1.000
	Right angle	0.838	0.933	0.817	0.800
	Diagonal	0.876	0.910	0.900	0.833
Middle	Vertical	0.771	0.813	0.817	0.700
	Horizontal	0.805	0.883	0.783	0.867
	Acute angle	0.910	0.910	0.950	0.833
	Right angle	0.848	0.950	0.783	0.933
	Diagonal	0.843	0.850	0.900	0.767
End	Vertical	0.938	0.950	0.950	0.933
	Horizontal	0.820	0.920	0.833	0.800
	Acute angle	0.686	0.753	0.733	0.667
	Right angle	0.781	0.883	0.717	0.867
	Diagonal	0.867	0.860	0.933	0.767

TABLE III
ΠΛ-CONTINUE CLASSIFICATION BY SEGMENT TYPE

Position	Type	P_{acc}	P_{prec}	P_{sen}	P_{spec}
Start	Vertical	0.629	0.670	0.783	0.467
	Horizontal	0.867	0.900	0.883	0.867
	Acute angle	0.840	0.900	0.867	0.800
	Right angle	0.700	0.740	0.833	0.600
	Diagonal	0.820	0.850	0.883	0.767
Middle	Vertical	0.967	0.950	1.000	0.933
	Horizontal	0.780	0.767	0.867	0.700
	Acute angle	0.780	0.817	0.800	0.767
	Right angle	0.560	0.550	0.733	0.367
	Diagonal	0.827	0.900	0.867	0.800
End	Vertical	0.805	0.763	1.000	0.533
	Horizontal	0.787	0.920	0.767	0.867
	Acute angle	0.667	0.700	0.733	0.567
	Right angle	0.867	0.950	0.817	0.933
	Diagonal	0.781	0.933	0.683	0.933

TABLE IV
ΠΛ-TRACE CLASSIFICATION (ONLY TRANSITION CORNERS)

Position	Type	P_{acc}	P_{prec}	P_{sen}	P_{spec}
Start	Π to Λ	0.756	0.680	0.690	0.810
	Λ to Π	0.729	0.650	0.700	0.753
Middle	Π to Λ	0.765	0.773	0.650	0.843
	Λ to Π	0.740	0.733	0.550	0.867
End	Π to Λ	0.740	0.798	0.630	0.833
	Λ to Π	0.811	0.883	0.617	0.927

TABLE V
ΠΛ-COPY CLASSIFICATION (ONLY TRANSITION CORNERS)

Position	Type	P_{acc}	P_{prec}	P_{sen}	P_{spec}
Start	Π to Λ	0.680	0.656	0.800	0.600
	Λ to Π	0.671	0.633	0.450	0.814
Middle	Π to Λ	0.860	0.920	0.750	0.933
	Λ to Π	0.727	0.633	0.650	0.786
End	Π to Λ	0.780	0.840	0.650	0.867
	Λ to Π	0.747	0.780	0.600	0.843

TABLE VI
ΠΛ-CONTINUE CLASSIFICATION (ONLY TRANSITION CORNERS)

Position	Type	P_{acc}	P_{prec}	P_{sen}	P_{spec}
Start	Π to Λ	0.698	0.667	0.417	0.867
	Λ to Π	0.718	0.684	0.667	0.753
Middle	Π to Λ	0.744	0.763	0.633	0.813
	Λ to Π	0.706	0.533	0.367	0.887
End	Π to Λ	0.847	0.933	0.667	0.960
	Λ to Π	0.711	0.727	0.600	0.760

the start position was found to be particularly weak in terms of predictive power. As a result, using these segments alone may not provide reliable and accurate diagnostic support.

V. Discussion

Overall, the segments with the best predictive power were the vertical lines and acute angles. Fig. IV depicts the most informative segments for each test. The better performance of the acute angles compared to the right angles might be attributed to the occurrence of more complex movements at the acute angles. Although their metrics were not low, the horizontal lines did not perform as well as the other two line types. This result might be because horizontal lines are usually shorter than the rest, meaning the number of data points is more diminutive. The results also align with the results from [10], which achieved an accuracy of 91.0%. Each test had the best segment from different positions. Interestingly, the segment's position did not seem to affect the results. The large count of informative segments in the ΠΛ-copy task suggests that the presence of the template may simplify the task for Parkinson's patients and thus make detection more complex. A similar finding was presented in [23], page 105, where the performance of gross motor skills was affected by adding the more structured template.

VI. Conclusions

Our successful implementation of handwriting- and drawing-based computer-aided analysis holds promise as a decision support tool for clinicians in neurodegenerative disease diagnostics, potentially alleviating the burden on the public health system. The proposed method automatically segments Luria's alternating series test and identifies the most informative parts for Parkinson's disease diagnostics. Results demonstrate varying discriminative power among different segments, The acute angle at the beginning of the ΠΛ-copy task performing the best (93.8% accuracy, 100.0% precision, 88.3% sensitivity, and 100.0% specificity). The vertical lines showed a stable performance across all three ΠΛ-tests, achieving high accuracies (up to 96.7%) in the middle and end positions. This suggests the significance of specific handwritten segments in Parkinson's disease detection.

In the future, we plan to expand our analysis to include other handwriting tests, such as Archimedean spiral and sentence writing. Additionally, we aim to explore deep learning-based analysis of the segments. Overall, this automatic test segmentation combined with machine learning-based decision support software could find practical application in clinical settings.

References

[1] L. V. Kalia and A. E. Lang, "Parkinson's disease," *The Lancet*, vol. 386, no. 9996, pp. 896 – 912, 2015.

[2] E. D. Louis and D. G. Machado, "Tremor-related quality of life: a comparison of essential tremor vs. parkinson's disease patients," *Parkinsonism & related disorders*, vol. 21, no. 7, pp. 729–735, 2015.

[3] E. Heremans, E. Nackaerts, G. Vervoort, S. Vercruysse, S. Broeder, C. Strouwen, S. P. Swinnen, and A. Nieuwboer, "Amplitude manipulation evokes upper limb freezing during handwriting in patients with parkinson's disease with freezing of gait," *PLOS ONE*, vol. 10, no. 11, pp. 1–13, 11 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0142874

[4] E. Smits, A. Tolonen, L. Cluitmans, M. Gils, B. Conway, R. C Zietsma, K. Leenders, and N. Maurits, "Standardized Handwriting to Assess Bradykinesia, Micrographia and Tremor in Parkinson's disease," *PloS one*, vol. 9, 05 2014.

[5] C. Marquardt and N. Mai, "A computational procedure for movement analysis in handwriting," *Journal of Neuroscience Methods*, vol. 52, no. 1, pp. 39 – 45, 1994.

[6] 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.

[7] D. Impedovo, "Velocity-based signal features for the assessment of parkinsonian handwriting," *IEEE Signal Processing Letters*, vol. 26, no. 4, pp. 632–636, 2019.

[8] M. T. Angelillo, D. Impedovo, G. Pirlo, and G. Vessio, "Performance-driven handwriting task selection for parkinson's disease classification," in *International Conference of the Italian Association for Artificial Intelligence*. Springer, 2019, pp. 281–293.

[9] P. Drotar, J. Mekyska, I. Rektorova, L. Masarova, 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.

[10] S. Nõmm, K. Bardõš, A. Toomela, K. Medijainen, and P. Taba, "Detailed analysis of the luria's alternating seriestests for parkinson's disease diagnostics," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec 2018, pp. 1347–1352.

[11] E. Valla, S. Nõmm, K. Medijainen, P. Taba, and A. Toomela, "Tremor-related feature engineering for machine learning based parkinson's disease diagnostics," *Biomedical Signal Processing and Control*, vol. 75, p. 103551, 2022.

[12] T.-L. Yang, C.-H. Lin, W.-L. Chen, H.-Y. Lin, C.-S. Su, and C.-K. Liang, "Hash transformation and machine learning-based decision-making classifier improved the accuracy rate of automated parkinson's disease screening," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 1, pp. 72–82, 2019.

[13] M. Isenkul, B. Sakar, and O. Kursun, "Improved spiral test using digitized graphics tablet for monitoring parkinson's disease," 05 2014.

[14] M. Diaz, M. Moetesum, I. Siddiqi, and G. Vessio, "Sequence-based dynamic handwriting analysis for parkinson's disease detection with one-dimensional convolutions and bigrus," *Expert Systems with Applications*, vol. 168, p. 114405, 2021.

[15] C. R. Pereira, S. A. Weber, C. Hook, G. H. Rosa, and J. P. Papa, "Deep learning-aided parkinson's disease diagnosis from handwritten dynamics," in *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. Ieee, 2016, pp. 340–346.

[16] I. Kamran, S. Naz, I. Razzak, and M. Imran, "Handwriting dynamics assessment using deep neural network for early identification of parkinson's disease," *Future Generation Computer Systems*, vol. 117, pp. 234–244, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X20330442

[17] S. Zarembo, S. Nõmm, K. Medijainen, P. Taba, and A. Toomela, "Cnn based analysis of the luria's alternating series test for parkinson's disease diagnostics," in *Asian Conference on Intelligent Information and Database Systems*. Springer, 2021, pp. 3–13.

[18] I. Kamran, S. Naz, I. Razzak, and M. Imran, "Handwriting dynamics assessment using deep neural network for early identification of parkinson's disease," *Future Generation Computer Systems*, vol. 117, pp. 234–244, 2021.

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

[20] S. Nomm, T. Kossas, A. Toomela, K. Medijainen, and P. Taba, "Determining necessary length of the alternating series test for parkinson's disease modelling," in *2019 International Conference on Cyberworlds (CW)*. IEEE, 2019, pp. 261–266.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection." [Online]. Available: https://arxiv.org/abs/1506.02640

[22] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning. 2nd Edition*, ser. Springer Series in Statistics. Springer, 2002.

[23] L. S. Vygotsky, *The collected works of LS Vygotsky: Problems of the theory and history of psychology*. Springer Science & Business Media, 1987, vol. 3.

Appendix 5

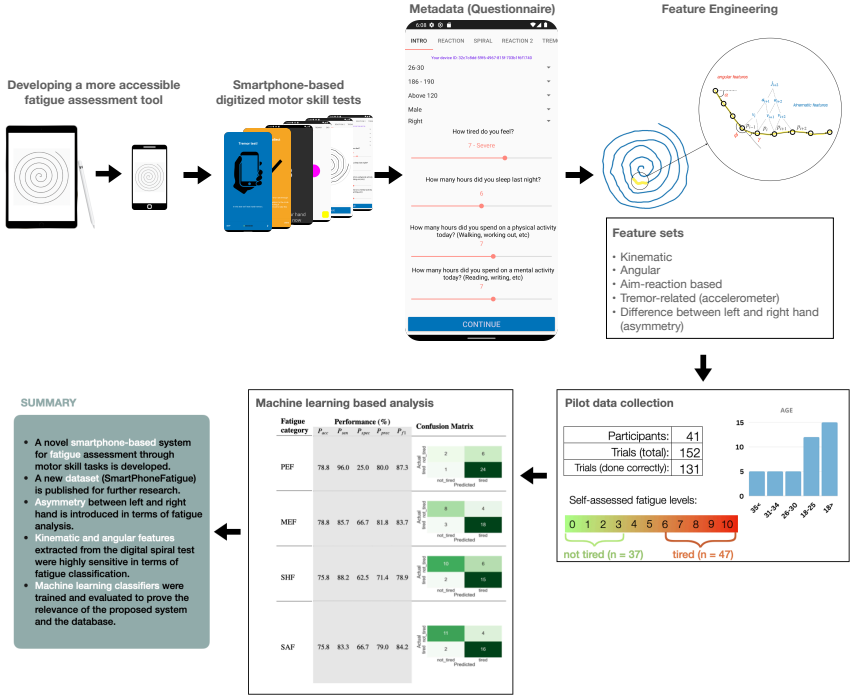
V

Elli Valla, Ain-Joonas Toose, Sven Nõmm, and Aaro Toomela. Transforming fatigue assessment: Smartphone-based system with digitized motor skill tests. *International journal of medical informatics*, 177:105152, 2023

Graphical Abstract

Transforming Fatigue Assessment: Smartphone-Based System with Digitized Motor Skill Tests

Elli Valla, Ain-Joonas Toose, Sven Nõmm, Aaro Toomela



International Journal of Medical Informatics 177 (2023) 105152

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Transforming fatigue assessment: Smartphone-based system with digitized motor skill tests

Elli Valla [a,*], Ain-Joonas Toose [a], Sven Nõmm [a], Aaro Toomela [b]

[a] Department of Software Science, School of Information Technology, Tallinn University of Technology (TalTech), Akadeemia tee 15a, 12618, Tallinn, Estonia
[b] School of Natural Sciences and Health, Tallinn University, Narva mnt. 25, 10120, Tallinn, Estonia

ARTICLE INFO

Keywords:
Fatigue
Motor skills
mHealth
Dataset
Machine learning

ABSTRACT

Background: The condition of fatigue is a complex and multifaceted disorder that encompasses physical, mental, and psychological dimensions, all of which contribute to a decreased quality of life. Smartphone-based systems are gaining significant research interest due to their potential to provide noninvasive monitoring and diagnosis of diseases.
Objective: This paper studies the feasibility of using smartphones to collect motor skill related data for machine learning based fatigue detection. The authors' main goal is to provide valuable insights into the nature of fatigue and support the development of more effective interventions to manage it.
Methods: An application for smartphones running on Android OS is developed. Two aim-based reaction tests, an Archimedean spiral test, and a tremor test, were assembled. 41 subjects participated in the study. The resulting dataset consists of 131 trials of fatigue assessment alongside digital signals extracted from the motor skill tests. Six machine learning classifiers were trained on computed features extracted from the collected digital signals.
Results: The collected dataset SmartPhoneFatigue is presented for further research. The real-world utility of this database was shown by creating a methodology to construct a fatigue predictive model. Our approach incorporated 60 distinct features, such as kinematic, angular, aim-based, and tremor-related measures. The machine learning models exhibited a high degree of prediction rate for fatigue state, with an accuracy exceeding 70%, sensitivity surpassing 90%, and an f1-score greater than 80%.
Conclusion: The results demonstrate that the proposed smartphone-based system is suitable for motion data acquisition in non-controlled environments and shows promise as a more objective and convenient method for measuring fatigue.

1. Introduction

In numerous industries such as transportation, healthcare, and manufacturing, fatigue presents a significant challenge that can result in accidents, mistakes, and decreased productivity. The cited studies [1] have revealed a concerning statistic: 3.6% of fatal road accidents are caused by physical exhaustion or drowsiness, highlighting the critical importance of recognizing early signs of fatigue. Fatigue can be described as a change in the psychobiological state caused by prolonged periods of demanding cognitive and physical activity, sleep deprivation, and other factors [2][3][4][5][6]. This state is associated with stress, aging, depression, illness, neurological disorder such as multiple sclerosis, Parkinson's disease, post-stroke, and Alzheimer's disease [7]. The prevalence and impact of fatigue have also been documented in cancer patients [8][9]. There are two main types of fatigue: peripheral (physical) and central (mental) fatigue [10]. Physical fatigue is commonly characterized as decreased ability to engage in physical tasks following prior physical exertion. Mental fatigue is identified through a decline in performance on tasks that demand alertness, as well as the retrieval and manipulation of information from memory [11]. Concepts like tiredness [12][13], or loss of focus [14] come up when talked about mental fatigue. Despite fatigue being a well-known phenomenon, its underlying mechanisms and effects are still not fully understood, necessitating further investigation [15]. By analyzing fatigue through the lens of fine motor ability, this paper seeks to deepen our understanding of this complex condition. The findings of this study hold potential for the development of improved hypotheses and more effective interventions for fatigue management.

* Corresponding author.
E-mail addresses: elli.valla@taltech.ee (E. Valla), ain-joonas.toose@taltech.ee (A.-J. Toose), sven.nomm@ttu.ee (S. Nõmm), aaro.toomela@tlu.ee (A. Toomela).

https://doi.org/10.1016/j.ijmedinf.2023.105152
Received 27 April 2023; Received in revised form 6 July 2023; Accepted 11 July 2023
Available online 20 July 2023

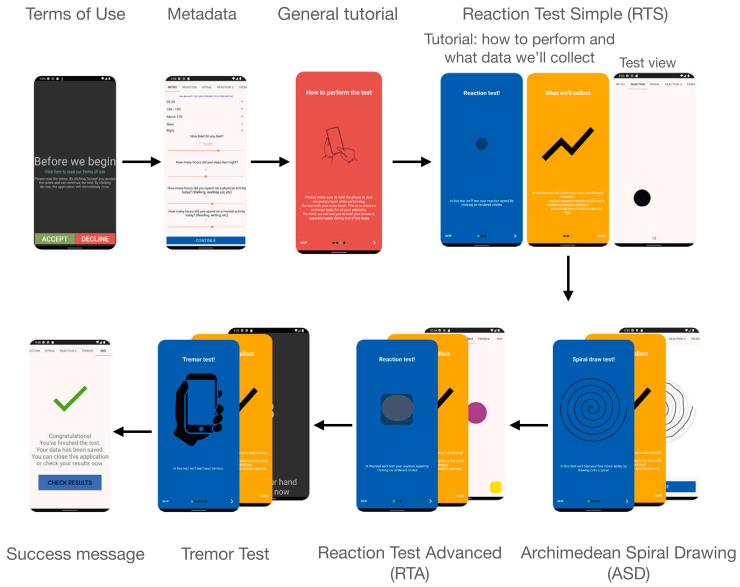


Fig. 1. The sequence of the Fatigue Application.

Technology advancements have led to the development of smartphone-based systems that use sensors and the capabilities of smartphones to capture data on an individual's physical and cognitive abilities. Several studies have introduced smartphone-based frameworks (m-Health) to monitor health and well-being in free-living environments [16][17][18][19][20][21].

1.1. Related work

Traditional approaches to fatigue assessment have limitations regarding accuracy and/or practicality. Authors of [22] have successfully tracked eye movement to assess mental fatigue through an extensive camera system. The blink duration, the mean velocity of the saccade, and the saccade duration were assessed. However, their tests were not focused on motor skills, nor could they provide test scenarios accessible to most people. In another study [23], the author used a digital version of the Stroop test [24] and heart rate monitors to measure athletes' mental fatigue before and after various physical tests. The study of [25] used electroencephalogram (EEG) signals to monitor the mental fatigue of construction workers. The proposed framework aligned cognitive fatigue state classifications with self-reported fatigue states, achieving an accuracy of 88.85%. Previous studies have analyzed how a state of fatigue impairs motor skills [26][27]. The results suggested that movement was slowed in the presence of mental fatigue. The authors observed a ~10% increase in actual movement duration, indicating an impairment of motor skills following a lengthy cognitively demanding task. Prior research has connected mental fatigue to spiral drawing tests. In [28], a tablet-based spiral drawing test demonstrated reproducible patterns for various levels of mental fatigue. Another study [29] described connections between muscular fatigue to hand tremors in a closed position test. However, the study's results by [30] showed no difference in hand tremors for mental fatigue.

This research uniquely combines motor skill abilities and fatigue analysis using a smartphone application offering a more convenient method for measuring fatigue. The hypothesis is that a feature set derived from smartphone data can effectively differentiate between three fatigue states: self-assessed tiredness, physical exertion, and mental exertion.

The novel contributions of this work are as follows:

1. We developed a smartphone system to collect fatigue-related data.
2. We collected data to evaluate the effectiveness of detecting fatigue from digitized fine motor skill tests.
3. We present the dataset describing numeric features acquired with the proposed framework.
4. We employ machine learning to analyze and classify fatigued state.

The rest of this paper is organized as follows. Sections 2 and 3 describe the materials and methods used to develop the proposed framework. Section 3.2 and Appendix A introduce the characteristics of the published dataset. Sections 4 and 5 report and discuss the experimental results to highlight the effectiveness of the proposed method. The paper is summarized in Section 6.

2. Smartphone application

To encourage repeated usage, the mobile application was designed with accessibility as a key requirement, as depicted in Fig. 1, illustrating its overall structure and workflow.

The participant's journey begins with accepting the Terms of Service agreement, which outlines the research topic, data collection goals, data characteristics, and withdrawal information. Following this, a general tutorial familiarizes participants with test-taking procedures and phone positioning. Prior to engaging in motor skill tests, participants are required to complete a questionnaire with non-personally identifiable information (non-PII) (see Fig. 2).

Alongside the collected features, the motor skill tests are described in the following sections.

2.1. Digitized motor tests

The mobile application includes four tests. Two of these tests (the first and third) are reaction tests, which have been proven as a valid assessment method for determining mental fatigue [31] [32].

2.1.1. Reaction test simple (RTS)

The first reaction test (see Fig. 3a) consists of 15 randomly rendered targets on the screen with varying sizes (45-150 pixels). The objective

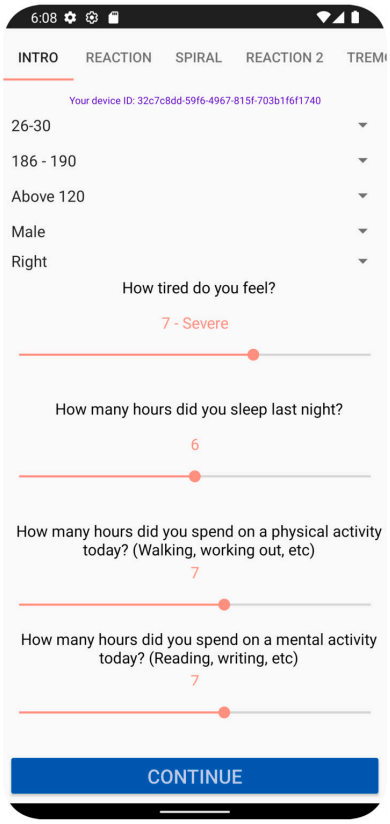


Fig. 2. Screen view of the questionnaire for collecting non-PII. The provided ranges are as follows - age: under 18, 18-25, 26-30, 31-35, above 35 (these ranges were selected to target the most probable participants' age range); height (cm): under 100, 101-150, 151-175, 176-185, 186-190, 191-205, above 205; weight (kg): under 50, 50-60, 61-75, 76-90, 91-105, 106-120, above 120; gender: female, male, other; dominant hand: left, right, ambidextrous; self-assessed severity of tiredness on a scale of 1-10 (descriptors were provided for each option of tiredness: 1 - No exhaustion, 2 - Very very slight, 3 - Very slight, 4 - Slight, 5 - Moderate, 6 - Somewhat severe, 7 - Severe, 8 - Very severe, 9 - Very very severe, 10 - Maximal); hours slept: 0-12 (any above 12 is considered 12 hours); hours spent on physical activity: 0-12; hours spent on mental activity: 0-12.

is to click on each target as quickly as possible. Data collection starts after the first target is hit, recording every input including its location, reaction time, and identifying missed moves.

2.1.2. Archimedean spiral drawing test (ASD)

The Archimedean spiral drawing (ASD) test (see Fig. 3b) is the second test of the suite, where the user must draw in the whitespace between the lines. Spiral analysis has long been classified as a clinically valid method for objectively evaluating disorders such as Parkinson's disease or tremor disorder [33][34]. Data collected from the drawing is acquired every 15 - 17 milliseconds, and the $x-$ and $y-$coordinates are recorded, along with timestamps and finger position. From this, a large selection of kinematic and geometric variables is computed (see Section 3.1).

2.1.3. Reaction test advanced (RTA)

Another version of the aim test was also included in the suite, which follows the principle of a reaction test but with added difficulty (see Fig. 3c). This version uses the color-matching logic from the Stroop test, requiring the test-taker to only react to a target that shares the same color as a rendered color guide. This version also records a point of reaction time from the correct color hit.

2.1.4. Tremor test

The final test in the suite is the tremor test (see Fig. 3d). While previous studies did not find a connection between hand tremors and mental fatigue, they did not assess different aspects of tremors relevant to fatigue. Since human fine motor skills are asymmetric, with the dominant hand typically being more precise, changes in motor asymmetry may occur under limited resources, such as fatigue. To explore this, we designed a test to measure the asymmetry between left- and right-hand tremor activity. In this test, the participant holds the phone in front with their hand fully extended, facing the tester, and waits for 10 seconds. A successful test completion requires passing the tremor test using both hands. The initial position serves as calibration, and all movement changes are recorded from that point. Acceleration data on the $X-$, $Y-$, and $Z-$planes are captured every 15-17 milliseconds throughout the test, providing a record of overall time and accelerations for all planes.

3. Data collection and analysis

Raw time series described in the previous section: finger position ($x-$ and $y-$coordinates) and timestamps can be used to compute many features. This section and Table 1 describe the engineered feature set.

3.1. Feature extraction and engineering

Kinematic features: Given a respective timestamp, we can calculate the velocity of the position vector $\vec{r} = [p_i, p_{i+1}]$. In other words, velocity is the rate at which displacement of the position vector changes with respect to time. Similarly, acceleration was computed as the rate of change in velocity and jerk as the rate of change in acceleration with respect to time. Following the sequence, we considered up to the sixth time derivative of the position vector. There are no universally accepted names for the fourth and higher time derivatives of displacement. However, the terms snap, crackle, and pop are used in literature for the fourth, fifth, and sixth time derivatives of displacement [35]. These high-order derivative features, which can be interpreted as microchanges in movement acceleration, were introduced in [36] in the context of Parkinson's disease diagnostics. Fig. 4 gives a visual representation of described differential-type features.

Angular features: Given the slope k of the position vector, we can extract angle α. Let N be the number of observation points and (x_i, y_i) are the coordinates of the point p_i, where $i \in \{1, 2, \ldots, N\}$, then the slope (k) and respective angle are represented as follows:

$$k = \frac{y_i - y_{i-1}}{x_i - x_{i-1}}, \tag{1}$$

$$\alpha = \arctan k, \tag{2}$$

Fig. 4 depicts all the angles that are considered in the current research:

$$\phi_i = \pi + \alpha_{i-1} - \alpha_i \tag{3}$$

$$\gamma_i = \alpha_i - \alpha_{i-1} \tag{4}$$

Yaw (γ) is described as the change in direction in which the point vector is pointing. The angular feature set was enriched with up to a third of respective time derivatives.

A study conducted by [37] showed that the tuple of integral-like features computed based on kinematic parameters and pressure possess sufficiently high discriminating power to distinguish Parkinson's disease patients from healthy control subjects. In [38], it was also demonstrated that these features might allow machine learning techniques to detect

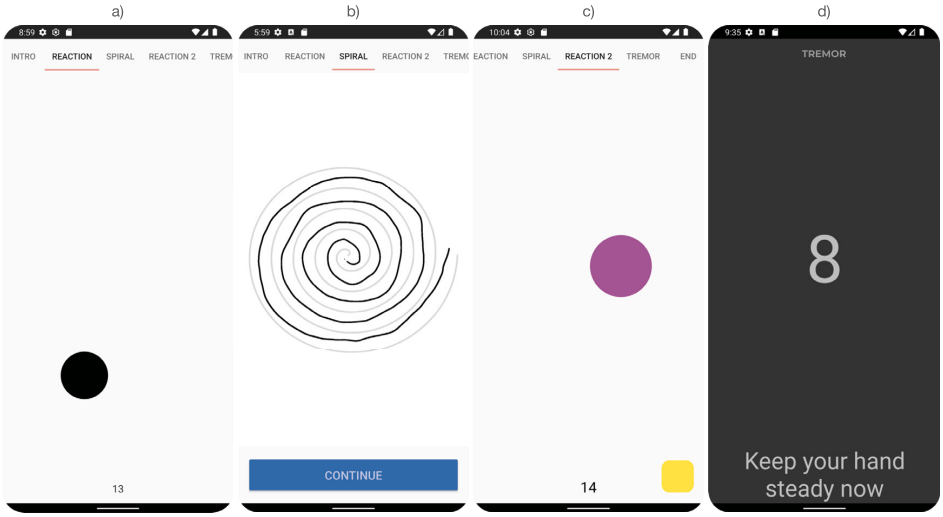


Fig. 3. Screen views of the four digitized motor skill tests.

Table 1
Subset of computed features. A total of 51 features were computed for the ASD test. The given subset describes the notation for better understanding. The tremor values were defined as an asymmetry between the left- and the right-hand points, computed as the subtraction.

Test name	Feature set	Description
ASD	distance	$d_i = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}$ (Euclidean distance)
	acceleration	Rate of change in velocity with respect to time. The second time derivative of the distance.
	ϕ_angle_mass	Mass of the angle ϕ (in radians), see Fig. 4
	x_jerk_mass	Mass of the rate of change in x-directional acceleration
	crackle_mass	Mass of the fifth time derivative of the distance
RTS and RTA	wasHitOnTarget	Boolean values True if the area of the touch overlaps with at least one pixel of the rendered circle.
	timeFromLastTouch	Time between touches
	timeFromFirstCorrect-ColorRender	The difference in time between two matching color renders
Tremor	x, y, z	Acceleration along $x-, y-, z-$axis
	absolute acceleration	$abs = \sqrt{x^2 + y^2 + z^2}$

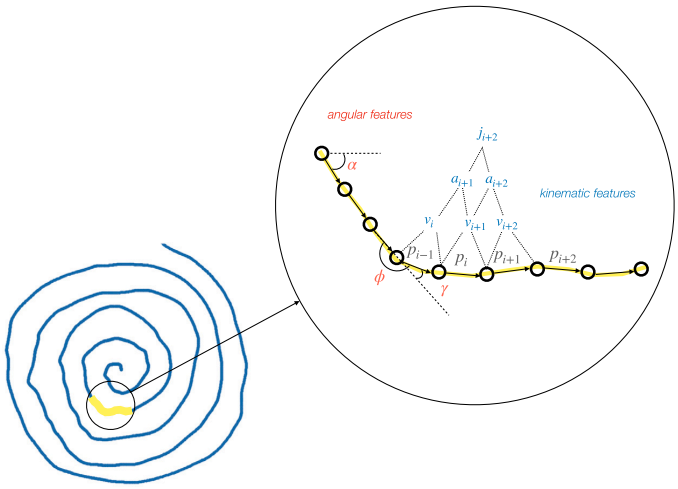


Fig. 4. Visual representation of the differential-type (a) and angular-type (b) features.

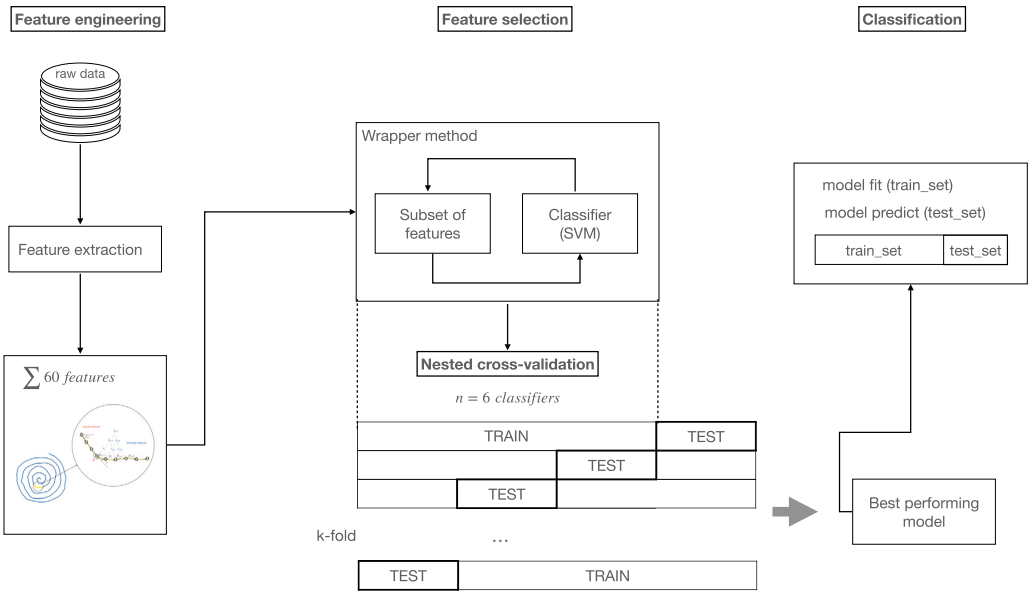


Fig. 5. Machine learning pipeline.

mental fatigue; therefore, these features are included in the present research. For the sake of self-sufficiency, the computational procedure of the motion parameters is described in the following paragraph.

Motion mass parameters: Motion mass parameters were introduced by [39] to describe the amount and smoothness of motion of a limb or some other group of joints. A sum of the absolute values at each observation point may be computed for each kinematic and geometric parameter that changes during the test. Let N be the number of observation points in the test (or a part of the test). Denote v_k the velocity along the directional vector of the stylus movement at observation point k where $k \in \{1, \dots, N\}$ then *velocity mass* is defined by equation

$$V_N = \sum_{k=1}^{N} |v_k| \tag{5}$$

3.2. Database for fatigue assessment through digital fine-motor skill tests (SmartPhoneFatigue)

We created a database of digital signals from 41 subjects completing motor skill tests and self-assessing their fatigue levels. The Tallinn University Board of Ethics 12.05.2021 decision nr 12 regulated the data collection process. Overall, 157 tests were collected from 41 test subjects. Data cleaning (removal of faulty tests, outlier detection) resulted in 131 trials eligible for further analysis. Detailed non-PI information about participants is described in Appendix A Table 5. Each row of the specific test corresponds to the digital signals collected for every timestamp (see Section 3.1). The shapes of the datasets for every test are described in Appendix A in Table 6.

3.3. Machine learning pipeline

A total of 60 features were engineered from the raw signals. Most discriminative predictors were selected to reduce dimensionality using wrapper-type feature selection procedures. We consider the SVM recursive feature elimination (SVM-RFE) wrapper method proposed by [40]. Six machine learning classifiers were used for the classification and research:

Table 2
Fatigue categories for classification.

Fatigue category	Threshold	Label
Physical exertion (PEF)	== 0	not tired (32)
	>= 1	tired (99)
Mental exertion (MEF)	== 0	not tired (84)
	>= 1	tired (47)
Sleep hours (SHF)	>= 7	not tired (62)
	<= 6	tired (69)
Self-assessed (SAF)	<= 3	not tired (37)
	>= 6	tired (47)

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Decision Tree (DT)
- Random Forest (RF)
- AdaBoost (AB)

These were trained and cross-validated in a nested k-fold loop. It means that supervised feature selection strategies are nested within the cross-validation iterations so that the most discriminating features are chosen based only on the training set, while the test set is kept only for validation [41]. This way, we avoid possible bias that may lead to model overfitting. Training and validation of the classifiers were performed using the scikit-learn library for Python [42]. Accuracy, precision, sensitivity, specificity, and f1-score were used to assess the performance of classifiers. The fatigue categories are presented in Table 2, where the target value (tired, not tired) is classified based on physical or mental exertion, sleep hours, and self-assessed tiredness.

The overall workflow is depicted in Fig. 5.

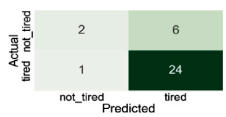
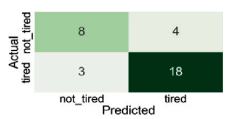
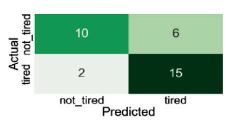
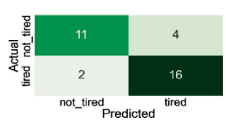
4. Results

Across the iterations of a 5-fold cross-validation scheme, the averaged mean accuracy, precision, sensitivity, specificity, and f1-score values are reported. Table 3 shows the resulting scores for

Table 3
Fatigue classification based on **self-assessed levels of tiredness**. Cross-validated model performance. The best scores for each test are presented in **bold**.

Test	Features	Classifier	P_{acc}	P_{prec}	P_{sen}	P_{spec}	P_{f1}
ASD test	ϕ_angle_mass, crackle_mass	LR	65.29%	65.29%	87.55%	37.50%	**73.96%**
		RF	60.59%	63.00%	76.44%	40.00%	68.10%
		KNN	**66.69%**	**73.30%**	72.00%	**58.57%**	70.03%
		SVM	62.94%	62.06%	**89.56%**	29.29%	65.46%
		DT	55.81%	60.74%	76.67%	29.64%	65.46%
		AB	59.41%	62.26%	74.44%	40.36%	67.01%
RTS test	wasHitOnTarget, timeFromLastTouch	LR	**59.34%**	60.16%	**78.44%**	34.64%	**68.01%**
		RF	45.37%	49.34%	54.89%	31.79%	51.07%
		KNN	55.88%	**60.24%**	61.11%	**48.57%**	60.20%
		SVM	52.35%	56.88%	63.11%	37.86%	58.96%
		DT	47.65%	51.02%	48.67%	46.07%	49.24%
		AB	54.63%	58.00%	63.11%	42.50%	59.83%
RTA test	timeFromLastTouch, timeFromFirstCorrect-ColorRender	LR	**66.76%**	65.23%	**86.87%**	40.36%	**74.40%**
		RF	58.38%	62.44%	65.56%	47.86%	63.58%
		KNN	61.91%	**65.50%**	70.22%	**51.07%**	67.22%
		SVM	63.09%	62.63%	86.67%	31.79%	72.08%
		DT	49.92%	55.41%	50.67%	48.21%	52.61%
		AB	54.93%	58.42%	63.33%	42.86%	60.35%
Tremor test	*y*, *abs*	LR	58.30%	58.26%	**89.33%**	18.93%	**70.34%**
		RF	**68.31%**	64.40%	58.89%	**57.14%**	59.74%
		KNN	55.88%	58.82%	67.56%	40.36%	62.27%
		SVM	51.18%	53.68%	89.11%	0.025%	66.97%
		DT	56.03%	65.71%	63.11%	46.43%	60.46%
		AB	65.51%	**69.59%**	71.55%	56.79%	69.36%
All tests joined	α_velocity_mass, crackle_mass, y_acceleration_mass, α_jerk, yaw_acceleration, velocity, *y*, *z*, *abs*, timeFromLastTouch	LR	74.00%	67.71%	88.00%	61.33%	**76.18%**
		RF	**74.00%**	**78.00%**	76.00%	**74.67%**	74.11%
		KNN	72.00%	64.88%	**92.00%**	53.33%	75.93%
		SVM	70.00%	63.81%	88.00%	54.00%	73.64%
		DT	64.00%	62.33%	72.00%	58.67%	65.21%
		AB	62.00%	61.67%	60.00%	65.33%	59.33%

Table 4
Best performing machine learning models for fatigue classification.

Fatigue category	Features	Classifier	Performance (%)					Confusion Matrix
			P_{acc}	P_{sen}	P_{spec}	P_{prec}	P_{f1}	
PEF	(ASD) ϕ_angle_mass, crackle_mass	KNN	78.8	96.0	25.0	80.0	87.3	
MEF	(ASD) *x*_jerk_mass, distance, acceleration	RF	78.8	85.7	66.7	81.8	83.7	
SHF	(RTA) timeFromLastTouch, timeFromFirstCorrectColorRender	RF	75.8	88.2	62.5	71.4	78.9	
SAF	(ALL TESTS) α_velocity_mass, crackle_mass, y_acceleration_mass, α_jerk, γ_acceleration, velocity, *y*, *z*, *abs*, timeFromLastTouch	RF	75.8	83.3	66.7	79.0	84.2	

self-assessed fatigue levels for a better understanding of the background procedures. Based on the cross-validated performance, the best-performing classifiers with the respective feature set were trained on the whole dataset (split 1/3 test/train set) for each fatigue category. Table 4 describes the final models with a respective performance review.

All six classifiers exhibited comparable performance, with a slight advantage observed for the KNN and RF classifiers. The ASD test, complemented by RTA, emerged as the most informative assessment for detecting fatigue. Moreover, for certain fatigue categories determined through self-assessment (SAF), the combination of all tests yielded the most favorable outcomes. Trajectory angles (e.g. ϕ_angle_mass) and micro-changes in acceleration (e.g. crackle_mass) as described in Section 4, proved to be highly informative in detecting fatigue. These features capture nuanced fluctuations in fine motor movement, providing valuable insights into the effects of fatigue on motor performance. This highlights the capability of machine learning algorithms to discern between these two states based on the study's utilized features, which, although subtle and imperceptible to the naked eye, possess informative value for classification.

5. Discussion

The work aims to investigate the relationships between motor skill tests and the level of tiredness, workload, and hours of sleep assessed by a person. First, we were interested in how well fatigue can be predicted by kinematic and angular features extracted from the ASD test. The proposed features exhibited near-perfect sensitivity in identifying fatigued individuals based on the characteristics of physical exertion. Furthermore, this finding highlights the importance of kinematic features in detecting fatigue, specifically trajectory angles and micro-changes in acceleration of fine motor movements. This suggests that subtle variations in movement patterns, undetectable to the unaided eye, and motor control can serve as valuable indicators of fatigue. Given their potential significance, it is imperative to conduct further research to fully explore the implications and mechanisms behind these subtle variations. Additionally, the introduction of *asymmetry* as a new characteristic is significant. Tracking its development may hold potential as an aspect of stroke rehabilitation, warranting further discussion.

To enhance the specificity score, it is necessary to include additional non-tired samples. The expected low specificity arises from the fact that patterns among tired individuals exhibit less variation compared to non-tired individuals. In our future research, we intend to develop models tailored to individual users, as human motor ability and its expression (the degree of "clumsiness") in a fatigued state can differ significantly [43]. This personalized approach aims to train the model using internal measurements specific to each individual, ultimately creating a personalized "fatigue meter." Furthermore, in the context of regular application usage, it is vital to consider proficiency as a separate variable. Although the limited number of trials in the present study did not significantly affect proficiency, future studies involving a greater number of repeated trials must account for its potential influence. By addressing this factor, we can achieve a more comprehensive understanding of the application's effectiveness.

Our study identifies weaknesses in the current approach. While the machine learning model achieved acceptable accuracy, we believe that more accurate oversight of test takers could further enhance results. Providing more detailed tutorials warrants further experimentation. Additionally, the exclusion of pressure signals poses a potential limitation, as pressure-related features are relevant in digitized ASD tests [34][36][28]. Future work will address these obstacles and incorporate analysis of pressure-related data.

Artificial intelligence has shown high accuracy (over 90%) in identifying motor-impairing diseases like Parkinson's disease [44]. However, distinguishing fatigue through motor performance poses challenges as fatigue's manifestations and definitions remain unclear. These questions drive current and future research in fatigue analysis. Smartphone-based systems offer real-time fatigue assessment, enabling proactive measures from preventing accidents to supporting cancer patients experiencing treatment-induced fatigue. Tracking fatigue levels over time with these systems can inform treatment effectiveness and enhance patient outcomes.

6. Conclusion

A framework was presented that utilizes a smartphone application to facilitate the collection and analysis of data for fatigue assessment. The proposed system incorporates four motor skill tasks for capturing digital signals, along with a questionnaire to gather subjective measures of fatigue. By examining the relationships between motor skill tests and various measures of fatigue, including a level of tiredness, workload, and hours of sleep, the authors aim to develop a more comprehensive understanding of the factors that contribute to fatigue. The SmartPhoneFatigue database, including digital signals from four motor skill tests performed by 41 subjects, is introduced and available upon request. Additionally, a methodology for creating a predictive fatigue model, using 60 distinct features such as kinematic, angular, aim-based, and tremor-related measures, has been developed to exhibit the database's practical applications. Through the integration of these features, our findings demonstrate an acceptable level of accuracy in classifying fatigue, surpassing 70%. Moreover, we observe high sensitivity (above 90%) and f1-score (above 80%). These results emphasize the importance of imperceptible variations in movement patterns and motor control as crucial indicators for fatigue detection. Our database and methodology exhibit substantial potential for enhancing our comprehension of fatigue and fostering the development of more effective management strategies. The potential benefits extend beyond accident prevention to broader healthcare applications. Standardized data collection and testing of subjects on multiple occasions is necessary, and it is the aim of our future research. Overall, the proposed system could potentially be used as a tool for monitoring fatigue levels in real-world settings, providing a more convenient and less resource-expensive approach compared to traditional methods.

7. Summary points

Problem: Fatigue presents a significant challenge that can result in accidents, mistakes, and decreased productivity across numerous industries. However, fatigue assessment in non-controlled environments is not trivial.

What is already known: Traditional approaches to fatigue assessment, such as self-reporting and physiological measures (eye movement trackers, heart monitor devices), have limitations regarding accuracy, accessibility, and/or practicality. Technology advancements have led to the development of smartphone-based fatigue assessment systems, offering a more objective and convenient method for measuring fatigue.

What this paper adds: Our research proposes and evaluates a novel smartphone-based application of digital motor skill tasks to assess fatigue. A total of 60 features (kinematic, geometric, and other) were engineered from the raw signals. Through machine learning-based evaluation, the set of features with the best performance was identified, resulting in the detection of fatigue with high levels of sensitivity (over 90%). The collected data is assembled in a novel dataset (SmartPhone-Fatigue) and is presented for further research.

CRediT authorship contribution statement

Elli Valla: Methodology, Software, Investigation, Formal analysis, Writing. **Ain-Joonas Toose:** Software (mobile application), Investigation, Formal analysis. **Sven Nõmm:** Methodology, Supervision, Conceptualization. **Aaro Toomela:** Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work of Sven Nõmm and Elli Valla in the project 'ICT programme' was supported by the European Union through the European Social Fund, Project number 2014-2020.4.05.19-0001.

Appendix A

Table 5
Detailed non-PII about participants.

id	height	weight	age	gender	tired scale	mental work scale	physical work scale	sleep scale	main hand	created
1	176 - 185	76-90	>35	Male	5	0	1	7	R	19:23:06
	176 - 185	76-90	>35	Male	6	0	1	8	R	19:27:42
	176 - 185	76-90	>35	Male	2	0	1	7	R	05:39:01
	176 - 185	76-90	>35	Male	8	0	1	7	R	19:45:22
	176 - 185	76-90	>35	Male	5	0	1	7	R	04:48:27
	176 - 185	76-90	>35	Male	4	0	1	7	R	04:57:43
2	151 - 175	50-60	31-35	Female	1	3	8	9	R	07:19:23
	151 - 175	50-60	31-35	Female	1	3	8	9	R	07:26:22
3	151 - 175	50-60	31-35	Female	4	5	2	8	R	15:33:42
4	176 - 185	76-90	31-35	Female	4	6	2	7	R	07:12:47
	176 - 185	76-90	31-35	Female	4	1	2	8	R	15:17:19
5	176 - 185	76-90	31-35	Male	2	3	1	8	R	10:39:25
6	176 - 185	76-90	>35	Male	6	6	2	6	R	14:02:24
7	186 - 190	>120	26-30	Male	6	6	7	6	R	23:48:06
	186 - 190	>120	26-30	Male	6	6	7	6	R	23:41:52
	101 - 150	50-60	18-25	Male	1	6	7	5	L	22:29:41
	186 - 190	>120	26-30	Male	6	6	7	6	R	00:02:21
	186 - 190	>120	26-30	Male	7	7	7	6	R	18:40:16
	186 - 190	>120	26-30	Male	6	6	7	6	R	22:59:20
	151 - 175	<50	26-30	Female	7	7	7	12	A	20:48:17
	186 - 190	>120	26-30	Male	6	6	7	6	R	23:58:56
	186 - 190	>120	26-30	Male	7	7	7	6	R	00:19:49
8	176 - 185	91-105	26-30	Male	5	7	3	7	R	21:18:13
	176 - 185	91-105	26-30	Male	2	0	0	10	R	08:50:10
	176 - 185	91-105	26-30	Male	4	0	0	9	R	07:08:03
	176 - 185	91-105	26-30	Male	5	6	2	7	R	22:06:37
	176 - 185	91-105	26-30	Male	1	5	0	7	R	08:10:34
	176 - 185	91-105	26-30	Male	5	6	3	7	R	21:27:49
9	176 - 185	76-90	<18	Male	8	5	1	5	R	11:30:49
10	151 - 175	50-60	<18	Female	5	1	0	4	R	09:14:46
11	186 - 190	61-75	18-25	Male	4	0	0	12	R	19:23:43
	186 - 190	61-75	18-25	Male	1	0	0	9	R	13:48:33
	186 - 190	61-75	18-25	Male	4	0	0	12	R	20:07:00
	186 - 190	61-75	18-25	Male	3	0	0	12	R	21:36:53
12	151 - 175	61-75	18-25	Female	6	6	1	6	L	08:32:35
	151 - 175	61-75	18-25	Female	6	7	1	4	L	19:24:06
	151 - 175	61-75	18-25	Female	6	8	11	5	L	23:26:45
13	176 - 185	61-75	18-25	Male	4	0	1	8	R	11:22:15
	176 - 185	61-75	18-25	Male	3	0	2	6	R	10:01:39
14	151 - 175	61-75	>35	Female	3	0	0	7	R	06:09:44
	151 - 175	61-75	>35	Female	3	0	3	7	R	21:06:47
	151 - 175	61-75	>35	Female	2	0	0	10	R	07:20:48
	151 - 175	61-75	>35	Female	6	1	3	10	R	22:03:05
	151 - 175	61-75	>35	Female	6	6	0	6	R	20:19:06
	151 - 175	61-75	>35	Female	4	0	0	7	R	04:34:05
	151 - 175	61-75	>35	Female	5	7	1	8	R	21:03:35
	151 - 175	61-75	>35	Female	3	0	0	8	R	04:32:43
	151 - 175	61-75	>35	Female	4	7	1	7	R	19:46:12
	151 - 175	61-75	>35	Female	2	0	0	7	R	04:36:52
	151 - 175	61-75	>35	Female	4	9	2	7	R	20:57:21
	151 - 175	61-75	>35	Female	2	0	0	6	R	04:49:18
	151 - 175	61-75	>35	Female	5	8	2	7	R	20:35:33
	151 - 175	61-75	>35	Female	2	9	0	7	R	04:20:16
15	151 - 175	50-60	18-25	Male	7	7	1	7	R	22:50:31
	151 - 175	50-60	18-25	Male	5	4	0	4	R	08:37:25
	151 - 175	50-60	18-25	Male	7	5	0	4	R	22:44:24
16	151 - 175	50-60	18-25	Female	1	0	0	6	R	08:32:07
	151 - 175	50-60	18-25	Female	1	2	0	7	R	08:35:22

Table 5 (*continued*)

id	height	weight	age	gender	tired scale	mental work scale	physical work scale	sleep scale	main hand	created
17	186 - 190	>120	26-30	Male	4	0	4	7	R	16:10:43
	186 - 190	>120	26-30	Male	5	9	2	9	R	21:18:46
	186 - 190	>120	26-30	Male	3	0	1	7	R	12:01:49
18	186 - 190	76-90	<18	Male	1	0	0	8	R	09:42:52
	186 - 190	76-90	<18	Male	1	1	1	7	R	09:14:27
	186 - 190	76-90	<18	Male	1	1	2	7	R	16:23:49
	186 - 190	76-90	<18	Male	1	0	0	8	R	22:16:27
19	151 - 175	50-60	<18	Female	6	0	7	6	R	05:22:17
	151 - 175	50-60	<18	Female	6	0	7	6	R	09:13:01
	151 - 175	50-60	<18	Female	6	0	7	6	R	16:27:27
	151 - 175	50-60	<18	Female	6	0	7	6	R	10:24:21
	151 - 175	50-60	<18	Female	6	0	7	6	R	16:52:46
	151 - 175	50-60	<18	Female	6	0	7	6	R	11:53:29
	151 - 175	50-60	<18	Female	10	0	1	5	R	17:35:24
	151 - 175	50-60	<18	Female	8	0	1	8	R	19:35:56
	151 - 175	50-60	<18	Female	6	0	7	6	R	14:26:01
20	186 - 190	61-75	<18	Male	1	0	0	0	R	09:08:47
21	-	76-90	31-35	Female	4	2	0	6	R	15:22:16
22	151 - 175	50-60	<18	Male	4	2	3	6	R	09:15:19
23	151 - 175	>120	>35	Male	3	1	1	6	R	07:45:18
	151 - 175	>120	>35	Male	4	4	3	5	R	18:33:48
	151 - 175	>120	>35	Male	7	7	1	6	R	21:01:18
	151 - 175	>120	>35	Male	3	2	0	7	R	06:52:22
	151 - 175	>120	>35	Male	5	4	1	5	R	18:43:04
	151 - 175	>120	>35	Male	1	6	1	6	R	07:05:41
	151 - 175	>120	>35	Male	5	4	3	7	R	18:37:31
	151 - 175	>120	>35	Male	7	7	1	6	R	21:20:48
	151 - 175	>120	>35	Male	7	5	1	4	R	02:57:52
	151 - 175	>120	>35	Male	6	6	0	9	R	19:34:23
	151 - 175	>120	>35	Male	6	6	0	7	R	19:33:33
	151 - 175	>120	>35	Male	4	1	1	8	R	06:49:35
	151 - 175	>120	>35	Male	4	5	1	6	R	03:00:32
	151 - 175	>120	>35	Male	7	4	1	7	R	03:02:55
	151 - 175	>120	>35	Male	5	6	0	6	R	04:02:22
24	176 - 185	61-75	<18	Female	7	4	1	3	L	11:31:15
25	151 - 175	<50	<18	Female	8	1	1	7	R	17:32:34
	151 - 175	<50	<18	Female	8	1	1	6	R	09:12:34
26	176 - 185	61-75	18-25	Female	4	0	0	10	R	07:42:08
27	176 - 185	91-105	18-25	Female	4	4	3	5	R	20:17:56
	176 - 185	91-105	18-25	Female	2	4	1	5	R	09:12:00
28	151 - 175	<50	<18	Female	6	3	0	5	R	22:19:46
29	151 - 175	50-60	<18	Male	4	0	7	5	R	21:46:17
	151 - 175	50-60	<18	Male	4	0	7	5	R	09:49:56
	151 - 175	50-60	<18	Male	4	0	7	5	R	07:57:23
	151 - 175	50-60	<18	Male	4	0	7	5	R	15:34:09
	151 - 175	50-60	<18	Male	4	0	7	5	R	18:22:47
30	151 - 175	91-105	26-30	Male	7	0	2	5	R	21:44:27
	151 - 175	91-105	26-30	Male	7	0	0	5	R	11:09:16
	151 - 175	91-105	26-30	Male	8	0	2	4	R	12:36:07
31	151 - 175	<50	18-25	Female	5	0	0	6	R	07:26:52
	151 - 175	<50	18-25	Female	7	1	1	7	R	22:00:45
32	151 - 175	<50	<18	Female	2	4	3	6	A	18:24:54
	151 - 175	<50	<18	Female	1	0	1	8	A	12:01:54
	151 - 175	<50	<18	Female	6	10	2	6	A	18:04:38
	151 - 175	<50	<18	Female	2	6	2	6	A	12:38:57
33	151 - 175	61-75	<18	Female	5	3	7	5	R	07:46:33
	151 - 175	61-75	<18	Female	5	3	7	5	R	15:25:48
34	176 - 185	76-90	<18	Male	3	6	2	5	R	09:09:44
	176 - 185	76-90	<18	Male	2	6	1	6	R	06:49:44
	176 - 185	76-90	<18	Male	5	2	1	8	R	14:35:05
35	191 - 205	91-105	<35	Male	4	7	1	5	R	09:46:18
	191 - 205	91-105	<35	Male	4	7	1	5	R	15:22:55
36	176 - 185	61-75	18-25	Male	7	4	4	4	R	11:27:41
37	151 - 175	<50	<18	Female	3	10	2	11	R	19:39:38
	151 - 175	<50	<18	Female	2	10	3	8	R	13:22:29
38	176 - 185	61-75	18-25	Male	3	1	1	5	R	10:01:25

(*continued on next page*)

Table 5 (*continued*)

id	height	weight	age	gender	tired scale	mental work scale	physical work scale	sleep scale	main hand	created
39	186 - 190	<120	26-30	Male	7	12	3	6	R	20:17:48
	186 - 190	>120	26-30	Male	7	12	3	6	R	01:35:57
	186 - 190	>120	26-30	Male	4	6	2	7	R	15:39:24
40	151 - 175	61-75	<18	Female	5	5	3	8	R	19:08:54
41	151 - 175	61-75	18-25	Female	7	12	2	6	R	16:38:51
	151 - 175	61-75	18-25	Female	4	2	2	12	R	14:37:46
	151 - 175	61-75	18-25	Female	4	8	2	6	R	17:17:44

Table 6
Characteristics of the database.

	Dataframe by test	Shape (timestamp x features)
1	Reaction Test Simple test (RTSdata)	2057 rows x 22 columns
2	Reaction Test Advanced (RTAdata)	2562 rows x 23 columns
3	Archimedean Spiral Drawing test (ASDdata)	131000 rows x 83 columns
4	Tremor test right hand (TTRdata)	66134 rows x 20 columns
5	Tremor test left hand (TTLdata)	66083 rows x 20 columns

Appendix B. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2023.105152.

References

[1] S. Dey, S.A. Chowdhury, S. Sultana, M.A. Hossain, M. Dey, S.K. Das, Real time driver fatigue detection based on facial behaviour along with machine learning approaches, in: 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON), IEEE, 2019, pp. 135–140.

[2] K. Martin, R. Meeusen, K.G. Thompson, R. Keegan, B. Rattray, Mental fatigue impairs endurance performance: a physiological explanation, Sports Med. 48 (2018) 2041–2051.

[3] S.M. Marcora, W. Staiano, V. Manning, Mental fatigue impairs physical performance in humans, J. Appl. Physiol. (2009).

[4] M.A. Boksem, T.F. Meijman, M.M. Lorist, Effects of mental fatigue on attention: an ERP study, Cogn. Brain Res. 25 (1) (2005) 107–116, https://doi.org/10.1016/j.cogbrainres.2005.04.011, https://www.sciencedirect.com/science/article/pii/S0926641005001187.

[5] M.M. Lorist, M.A. Boksem, K.R. Ridderinkhof, Impaired cognitive control and reduced cingulate activity during mental fatigue, Cogn. Brain Res. 24 (2) (2005) 199–205.

[6] V. Rozand, F. Lebon, P.J. Stapley, C. Papaxanthis, R. Lepers, A prolonged motor imagery session alter imagined and actual movement durations: potential implications for neurorehabilitation, Behav. Brain Res. 297 (2016) 67–75.

[7] M.E. Harrington, Neurobiological studies of fatigue, Prog. Neurobiol. 99 (2) (2012) 93–105.

[8] R.K. Portenoy, L.M. Itri, Cancer-related fatigue: guidelines for evaluation and management, The Oncologist 4 (1) (1999) 1–10.

[9] G.A. Curt, The impact of fatigue on patients with cancer: overview of fatigue 1 and 2, The Oncologist 5 (S2) (2000) 9–12.

[10] V.J. Gawron, J. French, D. Funke, An overview of fatigue, in: Stress, Workload, and Fatigue, 2000, pp. 581–595.

[11] A.G. Bills, Blocking: a new principle of mental fatigue, Am. J. Psychol. 43 (2) (1931) 230–245.

[12] M.A. Boksem, T.F. Meijman, M.M. Lorist, Mental fatigue, motivation and action monitoring, Biol. Psychol. 72 (2) (2006) 123–132, https://doi.org/10.1016/j.biopsycho.2005.08.007.

[13] M.A. Boksem, M. Tops, Mental fatigue: costs and benefits, Brains Res. Rev. 59 (1) (2008) 125–139.

[14] M.A. Boksem, T.F. Meijman, M.M. Lorist, Effects of mental fatigue on attention: an ERP study, Cogn. Brain Res. 25 (1) (2005) 107–116, https://doi.org/10.1016/j.cogbrainres.2005.04.011.

[15] D.M. Raizen, J. Mullington, C. Anaclet, G. Clarke, H. Critchley, R. Dantzer, R. Davis, K.L. Drew, J. Fessel, P.M. Fuller, E.M. Gibson, M. Harrington, W. Ian Lipkin, E.B. Klerman, N. Klimas, A.L. Komaroff, W. Koroshetz, L. Krupp, A. Kuppuswamy, J. Lasselin, L.D. Lewis, P.J. Magistretti, H.Y. Matos, C. Miaskowski, A.H. Miller, A. Nath, M. Nedergaard, M.R. Opp, M.D. Ritchie, D. Rogulja, A. Rolls, J.D. Salamone, C. Saper, V. Whittemore, G. Wylie, J. Younger, P.C. Zee, H. Craig Heller, Beyond the symptom: the biology of fatigue, Sleep (2023) zsad069, https://doi.org/10.1093/sleep/zsad069, https://academic.oup.com/sleep/advance-article-pdf/doi/10.1093/sleep/zsad069/50603785/zsad069.pdf.

[16] R. Guidoux, M. Duclos, G. Fleury, P. Lacomme, N. Lamaudière, P.-H. Manenq, L. Paris, L. Ren, S. Rousset, A smartphone-driven methodology for estimating physical activities and energy expenditure in free living conditions, J. Biomed. Inform. 52 (2014) 271–278.

[17] M. Kay, J. Santos, M. Takane, mHealth: new horizons for health through mobile technologies, World Health Organ. 64 (7) (2011) 66–71.

[18] A.Z. Antosik-Wójcińska, M. Dominiak, M. Chojnacka, K. Kaczmarek-Majer, K.R. Opara, W. Radziszewska, A. Olwert, Ł. Świkecicki, Smartphone as a monitoring tool for bipolar disorder: a systematic review including data analysis, machine learning algorithms and predictive modelling, Int. J. Med. Inform. 138 (2020) 104131.

[19] M. Kheirkhahan, S. Nair, A. Davoudi, P. Rashidi, A.A. Wanigatunga, D.B. Corbett, T. Mendoza, T.M. Manini, S. Ranka, A smartwatch-based framework for real-time and online assessment and mobility monitoring, J. Biomed. Inform. 89 (2019) 29–40.

[20] Y. Fukazawa, T. Ito, T. Okimura, Y. Yamashita, T. Maeda, J. Ota, Predicting anxiety state using smartphone-based passive sensing, J. Biomed. Inform. 93 (2019) 103151.

[21] V.P. Cornet, R.J. Holden, Systematic review of smartphone-based passive sensing for health and wellbeing, J. Biomed. Inform. 77 (2018) 120–132.

[22] X. Hu, G. Lodewijks, Exploration of the effects of task-related fatigue on eye-motion features and its value in improving driver fatigue-related technology, Transp. Res., Part F Traffic Psychol. Behav. 80 (2021) 150–171, https://doi.org/10.1016/j.trf.2021.03.014.

[23] G.M. Migliaccio, G. Di Filippo, L. Russo, T. Orgiana, L.P. Ardigò, M.Z. Casal, L.A. Peyré-Tartaruga, J. Padulo, Effects of mental fatigue on reaction time in sportsmen, Int. J. Environ. Res. Public Health 19 (21) (2022) 14360, https://doi.org/10.3390/ijerph192114360.

[24] J.R. Stroop, Studies of interference in serial verbal reactions, J. Exp. Psychol. 18 (6) (1935) 643.

[25] Y. Wang, Y. Huang, B. Gu, S. Cao, D. Fang, Identifying mental fatigue of construction workers using EEG and deep learning, Autom. Constr. 151 (2023) 104887.

[26] V. Rozand, F. Lebon, C. Papaxanthis, R. Lepers, Effect of mental fatigue on speed–accuracy trade-off, Neuroscience 297 (2015) 219–230.

[27] Y. Le Mansec, B. Pageaux, A. Nordez, S. Dorel, M. Jubeau, Mental fatigue alters the speed and the accuracy of the ball in table tennis, J. Sports Sci. 36 (23) (2018) 2751–2759.

[28] O. Senkiv, S. Nõmm, A. Toomela, Applicability of spiral drawing test for mental fatigue modelling, IFAC-PapersOnLine 51 (34) (2019) 190–195, https://doi.org/10.1016/j.ifacol.2019.01.064.

[29] O. Lippold, The tremor in fatigue, in: Human Muscle Fatigue: Physiological Mechanisms, 1981, pp. 234–248.

[30] F. Budini, L. Labanca, M. Scholz, A. Macaluso, Tremor, finger and hand dexterity and force steadiness, do not change after mental fatigue in healthy humans, PLoS ONE 17 (8) (2022) e0272033.

[31] G.M. Migliaccio, G. Di Filippo, L. Russo, T. Orgiana, L.P. Ardigò, M.Z. Casal, L.A. Peyré-Tartaruga, J. Padulo, Effects of mental fatigue on reaction time in sportsmen, Int. J. Environ. Res. Public Health 19 (21) (2022) 14360, https://doi.org/10.3390/ijerph192114360.

[32] D. Coutinho, B. Gonçalves, D.P. Wong, B. Travassos, A.J. Coutts, J. Sampaio, Exploring the effects of mental and muscular fatigue in soccer players' performance, Hum. Mov. Sci. 58 (March) (2018) 287–296, https://doi.org/10.1016/j.humov.2018.03.004.

[33] O. Senkiv, Fatigue recognition modelling, 2018.

[34] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, M. Faundez-Zanuy, Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease, Artif. Intell. Med. 67 (2016) 39–46.
[35] R.N. Jazar, Advanced Dynamics. Rigid Body, Multibody, and Aerospace Applications, John Wiley & Sons, Inc., 2007.
[36] E. Valla, S. Nõmm, K. Medijainen, P. Taba, A. Toomela, Tremor-related feature engineering for machine learning based Parkinson's disease diagnostics, Biomed. Signal Process. Control 75 (2022) 103551.
[37] S. Nõmm, K. Bardõš, A. Toomela, K. Medijainen, P. Taba, Detailed analysis of the Luria's alternating seriestests for Parkinson's disease diagnostics, in: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018, pp. 1347–1352.
[38] O. Senkiv, S. Nõmm, A. Toomela, Applicability of spiral drawing test for mental fatigue modelling, in: 2nd IFAC Conference on Cyber-Physical and Human Systems CPHS 2018, IFAC-PapersOnLine 51 (34) (2019) 190–195, https://doi.org/10.1016/j.ifacol.2019.01.064, http://www.sciencedirect.com/science/article/pii/S2405896319300679.
[39] S. Nõmm, A. Toomela, An alternative approach to measure quantity and smoothness of the human limb motions, Est. J. Eng. 19 (4) (2013) 298–308.
[40] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1) (2002) 389–422.
[41] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, 2nd edition, Springer Series in Statistics, Springer, 2002.
[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[43] T. Smyth, Impaired motor skill (clumsiness) in otherwise normal children: a review, Child Care Health Dev. 18 (5) (1992) 283–300.
[44] M. Diaz, M. Moetesum, I. Siddiqi, G. Vessio, Sequence-based dynamic handwriting analysis for Parkinson's disease detection with one-dimensional convolutions and BIGRUs, Expert Syst. Appl. 168 (2021) 114405.

Appendix 6

VI

Xuechao Wang, Junqing Huang, Marianna Chatzakou, Sven Nõmm, Elli Valla, Kadri Medijainen, Pille Taba, Aaro Toomela, and Michael Ruzhansky. Comparison of one-two-and three-dimensional CNN models for drawing-test-based diagnostics of the Parkinson's disease. *Biomedical Signal Processing and Control*, 87:105436, 2024

Biomedical Signal Processing and Control 87 (2024) 105436

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Comparison of one- two- and three-dimensional CNN models for drawing-test-based diagnostics of the Parkinson's disease

Xuechao Wang [a,*], Junqing Huang [a], Marianna Chatzakou [a], Sven Nõmm [b], Elli Valla [b], Kadri Medijainen [c], Pille Taba [d,e], Aaro Toomela [f], Michael Ruzhansky [a,g]

[a] Department of Mathematics: Analysis, Logic and Discrete Mathematics, Ghent University, Ghent, Belgium
[b] Department of Software Science, School of Information Technologies, Tallinn University of Technology, Akadeemia tee 15a, 12618, Tallinn, Estonia
[c] Institute of Sport Sciences and Physiotherapy, University of Tartu, Puusepa 8, Tartu 51014, Estonia
[d] Department of Neurology and Neurosurgery, University of Tartu, Puusepa 8, Tartu 51014, Estonia
[e] Neurology Clinic, Tartu University Hospital, Puusepa 8, Tartu 51014, Estonia
[f] School of Natural Sciences and Health Tallinn University, Narva mnt. 25, 10120, Estonia
[g] School of Mathematical Sciences, Queen Mary University of London, London, UK

ARTICLE INFO

Keywords:
Parkinson's disease
Drawing test
Artificial intelligence
Decision support system
Deep learning models
Convolutional neural networks
CNN

ABSTRACT

Subject: In this article, convolutional networks of one, two, and three dimensions are compared with respect to their ability to distinguish between the drawing tests produced by Parkinson's disease patients and healthy control subjects.
Motivation: The application of deep learning techniques for the analysis of drawing tests to support the diagnosis of Parkinson's disease has become a growing trend in the area of Artificial Intelligence.
Methods: The dynamic features of the handwriting signal are embedded in the static test data to generate one-dimensional time series, two-dimensional RGB images and three-dimensional voxelized point clouds, and then one-, two-, and three-dimensional CNN can be used to automatically extract features for effective diagnosis.
Novelty: While there are many results that describe the application of two-dimensional convolutional models to the problem, to the best knowledge of the authors, there are no results based on the application of three-dimensional models and very few using one-dimensional models.
Main result: The accuracy of the one-, two- and three-dimensional CNN models was 59.38%, 77.73% and 82.34% in the *DraWritePD* dataset (acquired by the authors) and 63.33%, 81.33% and 82.22% in the *PaHaW* dataset (well known from the literature), respectively. For these two data sets, the proposed three-dimensional convolutional classification method exhibits the best diagnostic performance.

1. Introduction

The present paper compares the one-, two- and three-dimensional deep convolutional neural network (CNN) models for the analysis of the drawing tests used to support the diagnosis of Parkinson's disease (PD). Parkinson's disease is one of the most common neurodegenerative disorders. Its symptoms like rigidity, tremor, and non-purposeful motions severely affect the quality of patient's life. Although at the time of writing of this article there is no cure for PD, proper therapy may allow one to eliminate these symptoms or reduce their effect on motion and improve the quality of daily life. Drawing tests have been used to diagnose PD and assess its severity for nearly a century [1]. These tests require one to continue, copy, or trace the repeating pattern or contour of an object. Only paper and pen were required to perform the test, while the evaluation was performed visually by the practitioner. This method is limited by the practitioner's experience, the ability of the naked eye, and the fact that the smoothness parameters of the drawing could not be recorded for future analysis and comparison. The seminal paper [2] laid the basis for computer-aided analysis of drawing and writing tests. In [2] it was suggested to use a digital table to acquire time-stamped coordinates of the tip of the stylus and compute the kinematic parameters that describe the movements of the tip of the stylus relative to the device screen. After that, kinematic parameters can be computed that describe the drawing movements observed during the test. In addition to providing kinematic and pressure descriptions

* Corresponding author.
E-mail addresses: xuechao.wang@ugent.be (X. Wang), junqing.huang@ugent.be (J. Huang), marianna.chatzakou@ugent.be (M. Chatzakou), sven.nomm@taltech.ee (S. Nõmm), elli.valla@taltech.ee (E. Valla), kadri.medijainen@ut.ee (K. Medijainen), pille.taba@kliinikum.ee (P. Taba), aaro.toomela@tlu.ee (A. Toomela), michael.ruzhansky@ugent.be, m.ruzhansky@qmul.ac.uk (M. Ruzhansky).

https://doi.org/10.1016/j.bspc.2023.105436
Received 31 March 2023; Received in revised form 9 August 2023; Accepted 12 September 2023
Available online 22 September 2023

of the testing process, digitisation of the tests offers the possibility of performing testing before the visit to the doctor, saving valuable time and providing access for medical professionals. The digitisation of the testing procedures has greatly expanded the set of features [3] and demonstrated the importance of features based on tremor [4] to achieve highly accurate results. The results mentioned above use statistical machine learning techniques, whereas the data are presented in tabular form. Without undermining the importance of these results, it is important to mention that the evaluation procedure is different from that used by the human practitioner. This makes it difficult for the practitioner to interpret the results of digitised tests. The core difference is that a human mostly assesses the shapes of the drawn contours, smoothness of the movements (as the naked eye can see), and presence or absence of the errors, whereas statistical learning algorithms use the set of values describing the kinematic and pressure parameters. One way to mimic the human practitioner is to employ deep learning techniques to classify drawings according to their shapes [5]. Such a solution would be closer to human assessment. However, it would ignore the advantages of performing the test on digital tables or tablet PCs that can capture the kinematic and pressure parameters that describe the test. The colouration of the drawn lines was proposed in [6] to encode the pressure parameter. This may be seen as the bridge between mimicking analysis made by human practitioners and novel approaches that are based on features described by kinematic and pressure properties of the motion. Later, [7,8] expanded on this idea and suggested varying the thickness of the drawn contours to encode one more kinematic or pressure parameter in the drawings. Later in [9], the "hand-crafted" and CNN-learnt features were compared. These approaches assume that the data are provided in the form of images and that the CNN classifier is used to estimate whether the test drawing was produced by the PD patient or the healthy control (HC) subject. In theory, encoding more kinematic and pressure parameters in the drawing could further increase the goodness of the diagnostic support model [10–12]. Following this idea, applications of three-dimensional convolutional neural networks seem to be a logical step. At the same time, one-dimensional CNN has been successfully applied to similar problems [13], leading to the idea of comparing the three cases. The CNN structures best suited for the particularities of the images resulting from the drawing testing procedures will be selected first. Then, different feature sets will be selected to encode in the original drawings. The selected models will then be trained and validated to determine the quality of the models and the required training time. The organisation of the paper follows the classical academic style. An overview of the literature necessary to position the current contribution and explain its novelty is presented in Section 2. Section 3 introduces the problem statement and elaborates the research motivation. Background information explaining how symptoms of PD influence the feature engineering process is presented, together with a description of the hardware and software settings used for data acquisition in Section 4. The choice of CNN architectures to compare with all other parameters of computational experiments is presented in Section 5. The main results are stated in Section 6. The results achieved and their medical interpretation are discussed in Section 7. Conclusions are drawn in the last section.

2. Literature overview and state of the art

The CNN concept was formalised in [23] for two-dimensional image recognition. Now, nearly 30 years later, there are numerous types of CNN architecture [24] used for image recognition and video processing. Although most CNN types are within two-dimensional CNN models, one can distinguish one-dimensional and three-dimensional CNN types [25].

Dynamic handwriting analysis benefits from the use of digital tablets and electronic pens [26]. Using these devices, it is possible to directly measure the temporal and spatial variables of handwriting, the pressure applied to the writing surface, the inclination of the pen, and the movement of the pen when it is not in contact with the surface (i.e. in air) [14]. Currently, the most popular approach to investigate the potential of automated handwriting analysis for the diagnosis of PD involves the use of dynamic information in the handwriting process to produce a more discriminating feature set of different dimensional data.

The applicability of kinematic, geometric, and non-linear dynamic characteristics was explored in a model of handwriting impairment in PD patients [6], and dynamic pressure features were encoded as colour. Furthermore, aerial movement during handwriting has a significant impact on the precision of disease classification [14]. The spectrum was used as the input of one-dimensional CNN, and the discrimination ability of different directions in the process of drawing motion was analysed, and the best diagnostic performance was obtained in X and Y directions [13]. In addition, a hybrid model [21] combining one-dimensional convolution network and Bidirectional Gated Recurrent Units (BiGRUs) was applied to the original features and derived features to capture the unique handwriting patterns that reflect Parkinson's disease. On the other hand, since dynamic analysis needs to consider not only the underlying generative process but also the geometry of handwritten patterns, graph analysis in a two-dimensional space is usually preferred to signal analysis in a one-dimensional space. Encouraged results have recently been reported to quantitatively assess the visual properties of handwritten motion samples from patients with PD using raw, filtered median and edge images [5]. However, according to others [10] a better understanding can be obtained using "dynamic augmentation" of static handwriting. Instead of simply using images of handwritten patterns, less realistic but more discriminating images are obtained by including additional dynamic information in the generation process. Furthermore, the application of three-dimensional CNN is limited in the area of analysis of drawing tests, as far as the author knows, this is the first attempt to use a three-dimensional CNN for Parkinson's diagnosis based on digitised handwriting tests. This leads to the exploration and comparison of the performance of one-, two- and three-dimensional CNN models in spiral drawing test classification.

One of the most serious problems to solve before applying deep learning techniques in the analysis of drawing and writing tests is the small size of the data sets available for training and validation. Due to the differences between testing protocols used in different medical centres and strict data handling requirements, acquiring sufficiently large data sets for training is not a viable solution. On the side of deep learning, there are two techniques that are used to overcome this problem. The first technique is data augmentation [27]. The augmentation procedure is based on the application of affine transformations, local nonlinear distortions, colour alternation, and noising to each image of the data set many times, whereas each alternated clone inherits the label of the original image. The set of transformations and their magnitude is chosen at random. This method was used in [7,8] and [28]. The latest has provided an analysis of different transformation types and their influence on modelling quality. Alternatively to this unsupervised technique, applications of generative adversary networks (GANs) [29,30] may be used. Table 1 summarises the main characteristics of previous works on PD diagnosis based on digital drawings: reference, dataset, feature set, method, performance, and year.

3. Problem statement

Main motivation of the present research is that, while 2D CNN is the dominant type when talking about image analysis, 3D models may allow for the encode of more kinematic and pressure parameters of the motion. Of course, one may suggest increasing the dimensionality even higher and using all the available parameters, but higher-dimensional convolutions are difficult to interpret. Therefore, CNNs with dimensions larger than three are left outside of the current research framework. On the contrary, one-dimensional CNNs have been used successfully before

Table 1
Overview of the related works.

Author(s)	Dataset	Features	Models	Accuracy	Year
Drotár et al. [14]	PaHaW	kinematic features	SVM	85.61	2014
Drotár et al. [15]	PaHaW	kinematic and spatio-temporal features	SVM	88.13	2015
Drotár et al. [3]	PaHaW	kinematic and pressure	SVM	87.4	2016
Pereira et al. [16]	HandPD	kinematic features	SVM	65.88	2016
Pereira et al. [17]	HandPD	time series based features	2D CNN	96.35	2018
Gil-Martín et al. [13]	[18]	kinematic features	1D CNN	96.5	2019
Diaz et al. [19]	PaHaW	dynamically enhaced static	2D CNN + SVM	75	2019
Naseer et al. [20]	PaHaW	fine-tuned-ImageNet features	AlexNet	98.28	2020
Diaz et al. [21]	PaHaW	raw and derived features	1D CNN-BiGRU	93.75	2021
Gazda et al. [22]	PaHaW	fine-tuned-ImageNet features	2D CNN	85.7	2022
	HandPD			92.7	

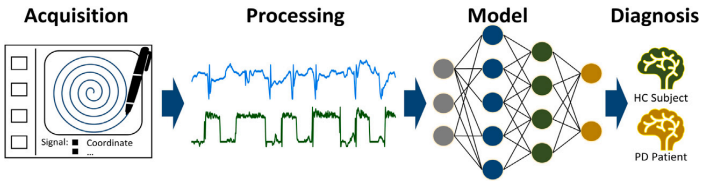


Fig. 1. The workflow for diagnosis of Parkinson's disease.

and are easy to interpret. This leads to the formal problem statement of the present investigation. Compare the performance of one-, two-, and three-dimensional CNN models for spiral drawing tests classification. This requires one to answer the following research questions.

1. Choose data set enhancement technique to encode different kinematic and drawing parameters into the drawing.
2. Choose the data set augmentation technique such that it acts in a similar way for one-, two-, and three-dimensional cases.
3. Choose the feature set(s) to encode.
4. Choose the CNN models structures which are the most similar among the one-, two-, and three-dimensional cases.

4. Materials

In this work, two data sets have been considered. The first data set, called *DraWritePD* [12], was acquired by the authors. The second data set, known as *PaHaW*, was kindly provided by the authors of [14,31]. Both datasets use similar digital signal acquisition equipment, and the handwriting data contains the same dynamic features (time sequences). In addition, they all contain a similar number of samples from each class, making the experiments more balanced.

4.1. DraWritePD

The "Drawing and handwriting tests for Parkinson's diagnostics" (*DraWritePD*) collects handwriting data from 25 patients with PD and 34 healthy control (HC) subjects of the same age and sex. For the group of patients with PD, the mean age was 74.1 ± 6.7 years. For the group of subjects with HC, the mean age was 74.1 ± 9.1 years. To acquire handwriting signals, special applications were designed for the Apple IPad Pro (9.6 inch, 2016 year) with the first generation of the Apple Pen. The application displays test instructions and the reference drawing on the IPad screen and records dynamic information from the Apple Pen tip accompanied by the time stamp. PD patients and their HC counterparts were asked to complete a series of handwriting tests consisting of 12 different drawing and writing tasks. When the task was completed, this dynamic information was stored in the file for later processing. It may be seen as a matrix with rows corresponding to the timestamps and columns corresponding to independent dynamic features, including: x coordinate (mm); y coordinate (mm); timestamp (sec); pressure (arbitrary unit of force applied on the surface: [0, ···, 6.0]); altitude (rad); azimuth (rad). In the present research, only digital versions of the Archimedes spiral drawing test (ASD) were considered.

4.2. PaHaW

The "Parkinson's disease handwriting database" (*PaHaW*) collects handwriting data from 37 patients with PD and 38 HC subjects [14,31]. No significant differences were found between the groups with respect to age or sex. The database was acquired in cooperation with the Movement Disorders Centre of the First Department of Neurology, Masaryk University, and St. Anne's University Hospital in Brno, Czech Republic. Each subject was asked to complete multiple handwriting tasks according to the prepared filled template at a comfortable speed. A tablet was overlaid with an empty paper template (containing only printed lines and a square box specifying the area for the Archimedean spiral), and a conventional ink pen was held in a normal fashion, allowing for immediate full visual feedback. Handwriting signals were recorded using an Intuos 4M (Wacom technology) digitising tablet at a sampling frequency of 150 Hz during pressure on the writing surface and movement over the writing surface. We denote these signals by on-surface movement and on-air movement, respectively. The recordings started when the pen touched the surface of the digitiser and finished when the task was completed. The tablet captured the following independent dynamic features: x coordinate; y coordinate; timestamp; button status; pressure; altitude; and azimuth. The button status was a binary variable, being 0 for pen-up state (in-air movement) and 1 for the pen-down state (on-surface movement). Although the task set presented in the *PaHaW* dataset is quite different from that used in the *DraWritePD* dataset, ASD was present in both datasets and was therefore used in this work.

5. Research workflow

In this section, we continue the idea of embedding dynamic handwriting features into static handwriting images. Specifically, more kinematic and pressure features are gradually encoded to generate higher-dimensional data representations, so that the corresponding one-, two-, and three-dimensional convolutional neural networks are used to analyse handwriting tasks to support the diagnosis of PD. Fig. 1 shows a general overview of the proposed automatic PD diagnosis system. Details of each stage are presented in the following subsections.

Table 2
Dynamic eigenvalues of three adjacent data points in the sample input.

t	*x*	*y*	*p*	*a*	*l*
533033966.322112	446.2969	−431.0742	0.723517	0.518733	1.059078
533033966.352263	449.875	−439.7695	0.739844	0.490138	1.059078
533033966.374942	454.7188	−448.125	0.800081	0.444148	1.059078

Note: The abbreviations x, y denote the x- and y- coordinate features; and a, l and p are the azimuth, altitude and pressure features, respectively; timestamp is represented by t.

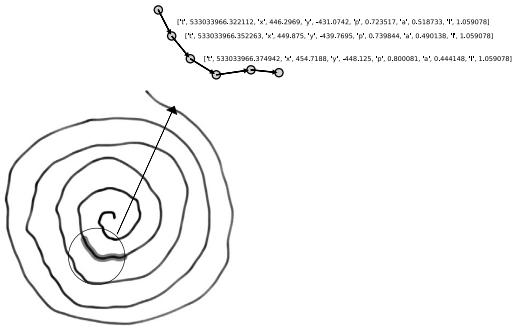


Fig. 2. Schematic diagram of the sample input, where each data point collects six independent features, and the arrow direction indicates the drawing direction. The abbreviations x, y denote the x- and y- coordinate features; and a, l and p are the azimuth, altitude and pressure features, respectively; timestamp is represented by t.

5.1. Data processing

Data processing consists of three main steps: data preparation, data enhancement, and data augmentation. Note that the following are introduced through the 1D, 2D, and 3D cases, respectively. In addition, the sample input is illustrated in Fig. 2. First, since the raw dataset inevitably contains some features that are not suitable for direct use, such as the units of the features being different, as illustrated in Table 2, the raw dynamic features are preprocessed with maximum and minimum normalisation before data enhancement to convert them into the same unit. Subsequently, the data enhancement gradually encodes dynamic features to generate enhanced data of different dimensions. Specifically, for the 1D case, the data encoding method is to directly regard the raw dynamic features (such as the x coordinate and the y coordinate) in the handwriting signal as 1D time series data. However, it should be noted that the encoding method of the timestamp feature is to replace the timestamp feature itself with the velocity feature calculated by combining the timestamp feature and the coordinate features. For the 2D RGB image encoding method, the coordinate features (the x coordinate and y coordinate) used in the 1D case are used as the pixel position information corresponding to each data point. Moreover, the azimuth, altitude, and pressure features of each data point are used as the red (R), green (G), and blue (B) colour information of the corresponding pixel. The velocity feature is encoded as line width information. For the 3D case, the only difference from the 2D case is the location information for each data point. It not only utilises the coordinate features (the x coordinate and y coordinate), but also combines the time feature (timestamp) to calculate the velocity feature of each data point, thus adopting (x coordinate, y coordinate, velocity) as the 3D position information of each data point. Afterwards, the generated raw point cloud data is voxelized into a matrix form acceptable to the CNN model with a fixed grid resolution (for convenience, hereinafter referred to as point cloud). It is worth noting that CNN has a powerful feature extraction ability [25], so the raw dynamic features were used directly for the enhancement of the data and no additional hand-crafted features were designed except for the velocity feature. Fig. 3 shows the results of the data enhancement in different dimensions. It is worth noting that velocity-based features can lead to better model performance [11,12], so the velocity is also replaced by acceleration or jerk respectively in the next experiments. Furthermore, a major challenge for the diagnosis of PD is the lack of suitable data. The direct application of CNN cannot effectively process raw handwriting signals collected from patients. One current approach to address this challenge is to augment data through data augmentation techniques or by combining multiple datasets [28], or employ pre-trained transfer learning strategies [20]. In the present work, data augmentation techniques are employed to significantly increase the diversity of PD handwriting samples, which can be roughly classified into the following categories; original, flipping, rotation, illumination, and jitter.

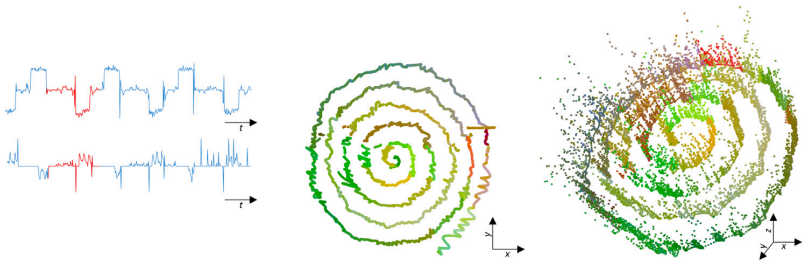
- Flipping: Flipping produces a mirror data, where the RGB image is flipped horizontally and vertically, and the point cloud is flipped along the x-axis and y-axis, respectively.
- Rotation: The data are rotated by a given angle, such as 90, 180, or 270 degrees, where the RGB image is rotated around the centre point, and the point cloud is rotated around the z-axis.
- Illumination: Illumination can be implemented by adjusting the colour map (RGB values), where random values are added to the R, G, and B channels of RGB images and point clouds.
- Jitter: Jitter can be implemented by adjusting the values of the coordinate features in the handwriting signal, where random values are added to the values of the coordinate features, and the resulting signal is enhanced.

Note that the first three data augmentation techniques are not suitable for the 1D case. Furthermore, relatively large data can negatively affect training time, while concise features can lead to under-fitting. Based on the available data, we first propose that the data sizes of 1D, 2D, and 3D are resized to 128, 128^2, and 128^3, respectively.

5.2. Neural network

The convolutional neural networks are bioinspired variants of multi-layer perceptrons (MLP) that can perform a variety of machine learning tasks without requiring the user to design and provide any hand-crafted features [25]. Recently, due to the development of new CNN variants, they have shown promising performance in traditionally challenging tasks with breakthrough progress [32,33]. The paradigm-shifting results provided by CNN are done in part with the help of extremely large training datasets. However, as mentioned earlier, one of the biggest limitations in the medical community is the inability to access larger, labelled, high-quality data that are sensitive, confidential, and difficult to collect. Due to the insufficient amount of data, in this work, in addition to using data augmentation techniques to augment data, we also incorporate a simplified version of the AlexNet [34] architecture, which consists of two main parts (convolutional layers for feature extraction and fully connected layers for classification), similar to [20]. Furthermore, it is worth pointing out that for fair comparisons, the one-, two-, and three-dimensional convolutional neural networks use the same network architecture, and the only difference is the size of the convolution kernel and the convolution method.

The simplified AlexNet architecture consists of four convolutional layers, a maximum pooling layer, a dropout layer, and three fully connected layers. The output of the fully connected layer is passed to



(a) 1D - Time Series (b) 2D - RGB Image (c) 3D - Point Cloud

Fig. 3. Enhanced data in different dimensional cases. In the one-dimensional (1D) case, the handwriting signal is enhanced into a time series, in which the raw dynamic features (such as x-coordinate and y-coordinate) are directly used; in the two-dimensional (2D) case, the handwriting signal is enhanced into an RGB image, in which the coordinate features (x-coordinate, y-coordinate) are used as (x,y) position information, and the (azimuth, altitude, pressure) features are used as (R,G,B) colour information, and the velocity feature is used as line width information; in the three-dimensional (3D) case, the handwriting signal is enhanced into a point cloud, in which the features (x-coordinate, y-coordinate, velocity) are used as (x,y,z) position information, and the (azimuth, altitude, pressure) features are used as (R,G,B) colour information.

Table 3
Architectural differences between one-, two-, and three-dimensional convolutional neural networks.

Layers	Filter	S	1D CNN		2D CNN		3D CNN	
			Input	K	Input	K	Input	K
Conv+ReLU	48	2	(6, 128)	5	(3, 128, 128)	(5, 5)	(3, 128, 128, 128)	(5, 5, 5)
MaxPooling	–	–	(48, 64)	–	(48, 64, 64)	–	(48, 64, 64, 64)	–
Conv+ReLU	128	2	(48, 32)	5	(48, 32, 32)	(5, 5)	(48, 32, 32, 32)	(5, 5, 5)
MaxPooling	–	–	(128, 16)	–	(128, 16, 16)	–	(128, 16, 16, 16)	–
Conv+ReLU	192	1	(128, 8)	3	(128, 8, 8)	(3, 3)	(128, 8, 8, 8)	(3, 3, 3)
Conv+ReLU	192	1	(192, 8)	3	(192, 8, 8)	(3, 3)	(192, 8, 8, 8)	(3, 3, 3)
MaxPooling	–	–	(192, 8)	–	(192, 8, 8)	–	(192, 8, 8, 8)	–
Flatten	–	–	(192, 4)	–	(192, 4, 4)	–	(192, 4, 4, 4)	–
FC+ReLU	–	–	768	–	3072	–	12288	–
Dropout	–	–	192	–	192	–	192	–
FC+ReLU	–	–	192	–	192	–	192	–
Dropout	–	–	128	–	128	–	128	–
FC+Softmax	–	–	128	–	128	–	128	–
Params			0.39MB		1.33MB		4.83MB	

Note: The abbreviations Conv, ReLU, and FC denote the convolutional layer, Rectified Linear Unit, and fully connected layer; and K and S are the kernel size and stride size; Params = parameters .

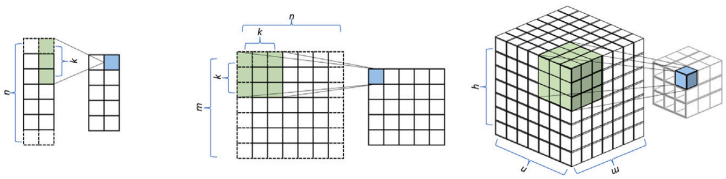
a softmax layer to produce a distribution on the labels of the 2 class. The first convolutional layer uses 48 kernels of size 5 with a stride of 2. The second convolutional layer takes as input the output of the first layer (ReLU activation and Max pooling) and filters it using 128 kernels of size 5. The third layer has 192 kernels of size 3 connected to the activated and pooled outputs of the second layer, while the fourth layer contains 192 kernels of size 3. The dropout layer in the fully connected layer temporarily removes nodes from the network with probability 50% during network training. The specific convolution kernel and the convolution operation are shown in Fig. 4. Details of the simplified AlexNet architecture deployed in our experiments are shown in Table 3 and Fig. 5.

5.3. Implementation

Model training and testing were carried out on a PC that has an Intel(R) Core(TM) i7 – 11700K CPU with 3.60 GHZ(8 CPU), 32GB RAM and an NVIDIA GEFORCE RTX3070Ti graphics card with 8 GB memory. It is worth noting that data partition scheme should be nested in five-fold cross-validation. The initial learning rate was set to $1e-4$, and Adam [35] optimiser was used to train the model. The cross-entropy loss function was used to optimise the model parameters and L2 regularisation is used to avoid overfitting. Various metrics were used to measure the performance of the model in more detail, including accuracy, precision, sensitivity, specificity, and F_1-score (see [36]).

6. Experimental results

In this section, we report a series of experimental results aimed at comparing the classification performance of the one-, two-, and three-dimensional convolutional neural network in the diagnosis of PD. Tables 4 and Table 5 show the dynamic features used in the data enhancement process and the corresponding model diagnostic results obtained. To obtain robust experimental conclusions, both the *DraWritePD* data set [37] and the *PaHaW* data set [3,14] were considered. In addition, we analysed the previous literature with the *PaHaW* database in order to contextualise our results. The related state-of-the-art results obtained on the *PaHaW* data set are shown in Table 6. First, we evaluated the impact of embedding different sets of dynamic features in the same-dimensional space on the diagnostic performance of the model. Specifically, in the baseline experiment of each dimension space, only the coordinate feature or its derived velocity feature is used to encode the position information, and then, on this basis, other dynamic features gradually encode the colour information and line width information in the enhanced data. The experimental results in Tables 4 and 5 show that, in the same dimensional space, overall diagnostic performance predictably presents the same upward trend, and the continuous improvement of performance confirms that encoding more dynamic features in the enhancement of the data helps to distinguish PD patients from HC subjects. In particular, the encoding of colour information and line width information makes the enhanced data more discriminating. There is, however, one exception. In the one-dimensional space, compared with the baseline experiment, the addition of velocity features failed to provide reasonable



(a) 1D - Conv1d Kernel (b) 2D - Conv2d Kernel (c) 3D - Conv3d Kernel

Fig. 4. The schematic diagram of convolution operation in different dimensions. The green area represents the convolution area, and the blue area represents the convolution result, where m, n, and h are the feature map sizes, and k represents the convolution kernel size.

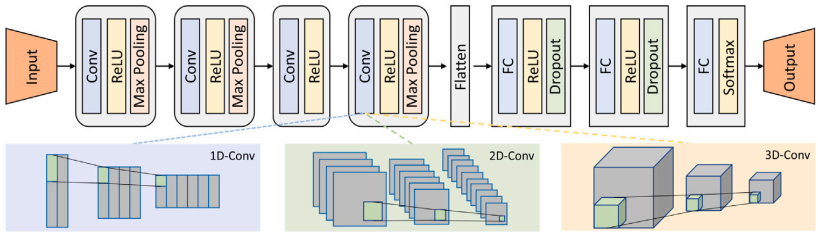


Fig. 5. The convolutional neural network model, in which the one-, two-, and three-dimensional convolutional neural networks use the same model architecture, only the convolution method is different.

Table 4
Performance comparison in the *DraWritePD* dataset.

Dimension	Dynamic features								Metrics (in%)				
	x	y	a	l	p	v	c	j	Precision	Sensitivity	Specificity	Accuracy	F_1 score
1D	✔	✔							50.50	53.32	50.25	51.67	52.03
	✔	✔				✔			51.67	61.75	52.67	56.93	54.51
	✔	✔	✔	✔	✔				52.25	63.32	51.67	57.73	56.67
	✔	✔	✔	✔	✔	✔			59.72	62.50	56.25	59.38	61.03
	✔	✔	✔	✔	✔		✔		63.72	67.25	59.67	62.56	65.21
	✔	✔	✔	✔	✔			✔	60.67	65.32	57.75	58.73	63.45
2D	✔	✔							53.25	56.32	68.67	62.56	58.61
	✔	✔				✔			66.67	62.75	75.25	69.38	67.14
	✔	✔	✔	✔	✔				68.72	75.00	78.67	73.67	72.67
	✔	✔	✔	✔	✔	✔			75.00	76.50	80.00	77.73	76.51
	✔	✔	✔	✔	✔		✔		77.50	78.25	81.75	80.38	79.32
	✔	✔	✔	✔	✔			✔	76.67	74.75	78.67	75.93	77.14
3D	✔	✔				✔			72.25	78.00	80.25	76.58	75.21
	✔	✔	✔	✔	✔	✔			77.50	**86.50**	81.75	82.34	81.45
	✔	✔	✔	✔	✔		✔		**82.50**	82.50	**87.25**	**85.38**	**85.51**
	✔	✔	✔	✔	✔			✔	77.25	86.25	80.00	83.34	81.95

Note: The abbreviations x, y denote the x- and y- coordinate features; and a, l and p are the azimuth, altitude and pressure features, respectively; velocity, acceleration, and jerk are represented by v, c, and j, respectively.

predictions, possibly because the velocity features were derived from coordinate features, resulting in redundant discriminating information. Furthermore, we compared the diagnostic performance of convolutional neural networks in different dimensions. It is worth noting that for a fair comparison, all convolutional neural network models used the same network architecture. First, the experimental results in Tables 4 and 5 confirm that, with the same dynamic feature encoding, on average, convolutional neural networks achieve increasingly better diagnostic performance with increasing dimensionality, with the most competitive diagnostic results obtained by 3D convolutional neural networks. For example, in terms of diagnostic accuracy, the 1D, 2D and 3D convolutional networks achieved sequentially increasing diagnostic performance of 63.33%, 81.33% and 84.67%, respectively, in the *PaHaW* data set. In addition, interestingly, the diagnostic performance of 3D convolutional neural networks is almost comparable to that of low-dimensional convolutional neural networks even if only location information is given. For example, in the DaWritePD data set, the 3D convolutional neural network can achieve 75.21% diagnostic accuracy only in the baseline experiment; on the contrary, the optimal diagnostic accuracy of 1D and 2D convolutional neural network 1 D and 2 D is just 65.21% and 79.32% respectively.

7. Discussion

Comparing Tables 4 5 that summarise model goodness metrics (computed on the basis of testing data) for the different feature sets and model dimensionalities one can observe that encoding more features into the image usually causes model goodness to increase and increasing dimensionality of the convolutional kernel also leads to better models. In the case of the *DraWritePD* data set, exceptions occur with specificity and precision. In the case of the *PaHaW* data set, exceptions also occur in the sensitivity of the models. Such anomalies in the behaviour of goodness metrics may be caused by the presence and absence of the feature describing the velocity that is related to the amount of tremor in the motions. Currently, there are many new features (spiral specific features) or improvements to existing features [12] that can

Table 5
Performance comparison in the *PaHaW* dataset.

Dimension	Dynamic features								Metrics (in%)				
	x	y	a	l	p	v	c	j	Precision	Sensitivity	Specificity	Accuracy	F_1 score
1D	✔	✔							50.00	57.14	50.00	53.33	53.33
	✔	✔				✔			53.93	57.14	61.75	56.67	57.33
	✔	✔	✔	✔	✔				58.14	58.48	62.50	60.67	59.14
	✔	✔	✔	✔	✔	✔			59.03	71.43	56.25	63.33	64.58
	✔	✔	✔	✔	✔		✔		62.31	75.71	62.50	64.22	65.29
	✔	✔	✔	✔	✔			✔	57.93	68.73	56.25	60.67	63.92
2D	✔	✔							56.93	75.71	57.25	64.33	63.16
	✔	✔				✔			72.31	81.48	75.00	75.33	73.43
	✔	✔	✔	✔	✔				82.33	71.43	82.50	80.00	78.92
	✔	✔	✔	✔	✔	✔			76.25	85.71	75.50	81.33	80.51
	✔	✔	✔	✔	✔		✔		81.03	84.73	78.75	83.67	82.29
	✔	✔	✔	✔	✔			✔	79.61	83.67	76.25	80.33	79.97
3D	✔	✔				✔			62.31	82.73	61.25	68.67	73.29
	✔	✔	✔	✔	✔	✔			75.93	**90.48**	75.00	82.22	82.50
	✔	✔	✔	✔	✔		✔		**83.61**	87.31	**85.50**	**84.67**	**85.71**
	✔	✔	✔	✔	✔			✔	82.71	85.71	80.25	81.73	81.50

Note: The abbreviations x, y denote the x- and y- coordinate features; and a, l and p are the azimuth, altitude and pressure features, respectively; velocity, acceleration, and jerk are represented by v, c, and j, respectively.

Table 6
Comparisons with state-of-the-art works.

Author(s)	Dataset	Features	Models	Accuracy (in %)
Drotár et al. [31]	PaHaW	hand-crafted	SVM	62.80
Diaz et al. [21]	PaHaW	1D CNN-extracted	1D CNN + BiGRU	93.75
Diaz et al. [19]	PaHaW	2D CNN-extracted	2D CNN + SVM	75.00
Present work	PaHaW	1D CNN-extracted	1D CNN	64.22
		2D CNN-extracted	2D CNN	83.67
		3D CNN-extracted	3D CNN	84.67

further improve the classification performance, similar to the diagnostic accuracy in 90% obtained using multiple raw and derived features in [21]. Although such an investigation is beyond the scope of this paper, it may constitute the direction of future studies. Another point to discuss is that while the differences in performance metrics are similar between one- and two-dimensional CNNs, the difference between two- and three-dimensional CNNs is greater for the *DraWritePD* dataset. This may be triggered by the fact that in the case of *PaHaW* no reference drawing is provided, but in the case of *DraWritePD* a reference drawing is presented, making it easier to complete the test. The direction of the drawing and the age groups of the subjects tested may also contribute to this difference.

8. Conclusions

The application of one-, two-, and three-dimensional deep convolutional neural networks for the analysis of spiral drawing tests to support the diagnosis of Parkinson's disease has been investigated. Through comparative experiments on the two datasets, our hypothesis is confirmed that data representation and classification models in high-dimensional space are more beneficial to distinguish PD patients from HC subjects. Additionally, although we adapted the raw dynamic feature set for feature encoding to obtain high diagnostic accuracy, we believe that there is still room for improvement. Future research will be directed towards determining the specific feature sets to be encoded.

Ethics statement

The data acquisition process was carried out with strict guidance from the privacy law. The study was approved by the Research Ethics Committee of the University of Tartu (No. 1275T – 9).

CRediT authorship contribution statement

Xuechao Wang: Workflow design, Computational experiments, Manuscript writing. **Junqing Huang:** Curating of the computational experiments, Manuscript curating. **Marianna Chatzakou:** Curating of the computational experiments, Manuscript curating. **Sven Nõmm:** Problem conceptualisation, Workflow curating, Manuscript editing. **Elli Valla:** Conceptualisation of the 3D data representation. **Kadri Medijainen:** Performed data acquisition. **Pille Taba:** Medical protocols curating. **Aaro Toomela:** Medical protocols curating. **Michael Ruzhansky:** Manuscript, Workflow, Computations curating, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

Acknowledgements

The authors thank P. Drotar and his team for sharing the PaHaW dataset. This work was partially supported by the FWO Odysseus 1 grant G.0H94.18N: Analysis and Partial Differential Equations and the Methusalem Programme of the Ghent University Special Research Fund (BOF) (Grant number 01M01021). Michael Ruzhansky is also supported by the EPSRC, United Kingdom grant EP/R003025/2. Marianna Chatzakou is a postdoctoral fellow of the Research Foundation – Flanders (FWO) under the postdoctoral grant No 12B1223N. The work of Sven Nõmm and Elli Valla in the project "ICT programme" was supported by the European Union through the European Social Fund. Pille Taba's work was supported by the PRG grant 957 from the Estonian Research Council.

References

[1] J. Phillips, G.E. Stelmach, N. Teasdale, What can indices of handwriting quality tell us about parkinsonian handwriting? Hum. Mov. Sci. 10 (2–3) (1991) 301–314.

[2] C. Marquardt, N. Mai, A computational procedure for movement analysis in handwriting, J. Neurosci. Methods 52 (1) (1994) 39–45, http://dx.doi.org/10.1016/0165-0270(94)90053-1.

[3] P. Drotar, J. Mekyska, I. Rektorova, L. Masarova, Z. k Smékal, M. Faundez-Zanuy, Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease, Artif. Intell. Med. 67 (2016) 39–46, http://dx.doi.org/10.1016/j.artmed.2016.01.004.

[4] E. Valla, S. Nõmm, K. Medijainen, P. Taba, A. Toomela, Tremor-related feature engineering for machine learning based Parkinson's disease diagnostics, Biomed. Signal Process. Control 75 (2022) 103551, http://dx.doi.org/10.1016/j.bspc.2022.103551.

[5] M. Moetesum, I. Siddiqi, N. Vincent, F. Cloppet, Assessing visual attributes of handwriting for prediction of neurological disorders—A case study on Parkinson's disease, Pattern Recognit. Lett. 121 (2019) 19–27.

[6] C.D. Rios-Urrego, J.C. Vásquez-Correa, J.F. Vargas-Bonilla, E. Nöth, F. Lopera, J.R. Orozco-Arroyave, Analysis and evaluation of handwriting in patients with Parkinson's disease using kinematic, geometrical, and non-linear features, Comput. Methods Programs Biomed. 173 (2019) 43–52.

[7] S. Nõmm, S. Zarembo, K. Medijainen, P. Taba, A. Toomela, Deep CNN based classification of the archimedes spiral drawing tests to support diagnostics of the Parkinson's disease, IFAC-PapersOnLine 53 (5) (2020) 260–264, http://dx.doi.org/10.1016/j.ifacol.2021.04.185, 3rd IFAC Workshop on Cyber-Physical & Human Systems CPHS 2020. URL https://www.sciencedirect.com/science/article/pii/S2405896321003360.

[8] S. Zarembo, S. Nõmm, K. Medijainen, P. Taba, A. Toomela, CNN based analysis of the Luria's alternating series test for Parkinson's disease diagnostics, in: Asian Conference on Intelligent Information and Database Systems, Springer, 2021, pp. 3–13.

[9] Z. Galaz, P. Drotar, J. Mekyska, M. Gazda, J. Mucha, V. Zvončák, Z. Smekal, M. Faundez-Zanuy, R. Castrillon, J.R. Orozco-Arroyave, et al., Comparison of CNN-learned vs. Handcrafted features for detection of Parkinson's disease dysgraphia in a multilingual dataset, Front. Neuroinform. (2022) 35.

[10] J. Galbally, M. Diaz-Cabrera, M.A. Ferrer, M. Gomez-Barrero, A. Morales, J. Fierrez, On-line signature recognition through the combination of real dynamic data and synthetically generated static data, Pattern Recognit. 48 (9) (2015) 2921–2934.

[11] S. Nõmm, K. Bardõš, A. Toomela, K. Medijainen, P. Taba, Detailed analysis of the luria's alternating seriestests for parkinson's disease diagnostics, in: 2018 17th IEEE International Conference on Machine Learning and Applications, ICMLA, IEEE, 2018, pp. 1347–1352.

[12] E. Valla, S. Nomm, K. Medijainen, P. Taba, A. Toomela, Tremor-related feature engineering for machine learning based Parkinson's disease diagnostics, Biomed. Signal Process. Control 75 (2022) 103551.

[13] M. Gil-Martín, J.M. Montero, R. San-Segundo, Parkinson's disease detection from drawing movements using convolutional neural networks, Electronics 8 (8) (2019) 907.

[14] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, M. Faundez-Zanuy, Analysis of in-air movement in handwriting: A novel marker for Parkinson's disease, Comput. Methods Programs Biomed. 117 (3) (2014) 405–411.

[15] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, M. Faundez-Zanuy, Decision support framework for Parkinson's disease based on novel handwriting markers, IEEE Trans. Neural Syst. Rehabil. Eng. 23 (3) (2014) 508–516.

[16] C.R. Pereira, D.R. Pereira, F.A. Silva, J.P. Masieiro, S.A. Weber, C. Hook, J.P. Papa, A new computer vision-based approach to aid the diagnosis of Parkinson's disease, Comput. Methods Programs Biomed. 136 (2016) 79–88.

[17] C.R. Pereira, D.R. Pereira, G.H. Rosa, V.H. Albuquerque, S.A. Weber, C. Hook, J.P. Papa, Handwritten dynamics assessment through convolutional neural networks: An application to Parkinson's disease identification, Artif. Intell. Med. 87 (2018) 67–77.

[18] M. Isenkul, B. Sakar, O. Kursun, et al., Improved spiral test using digitized graphics tablet for monitoring Parkinson's disease, in: The 2nd International Conference on E-Health and Telemedicine, Vol. 5, ICEHTM-2014, 2014, pp. 171–175.

[19] M. Diaz, M.A. Ferrer, D. Impedovo, G. Pirlo, G. Vessio, Dynamically enhanced static handwriting representation for parkinson's disease detection, Pattern Recognit. Lett. 128 (2019) 204–210.

[20] A. Naseer, M. Rani, S. Naz, M.I. Razzak, M. Imran, G. Xu, Refining Parkinson's neurological disorder identification through deep transfer learning, Neural Comput. Appl. 32 (2020) 839–854.

[21] M. Diaz, M. Moetesum, I. Siddiqi, G. Vessio, Sequence-based dynamic handwriting analysis for Parkinson's disease detection with one-dimensional convolutions and BiGRUs, Expert Syst. Appl. 168 (2021) 114405.

[22] M. Gazda, M. Hireš, P. Drotár, Multiple-fine-tuned convolutional neural networks for parkinson's disease diagnosis from offline handwriting, IEEE Trans. Syst. Man Cybern. 52 (1) (2021) 78–89.

[23] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324, http://dx.doi.org/10.1109/5.726791.

[24] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, R. Socher, Deep learning-enabled medical computer vision, NPJ Digit. Med. 4 (1) (2021) 5.

[25] L. Alzubaidi, J. Zhang, A.J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M.A. Fadhel, M. Al-Amidie, L. Farhan, Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions, J. Big Data 8 (2021) 1–74.

[26] C.R. Pereira, D.R. Pereira, S.A. Weber, C. Hook, V.H.C. De Albuquerque, J.P. Papa, A survey on computer-assisted Parkinson's disease diagnosis, Artif. Intell. Med. 95 (2019) 48–63.

[27] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (1) (2019) 1–48.

[28] I. Kamran, S. Naz, I. Razzak, M. Imran, Handwriting dynamics assessment using deep neural network for early identification of Parkinson's disease, Future Gener. Comput. Syst. 117 (2021) 234–244.

[29] E. Dzotsenidze, E. Valla, S. Nõmm, K. Medijainen, P. Taba, A. Toomela, Generative adversarial networks as a data augmentation tool for CNN-based Parkinson's disease diagnostics, IFAC-PapersOnLine 55 (29) (2022) 108–113, http://dx.doi.org/10.1016/j.ifacol.2022.10.240, 15th IFAC Symposium on Analysis, Design and Evaluation of Human Machine Systems HMS 2022. URL https://www.sciencedirect.com/science/article/pii/S2405896322022674.

[30] Y. Chen, X.-H. Yang, Z. Wei, A.A. Heidari, N. Zheng, Z. Li, H. Chen, H. Hu, Q. Zhou, Q. Guan, Generative adversarial networks in medical image augmentation: a review, Comput. Biol. Med. (2022) 105382.

[31] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, M. Faundez-Zanuy, Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease, Artif. Intell. Med. 67 (2016) 39–46.

[32] Z. Wu, Y. Huang, L. Wang, X. Wang, T. Tan, A comprehensive study on cross-view gait based human identification with deep CNNs, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2) (2016) 209–226.

[33] S. Miao, S. Piat, P. Fischer, A. Tuysuzoglu, P. Mewes, T. Mansi, R. Liao, Dilated FCN for multi-agent 2D/3D medical image registration, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, No. 1, 2018.

[34] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.

[35] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[36] M. Grandini, E. Bagli, G. Visani, Metrics for multi-class classification: an overview, 2020, arXiv preprint arXiv:2008.05756.

[37] S. Nõmm, K. Bardõš, A. Toomela, K. Medijainen, P. Taba, Detailed analysis of the Luria's alternating SeriesTests for Parkinson's disease diagnostics, in: 2018 17th IEEE International Conference on Machine Learning and Applications, ICMLA, 2018, pp. 1347–1352, http://dx.doi.org/10.1109/ICMLA.2018.00219.

Appendix 7

VII

Elli Valla, Gert Kanter, Sven Nõmm, Anton Osvald Kuusk, Peeter Maran, Karl Mihkel Seenmaa, Killu Mägi, and Aaro Toomela. Enhancing cerebral palsy gait analysis with 3D computer vision: A dual-camera approach. In *2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 1352–1357, 2024

2024 10th International Conference on Control, Decision and Information Technologies
CoDIT 2024 | Valletta, Malta / July 01-04, 2024 Technically co-sponsored by IEEE & IFAC

Enhancing Cerebral Palsy Gait Analysis with 3D Computer Vision: A Dual-Camera Approach

Elli Valla
Gert Kanter
Sven Nõmm
Department of Software Science,
School of Information Technology
Tallinn University of Technology
Tallinn, Estonia
e-mail: elli.valla@taltech.ee

Anton Osvald Kuusk
Peeter Maran
Karl Mihkel Seenmaa
School of Information Technology
Tallinn University of Technology
Tallinn, Estonia

Killu Mägi
Haapsalu Neurological
Rehabilitation Centre
Haapsalu, Estonia

Aaro Toomela
Shool of Natural Sciences
and Health
Tallinn University
Tallinn, Estonia
e-mail: aaro.toomela@tlu.ee

2024 10th International Conference on Control, Decision and Information Technologies (CoDIT) | 979-8-3503-7397-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/CoDIT62066.2024.10708137

***Abstract*—This paper introduces a framework for analyzing cerebral palsy (CP) gait using a markerless 3D computer vision system equipped with two RGB cameras. The system employs advanced pose estimation algorithms and machine learning techniques to analyze gait dynamics. Although limited by the model's simplicity at certain joints—particularly the pelvis, which is represented by just two points—the system integrates the four most clinically relevant out of 11 measured kinematic variables. The system excels in capturing large joint angles like knee flexion/extension but faces challenges with smaller angles such as hip abduction/adduction due to the limitations of single-point representation. Comprehensive gait metrics, including cadence and walking speed, are derived by projecting foot movements onto the floor plane, providing valuable insights into gait mechanics. This research simplifies the technology required for precise gait analysis and enhances its accessibility in clinical settings, offering significant advancements in the diagnosis and treatment of movement disorders.**

I. INTRODUCTION

Cerebral palsy (CP) represents the most common childhood motor disability, where precise gait analysis is key to rehabilitation [1]. Traditional marker-based motion capture systems, despite being the gold standard, come with limitations such as high costs, time-intensive setups, and discomfort for patients due to marker attachment [2] [3] [4]. Marker placement for gait analysis in CP patients, requiring 1 to 2.5 hours for setup and additional hours for analysis, demands precision and expertise from clinicians. This process, illustrated in Fig 1a, b, c, involves extensive measurements and can be uncomfortable for young patients, sometimes prolonging sessions. Additionally, the heat from infrared cameras used in analysis may increase discomfort, adding to the procedure's strain.

Acknowledging the challenges of traditional gait analysis, including clinician subjectivity and patient discomfort, this study introduces a less invasive markerless computer vision technique. By employing two standard RGB cameras, we've developed a system for 3D pose estimation, enhancing gait analysis accessibility for CP patients. This innovation minimizes the need for extensive human interaction from clinical staff, paving the way for future advancements in analysis tools.

While markerless gait analysis systems are not new [4] [5] [6] [7], the novelty of our work lies in its application and validation for CP gait analysis using only two regular cameras for 3D pose estimation. We have successfully demonstrated that kinematic parameters extracted from the keypoints can be used to reconstruct relevant metrics, such as sagittal and frontal angles of the lower limbs. This advancement represents a significant step forward in leveraging computer vision for medical applications, particularly in the context of CP, where detailed analysis of gait patterns can inform more targeted and effective rehabilitation strategies.

A. Related Work

Recent advancements in deep learning and computer vision have shown significant potential in analyzing human movement from video footage, offering substantial benefits for studying disease populations. Integrating these technologies with smartdevices could transform rehabilitation practices by enabling the assessment of real-world movements without the need for specialized hardware.

Standard 3D gait analysis uses multi-camera systems and skin markers, limited by cost, setup complexity, and patient discomfort [8] [9] [10]. Markerless technologies offer alternatives, though multi-camera setups are impractical for ambulatory settings [2] [11] [12] [13]. Single- and dual camera methods, utilizing RGB-depth cameras, provide a simpler solution for 2D kinematic analysis, essential for screening and treatment evaluation [14] [15]. Most open-source deep learning methods, such as AlphaPose and OpenPose [16] [17] [18] are trained on general movement data rather than specific gait patterns, and do not adhere to clinical gait analysis standards. Additionally, their training datasets often exclude individuals with gait impairments, compromising the optimization of these methods for clinical use and their validity in clinical settings.

A study by Moro *et al.* [4] explored the efficacy of markerless versus marker-based motion tracking systems, employing three cameras to achieve results on par with traditional methods that typically require eight cameras. This advancement, propelled by machine learning technologies, suggests the feasibility of using as few as three cameras for accurate tracking of joint and limb movements, challenging the long-standing need for more extensive camera setups for precise motion tracking.

In their study, Ma *et al.* [19] assessed the Microsoft Kinect camera's effectiveness for gait analysis in children with CP. It compares Kinect's lower limb joint kinematics against a traditional marker-based system. Initial results show modest to poor correlation, but calibration with linear regression and Long Short Term Memory (LSTM) algorithms significantly improves accuracy, especially in hip and knee sagittal kinematics. The study by Nieto *et al.* [7] demonstrated the use of a smartphone with cloud computing to accurately extract gait features, achieving 95% accuracy for side views and 80% for frontal views. Meanwhile, Maex *et al.* [5] successfully utilized a single video camera for lower body sagittal plane kinematic analysis, revealing high correlations up to 0.99. These results are promising; however, for comprehensive kinematic analysis across all anatomical planes, not limited to the sagittal plane, a setup with at least two cameras is necessary to capture 3D data.

The study of Haberfehler *et al.* [20] explores using machine learning and video analysis to assess dystonia in dyskinetic cerebral palsy patients, aiming to automate the scoring process currently reliant on clinical evaluation. By extracting 2D stick figures from videos and analyzing them with DeepLabCut [21] for pose estimation, the research achieved tracking accuracy and trained models to predict clinical scores, comparing favorably with human scoring. This proof-of-concept demonstrates the potential for machine learning to efficiently and accurately assess dystonia, offering a scalable and less subjective alternative to traditional methods.

II. PROBLEM STATEMENT

This study advances cerebral palsy (CP) gait analysis by integrating a dual-camera system with advanced pose estimation algorithms to construct precise 3D gait models. It focuses on the critical question:

How effectively can advanced computer vision techniques and automated keypoint extraction analyze and model the complex gait dynamics of cerebral palsy patients?

Our markerless dual-camera setup simplifies traditional methods, reducing setup times and discomfort for CP patients. We evaluate its ability to autonomously extract and analyze kinematic features, comparing it directly with the established Vicon system to assess precision and accessibility.

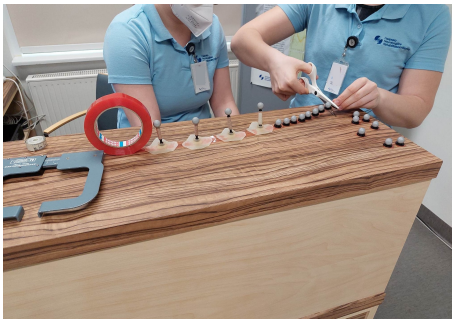
III. MATERIALS AND METHODOLOGY

This section outlines the study's methodical strategy in creating a markerless system for analyzing cerebral palsy gait.

A. Data acquisition and experimental setting

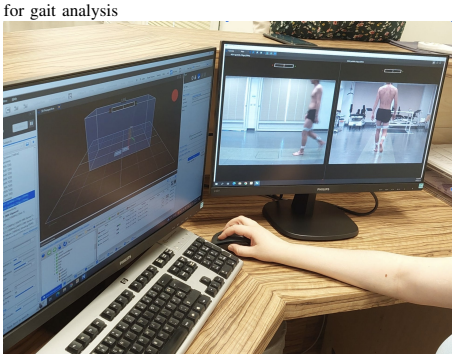
At the Haapsalu Neurological Rehabilitation Centre (HNRC) [22], the clinical motion and gait analysis laboratory plays a crucial role in assessing patients' specific symptoms. Equipped with advanced technology, including 8 MX T-Series infrared cameras, 2 Basler pilot piA640-210gc video cameras (positioned to record the patient from both the side and front), and 2 AMTI force plates, the laboratory utilizes Vicon's comprehensive software suite (Vicon Polygon, Vicon Nexus, among others) for data recording, processing, and reporting. Central to the analysis process are reflective markers attached to the patient, tracked meticulously by the infrared cameras to monitor movement.

This study employs two Basler pilot piA640-210gc RGB cameras to capture gait from side and front views at HNRC,



(a) Clinicians preparing reflective markers for gait analysis

(b) Utilizing a laser for enhanced precision in marker placement for gait analysis



(c) Comparative analysis of gait dynamics: Video footage and 3D model visualization

Fig. 1: Reflective marker setup (a, b), laser alignment for accuracy (b), and 3D model evaluation (c) in HNRC's gait lab.

aiming to achieve the comprehensive analysis typically performed with eight MX T-Series infrared cameras. The exploration seeks to evaluate the effectiveness of utilizing fewer cameras alongside force plates, compared to the conventional eight MX T-series camera setup.

B. Pose Estimation Frameworks

The following frameworks were examined and tested for potential use: MediaPipe [23], AlphaPose [24], OpenPose [25], Metrabs [26], TensorFlow Movenet [27], HRNet [28], and Detectron2 [29]. These frameworks offer capabilities such as real-time multimedia processing, pose estimation, multi-person pose estimation, object tracking, and computer vision tasks like image classification and instance segmentation.

1) The number of keypoints: Our primary focus is on keypoints, particularly their body coverage and accuracy. Most freely available models only provide the basic 17 keypoints, excluding critical data like the foot. MediaPipe, however, offers a more comprehensive solution with 32 keypoints covering the entire body, making it a standout choice for our needs (see Fig. 3a).

2) Stability test: In our stability evaluation, we converted a single image into a 768-frame video at 30 fps and 1280x720 resolution to simulate continuous movement for testing different pose estimation frameworks. Each framework processed this video to assess the consistency of keypoint tracking across frames. The standard deviation of keypoint placement, detailed in Table I, quantifies the precision of each framework, with 2 pixels roughly equivalent to 1 centimeter.

TABLE I: Standard deviation for different models

Model	x (pixels)	y (pixels)
MediaPipe	0.2459	0.4226
Detectron2	0.5519	0.4298
TensorFlow Movenet	2.1118	2.9386
HRNet	26.7547	28.8701

3) Conclusive analysis: Table II compares various pose estimation frameworks, highlighting their suitability for single versus multi-person detection, compatibility with Nvidia GPUs, and licensing terms. In our single-patient lab setting, multi-person detection was not necessary, and MediaPipe's capability for single-person detection aligned well with our requirements. MediaPipe stands out due to its accuracy, comprehensive model availability for free use, and near real-time processing speed, all without the necessity of Nvidia GPUs. This framework performs well even with lower-quality images and features a user-friendly API, making it the preferred choice for our research purposes.

C. Gait cycle prediction

This study employs machine learning techniques for gait cycle analysis, focusing on classifying gait phases [30] through body landmarks identified by pose estimation. Key to this process is selecting specific lower-body landmarks: 27-32 in Fig. 3a, and applying median filtering to reduce noise,

TABLE II: Pose estimation framework comparison

Framework	Single	Multiple	Nvidia GPU	License
MediaPipe	+	-	-	+
Hrnet	+	+	+	+
OpenPose	+	+	+	-
AlphaPose	+	+	-	-
Metrabs	+	+	-	-
Tensorflow	+	-	-	+
Detectron2	+	+	-	+

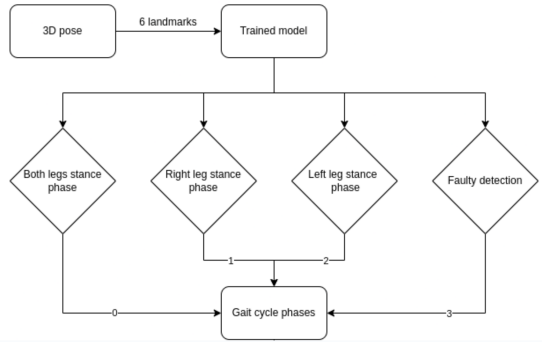


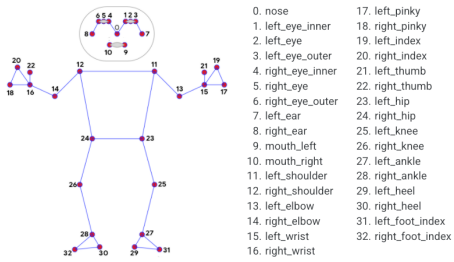
Fig. 2: Process for labelling frames for machine learning based gait cycle prediction.

thus ensuring precise phase identification. Movement is captured through the calculation of coordinate changes between frames, essential for distinguishing different gait phases. The methodology involves detailed frame-by-frame data annotation, categorizing each frame into one of four classes based on leg positions and movement accuracy, enhancing the quality and reliability of gait analysis. This systematic annotation, illustrated in Fig 2, uses a coding system to represent the gait cycle accurately, promoting consistency across studies.

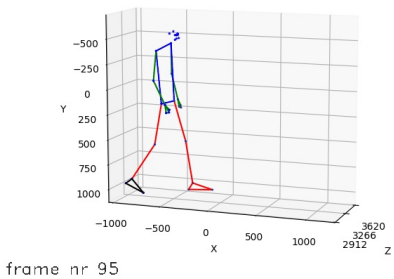
D. 3D model construction

Creating a 3D model of a subject's gait is a complex process that begins with capturing video from multiple angles. It involves extracting 2D poses from these videos using pose detection frameworks, and then calculating the 3D pose by triangulating these 2D poses. Central to this process is rigorous camera calibration, determining intrinsic parameters like focal lengths, skew factor, and principal points. This calibration is crucial for accurately mapping 2D images to a 3D space, ensuring the model reflects true dimensions and orientations. Calibration relies on images featuring a chessboard pattern, with around 30 images optimizing calibration accuracy and laying a solid foundation for subsequent 3D model synthesis.

Following calibration (see Fig. 4a), the cameras' relative positions and orientations are accurately determined using images of the calibration device captured simultaneously by both cameras. This process involves calculating translation and rotation matrices that describe the positioning of one camera relative to the other with about five images being optimal.

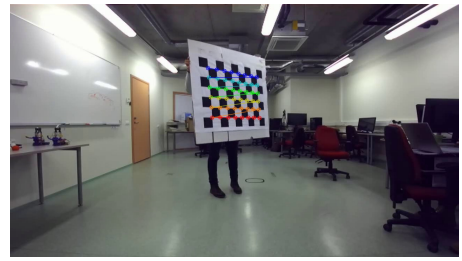


(a) 2D keypoints extracted from the pose estimation algorithm (MediaPipe)

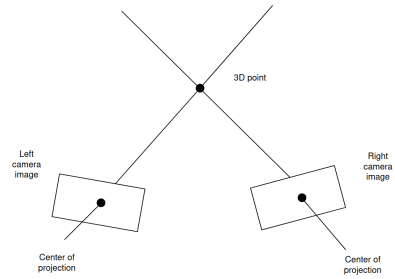


(b) Constructed 3D model

Fig. 3: 3D model illustration.



(a) Calibration using chessboard



(b) Triangulation scheme

Fig. 4: Steps in the process of the 3D model creation: camera calibration (a) and triangulation (b).

The triangulation (see Fig. 4b) phase is central for creating the 3D model, involving precise estimation of the 3D location of points in space based on their 2D projections from each camera. This step is critical for transforming 2D data into a comprehensive 3D model, requiring adjustments for lens distortions and addressing inherent imprecisions in projecting complex human movements into 2D images.

E. Kinematic variables and general gait parameters

This study evaluates motion across the sagittal, frontal, and transverse planes to calculate kinematic variables, offering a deep dive into the intricacies of human gait mechanics. Kinematic variables, or gait kinematics, are essential measurements that describe the motion of the body's segments throughout the gait cycle. These measurements are categorized based on three anatomical planes that correspond to the human body's movement directions: the sagittal plane, which divides the body into right and left portions; the frontal (or coronal) plane, dividing it into front and back portions; and the transverse plane, splitting it into upper and lower portions.

The study's analysis of kinematic variables within these planes includes:

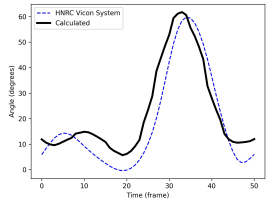
- **Sagittal plane**:
  - Pelvis anterior/posterior
  - Hip flexion/extension
  - Knee flexion/extension
  - Ankle dorsiflexion/plantar flexion
- **Frontal plane**:
  - Pelvis superior/inferior
  - Knee varus/valgus
  - Hip abduction/adduction
- **Transverse plane**:
  - Foot progression angle
  - Pelvis transverse rotation
  - Hip internal/external rotation
  - Knee internal/external rotation

The Vicon system facilitates the calculation of these variables, though its documentation, focused on marker-based measurements, required adaptations for our marker-less system.

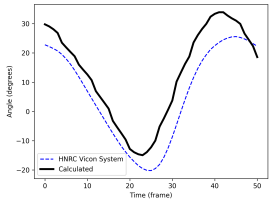
IV. Results and Discussion

Our developed 3D model (see Fig. 3b) captures the nuances of human gait, providing a basis for in-depth analysis and understanding of movement patterns. Out of 11 kinematic variables (see Section III-E) measured at HNRC, four measurements are integrated into the proposed system. This limitation stems largely from the model having insufficient points on certain joints. For example, the model's pelvis is represented by just two points, making it essentially a 2D object and unable to calculate full three-dimensional rotation. However, integrating four out of the 11 variables is notable as

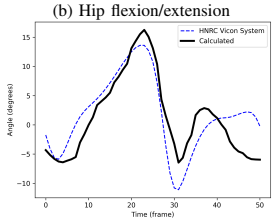
it includes the variables most commonly used by clinicians at HNRC, allowing for a comprehensive analysis of the system's capabilities.



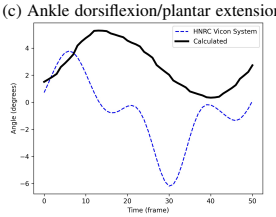
(a) Knee flexion/extension



(b) Hip flexion/extension



(c) Ankle dorsiflexion/plantar extension



(d) Hip abduction/adduction

Fig. 5: Gait analysis - ground truth comparison. The black/bold line represents data calculated using the proposed framework, while the dashed line shows the ground truth data from the Vicon system.

As shown in Fig. 5 and Table III, our study's system performs well with large joint angles, such as knee flexion/extension (Fig. 5a), but less so with smaller angles like hip abduction/adduction. The primary issue stems from MediaPipe's detection process, which does not diminish its overall accuracy but points to the difficulty of accurately depicting complex joints with a single point. At HNRC, gait analysis focuses on assessing patients' kinematic variables and gait parameters. This includes comparing numerical data and analyzing graphs to observe joint angle variations throughout a gait cycle. Special attention is given to the graphical representation of gait patterns, emphasizing the symmetry between the patient's left and right sides and their alignment with standard control data. Given the HNRC clinicians' emphasis on graph shapes, the authors highlight the effectiveness of the proposed framework for analyzing larger angles that significantly influence gait mechanics.

Furthermore, authors report a range of general gait parameters including cadence, single support, double support, final contact, step length, step width, walking speed, and limp index, calculated in close collaboration with HNRC clinicians. These parameters were obtained by projecting foot movements onto the floor plane and aligning walking trajectories along the x-axis, facilitating straightforward calculations, as illustrated in Fig. 6. This method combines in-depth kinematic analysis with essential gait metrics, offering a comprehensive insight into gait mechanics crucial for enhancing the diagnosis and treatment of movement disorders.

V. CONCLUSIONS

This study introduces a markerless system for analyzing cerebral palsy gait using dual RGB cameras and MediaPipe for pose estimation, aimed at enhancing rehabilitation through 3D modeling. It achieved promising correlations for kinematic variables like knee, hip, and ankle in the sagittal plane. The authors detail essential gait parameters like cadence and step length, offering a thorough view into gait mechanics vital for improving movement disorder treatments. Despite limitations with smaller angle changes due to markerless detection, this study lays the groundwork for a cost-effective, efficient gait analysis system, offering a foundation for future research and potential improvements in clinical practice. Future work stands to improve our understanding of motor performance in natural settings, mitigating the influence of unnatural behaviors often observed in clinical assessments.

ACKNOWLEDGMENT

This research is partially supported by the Estonian Research Council ETAG through the research project PRG 2100 "Explainable Artificial Intelligence-based analysis of motor tests for the evaluation of human motor and cognitive functions"

REFERENCES

[1] "What is cerebral palsy?" https://www.cdc.gov/ncbddd/cp/facts.html, Centers for Disease Control and Prevention, Accessed 2023.
[2] E. Ceseracciu, Z. Sawacha, and C. Cobelli, "Comparison of markerless and marker-based motion capture technologies through simultaneous data collection during gait: proof of concept," *PloS one*, vol. 9, no. 3, p. e87640, 2014.
[3] V. Medved, *Measurement of human locomotion*. CRC press, 2000.
[4] M. Moro, G. Marchesi, F. Hesse, F. Odone, and M. Casadio, "Markerless vs. marker-based gait analysis: A proof of concept study," *Sensors*, vol. 22, no. 5, p. 2011, 2022.
[5] R. Maex, A. Castelli, G. Paolini, A. Cereatti, and U. Della Croce, "A 2d markerless gait analysis methodology: Validation on healthy subjects," *BioMed Research International*, vol. 2015, p. 186780, 2015.
[6] S. X. Yang, M. S. Christiansen, P. K. Larsen, T. Alkjær, T. B. Moeslund, E. B. Simonsen, and N. Lynnerup, "Markerless motion capture systems for tracking of persons in forensic biomechanics: an overview," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 2, no. 1, pp. 46–65, 2014.

TABLE III: Pearson correlation between the system proposed in this study and the current Vicon system at HNRC

Angle	Knee Flex/Ext	Hip Flex/Ext	Ankle Flex/Ext	Hip Abd/Add
Trial 1.	0.939	0.969	0.825	0.284
Trial 2.	0.918	0.983	0.819	0.374
Trial 3.	0.853	0.957	0.874	0.430
Avg	**0.903**	**0.969**	**0.839**	**0.362**

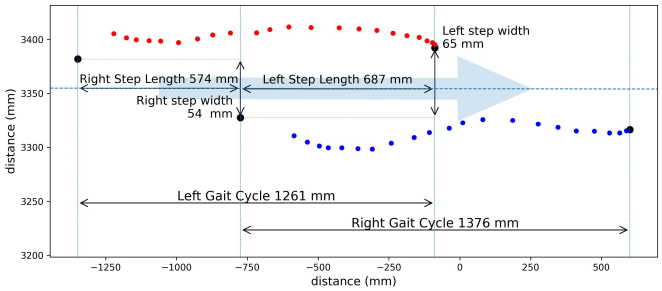


Fig. 6: Projection of heel movement onto the laboratory floor plane for both feet at the HNRC laboratory. The red dots represent left heel movement, the blue dots represent right heel movement, and the black dots indicate stationary heels. The arrow in the background indicates the direction of movement.

[7] M. Nieto-Hidalgo, F. J. Ferrández-Pastor, R. J. Valdivieso-Sarabia, J. Mora-Pascual, and J. M. García-Chamizo, "Gait analysis using computer vision based on cloud platform and mobile device," *Mobile Information Systems*, 2018. [Online]. Available: https://doi.org/10.1155/2018/7381264

[8] R. B. Davis III, S. Ounpuu, D. Tyburski, and J. R. Gage, "A gait analysis data collection and reduction technique," *Human movement science*, vol. 10, no. 5, pp. 575–587, 1991.

[9] A. Leardini, Z. Sawacha, G. Paolini, S. Ingrosso, R. Nativo, and M. G. Benedetti, "A new anatomically based protocol for gait analysis in children," *Gait & posture*, vol. 26, no. 4, pp. 560–571, 2007.

[10] M. P. Kadaba, H. Ramakrishnan, and M. Wootten, "Measurement of lower extremity kinematics during level walking," *Journal of orthopaedic research*, vol. 8, no. 3, pp. 383–392, 1990.

[11] E. D'Antonio, J. Taborri, I. Mileti, S. Rossi, and F. Patané, "Validation of a 3d markerless system for gait analysis based on openpose and two rgb webcams," *IEEE Sensors Journal*, vol. 21, no. 15, pp. 17 064–17 075, 2021.

[12] J. Colombel, D. Daney, V. Bonnet, and F. Charpillet, "Markerless 3d human pose tracking in the wild with fusion of multiple depth cameras: comparative experimental study with kinect 2 and 3," *Activity and behavior computing*, pp. 119–134, 2021.

[13] M. Sandau, H. Koblauch, T. B. Moeslund, H. Aanæs, T. Alkjær, and E. B. Simonsen, "Markerless motion capture can provide reliable 3d gait kinematics in the sagittal and frontal plane," *Medical engineering & physics*, vol. 36, no. 9, pp. 1168–1175, 2014.

[14] D. Balta, G. Figari, G. Paolini, E. Pantzar-Castilla, J. Riad, U. D. Croce, and A. Cereatti, "A model-based markerless protocol for clinical gait analysis based on a single rgb-depth camera: Concurrent validation on patients with cerebral palsy," *IEEE Access*, vol. 11, pp. 144 377–144 393, 2023.

[15] Y. Ma, K. Mithraratne, N. C. Wilson, X. Wang, Y. Ma, and Y. Zhang, "The validity and reliability of a kinect v2-based gait analysis system for children with cerebral palsy," *Sensors*, vol. 19, no. 7, p. 1660, 2019.

[16] M. Yamamoto, K. Shimatani, M. Hasegawa, Y. Kurita, Y. Ishige, and H. Takemura, "Accuracy of temporo-spatial and lower limb joint kinematics parameters using openpose for various gait patterns with orthosis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 2666–2675, 2021.

[17] X. Lv, S. Wang, T. Chen, J. Zhao, D. Chen, M. Xiao, X. Zhao, and H. Wei, "Human gait analysis method based on sample entropy fusion alphapose algorithm," in *2021 33rd Chinese Control and Decision Conference (CCDC)*. IEEE, 2021, pp. 1543–1547.

[18] J. Stenum, C. Rossi, and R. T. Roemmich, "Two-dimensional video-based analysis of human gait using pose estimation," *PLoS computational biology*, vol. 17, no. 4, p. e1008935, 2021.

[19] Y. Ma, K. Mithraratne, N. Wilson, X. Wang, and Y. Zhang, "The validity and reliability of a kinect v2-based gait analysis system for children with cerebral palsy," p. 1660, 2019, accessed: April 7, 2023. [Online]. Available: https://doi.org/10.3390/s19071660

[20] H. Haberfehlner, S. S. van de Ven, S. A. van der Burg, F. Huber, S. Georgievska, I. Aleo, J. Harlaar, L. A. Bonouvrié, M. M. van der Krogt, and A. I. Buizer, "Towards automated video-based assessment of dystonia in dyskinetic cerebral palsy: A novel approach using markerless motion tracking and machine learning," *Frontiers in Robotics and AI*, vol. 10, p. 1108114, 2023.

[21] "deeplabcut documentation." [Online]. Available: http://www.mackenziemathislab.org/deeplabcut

[22] Haapsalu Neurological Rehabilitation Centre. https://www.hnrk.ee/?lang=en. Accessed: April 7, 2023.

[23] mediapipe Contributors, "mediapipe git Documentation," Accessed: 2023. [Online]. Available: https://github.com/google/mediapipe

[24] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, "Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7157–7173, 2023.

[25] OpenCV.org, "OpenCV – open source computer vision library," accessed on April 10, 2023. [Online]. Available: https://opencv.org/

[26] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe, "Metrabs: Metric-scale truncation-robust heatmaps for absolute 3d human pose estimation," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 1, pp. 16–30, 2021.

[27] tensorflow movenet tutorial Contributors, "tensorflow movenet tutorial Documentation," Accessed: 2023. [Online]. Available: https://www.tensorflow.org/hub/tutorials/movenet

[28] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *CVPR*, 2020.

[29] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[30] W. Pirker and R. Katzenschlager, "Gait disorders in adults and the elderly: A clinical guide," *Wiener Klinische Wochenschrift*, vol. 129, no. 3-4, pp. 81–95, 2017.

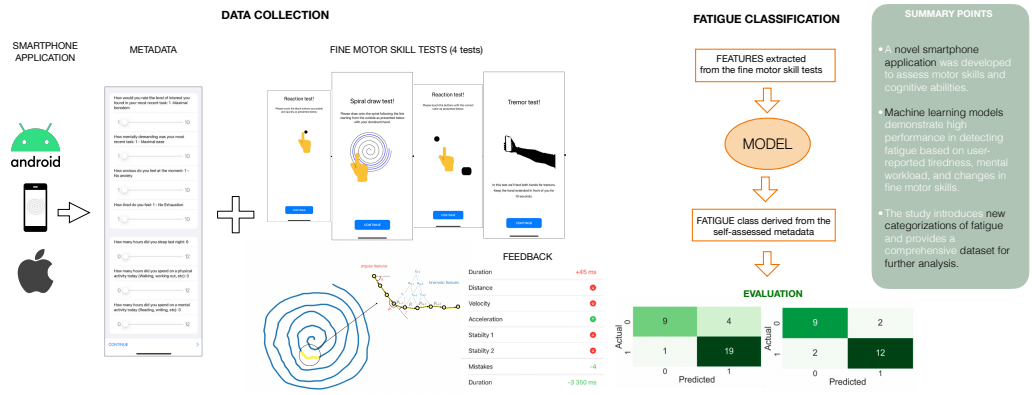
Appendix 8

VIII
Elli Valla, Ain-Joonas Toose, Sven Nõmm, and Aaro Toomela. Transforming fatigue assessment: Smartphone-based system with digitized motor skill tests. *International journal of medical informatics*, 177:105152, 2023

Graphical Abstract

Smartphone-based microkinematic feature analysis for mental fatigue detection using machine learning

Elli Valla, Lilian Väli, Sven Nõmm, Aaro Toomela



Highlights

Smartphone-based microkinematic feature analysis for mental fatigue detection using machine learning

Elli Valla, Lilian Väli, Sven Nõmm, Aaro Toomela

- The study presents a new digital dataset of fine motor skills and self-assessed metadata (SmartPhoneFatigueV2) to the scientific community.

- Machine learning models demonstrate high performance in detecting fatigue based on user-reported fatigue, mental workload, and changes in fine motor skills.

- A new smartphone application was developed to assess motor skills and cognitive health, providing an effective tool for real-time fatigue assessment in various settings.

- The study introduces new categorisations of fatigue enhancing the conceptual understanding and classification of this complex condition.

Smartphone-based microkinematic feature analysis for mental fatigue detection using machine learning

Elli Valla

Department of Software Science, School of Information Technology, Tallinn University of Technology (TalTech), Akadeemia tee 15a, 12618, Tallinn, Estonia

Lilian Väli

Department of Software Science, School of Information Technology, Tallinn University of Technology (TalTech), Akadeemia tee 15a, 12618, Tallinn, Estonia

Sven Nõmm

Department of Software Science, School of Information Technology, Tallinn University of Technology (TalTech), Akadeemia tee 15a, 12618, Tallinn, Estonia

Aaro Toomela

School of Natural Sciences and Health, Tallinn University, Narva mnt. 25, 10120, Tallinn, Estonia

Abstract

Background: Mental fatigue significantly affects cognitive functions and productivity. Recognized in healthcare, education, and workplace safety, there is a growing need for objective fatigue detection tools. Smartphones, with their advanced capabilities, are promising platforms for innovative fatigue detection systems.
Objective: This study aims to develop a smartphone-based system for iOS and Android to detect mental fatigue using fine motor skill tests and a questionnaire, collecting data to train machine learning models.
Methods: Participants completed tasks before and after activities inducing cognitive fatigue, using 166 devices resulting in 347 sessions. From raw signals, 60 features were engineered. Dimensionality reduction used wrapper-type feature selection. Six machine learning algorithms were employed: Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Random Forest, and AdaBoost, cross-validated with a nested k-fold approach.
Results: The study achieved 0.86 sensitivity using machine learning models with self-assessed fatigue and mental work hours. Key features included anxiety, effort scales, and the kinematic feature linked to handwriting changes and hand tremors. The most effective model combined these features with data from fine motor skill tests, highlighting a multi-dimensional approach to fatigue classification.
Conclusion: The findings offer applications in workplace safety, education, and healthcare, and the dataset provides a valuable resource for further research into cognitive and motor functions.

Keywords: fatigue, machine learning, fine motor skills, microkinematics, smartphone application, eHealth, dataset

1. Introduction

Fatigue poses significant challenges in various industries, contributing to accidents and decreased productivity. Studies have shown that fatigue, often resulting from prolonged cognitive or physical activity, sleep deprivation, and other factors, is linked to stress, aging, depression, illness, and neurological disorders [1][2][3]. It manifests in two main types: peripheral (physical) and central (mental) [4], affecting performance in tasks demanding alertness and information retrieval [5].

Email addresses: elli.valla@taltech.ee (Elli Valla), livali@taltech.ee (Lilian Väli), sven.nomm@taltech.ee (Sven Nõmm), aaro.toomela@tlu.ee (Aaro Toomela)

Extensive research has explored fatigue's nuanced effects on cognitive and physical performance. Studies have demonstrated its detrimental impact, such as increased reaction times and errors in tasks requiring prolonged attention [6], and impaired physical performance following cognitive tasks [7]. Additionally, a substantial study involving 228 participants revealed fatigue's progressive impact on attention [8], while another examined varying cognitive tasks, linking task complexity with fatigue severity [9]. In real-world contexts, occupational fatigue has been shown to significantly impair cognitive functioning [10]. Another interesting study analysed the effect of fatigue on speech over 24 hours, noting changes in speech patterns [11].

Advancements in technology have enabled the use of smartphone sensors in monitoring physical and cognitive abilities, offering new avenues in health and well-being monitoring in free-living environments [12][13][14][15][16][17]. This paper explores the concept of fatigue through fine motor ability, aiming to contribute to fatigue management strategies.

1.1. Related Work

Recent studies have leveraged machine learning (ML) and smartphones for fatigue detection. Hooda *et al.* [18] reviewed 67 articles on fatigue detection using ML, finding that combining biological and physical features yields high accuracy. Jasim *et al.* [19] compared 48 papers on drowsiness detection, recommending a hybrid approach combining physiological and behavioral features for optimal results.

In Parkinson's disease research, handwriting tasks have been analysed using ML techniques to diagnose the condition [20]. Similarly, a mobile app was developed for assessing Multiple Sclerosis-related fatigue [21], offering real-time symptom monitoring. The Untire mHealth app was found effective in managing fatigue among cancer patients and survivors [22]. For Barth syndrome, a phone app measured fatigue in real-time, differentiating between affected and control participants [23].

In contrast to these qualitative approaches, this study combines qualitative and quantitative data from smartphone sensors. In Valla *et al.* [24] authors used an Android app for fine motor skill assessment, achieving promising results in fatigue prediction. Research has also linked fatigue with impaired motor skills. Studies [25][26] observed slowed movements and increased movement duration under mental fatigue. Senkiv *et al.* [27] employed a spiral drawing test to evaluate mental fatigue levels, and Lippold *et al.* [28] connected muscular fatigue to hand tremors, though Budini *et al.* [29] found no significant tremor differences with mental fatigue.

These works demonstrate the potential of ML in fatigue detection for various applications. However, existing solutions require improvements for broader application and increased accessibility.

This research aims to combine motor skills analysis and fatigue assessment using a smartphone application, hypothesizing that data from smartphones can effectively differentiate various fatigue states: self-assessed fatigue, physical exertion, and mental exertion.

1.2. Problem Statement

The primary aim of this research is to explore the possibility of detecting mental fatigue using a smartphone application designed for fine motor skill assessment. The central research question guiding this investigation is: **Can mental fatigue be effectively identified through fine motor skill tests administered via a smartphone application?**

This study strives to broaden the application of smartphone apps for tasks that assess fine motor skills, employing machine learning (ML) models to identify signs of mental fatigue. A critical element of this study is the enhancement of data collection methods, focusing on evaluating changes in fine motor skills before and after tasks that induce mental fatigue. Additionally, the research aims to refine user questionnaires to broaden the dataset, investigating if these improvements can boost the ML model's accuracy in detecting fatigue.

Furthering previous endeavors, this study will analyse an existing Android application [24] and develop an updated iOS version to widen the scope of participant diversity. The novel contributions of this research are multifaceted:

1. Advanced application for analysing motor skills and fatigue, offering an innovative tool for fatigue assessment.

2. Collection of new data to evaluate the effectiveness of detecting fatigue through digitised fine motor skill tests, contributing to the empirical understanding of fatigue detection.

3. Presentation of the collected dataset to the scientific community for further research, facilitating collaborative advancements in the field.

4. Introduction of new proposed categorisations of fatigue, enhancing the conceptual understanding and classification of this complex condition.

5. Development of new machine learning models with improved performance in detecting fatigue, showcasing the potential of ML in health and well-being applications.

This approach not only leverages technological advancements in smartphone capabilities but also contributes significantly to the ongoing discourse in fatigue research through innovative data collection and analysis techniques.

The rest of this paper is organised as follows. Section 2 describes the materials and methods used to develop the proposed framework. Section 2.9 and Appendix 1 introduce the characteristics of the published dataset. Sections 4 and 5 report and discuss the experimental results to highlight the effectiveness of the proposed method. The paper is summarised in Section 6.

2. Methods

The subsequent sections provide a comprehensive overview of the development process and functionalities of the smartphone application designed for fatigue detection. Additionally, these sections detail the data collection procedures, describe the collected dataset, and introduce the ML pipeline employed in this study.

2.1. Development of the Smartphone Application for Fatigue Detection

The research outlined in [24] has been extended to include further development of two key software components: the back-end application and the Android mobile application. Concurrently, a dedicated mobile application tailored for iOS devices (see Appendix 3) was initiated from scratch.

The back-end system underwent a significant update, integrating advanced logic within its controller. This logic is vital for discerning whether a device is being utilised for initial or subsequent test attempts within the application. A critical feature of this update is the enforcement of a time interval between test attempts. Moreover, novel methodologies were implemented to provide users with feedback regarding any improvements or regressions between their first and second attempts. An additional endpoint was also established to facilitate the retrieval of test data over specified date ranges.

A depiction of the application workflow is provided in Figure 1. This high-level flow chart illustrates the interactions among the different components of the system.

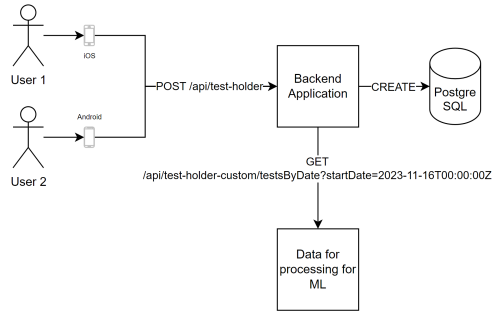


Figure 1: Back-end workflow diagram for the smartphone-based mental fatigue detection system

In total, the user is required to complete four different tests within the application. This section presents the application's workflow in chronological order together with screenviews captured from the iOS application. When opening the application, users are first prompted to agree to the terms of use, a prerequisite for further interaction with the app. The user can read the terms of use by tapping on "Click here to read our Terms of Use" which directs the user to the document. The terms of use document is brought out in Appendix 2. After accepting these terms, users are given general instructions for performing the tests.

The questionnaire designed to collect qualitative metadata from the users before the first completion of the test is depicted in Figure 2. The metadata includes the participants's gender, age, height, weight, education level, dominant hand, self-evaluation of fatigue, and assessment of the nature of daily activities. After completing the fatigue-inducing task (such as a lecture, meeting, or exam), the user's level of interest in their most recent task, assessment of its mental demands, anxiety level, and exhaustion level are recorded. These assessments range from 0 to 10. Additionally, information regarding the number of hours spent on physical and mental activities during the day and the previous night's sleep is collected, with values ranging from 0 to 12.

2.2. The Reaction Test - Simple

The Reaction Test Simple (RTS) is the first test within the application and is designed to evaluate the user's response times, accuracy, and mistakes. In this test, the user is expected to tap on black dots that appear at various locations on the screen in a randomised manner, each differing in size. The total count of these black dots that the user must hit is fifteen. Instruction for the

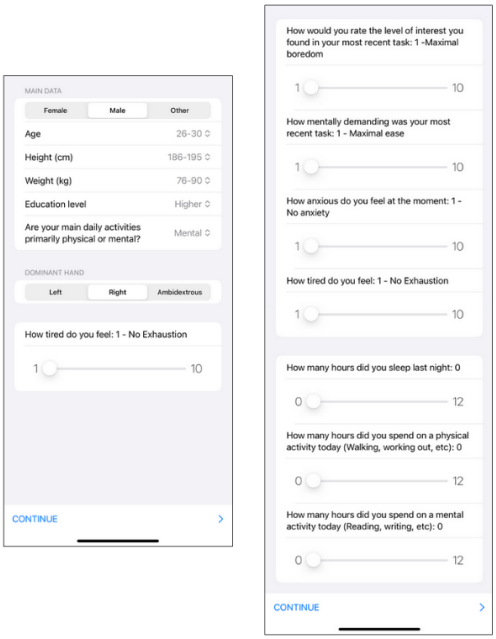


Figure 2: Questionnaire for collecting qualitative metadata from users before test initiation.

user on how to execute the RTS is provided through an animated tutorial, which demonstrates the appropriate method for undertaking the test. The user's workflow in this test is brought out in Figure 3.

The application records several parameters during the test: each screen tap, the coordinates of these taps, the accuracy of tapping directly on the black dots, the elapsed time in milliseconds between taps, and the dimensions of the screen of the user's smartphone. Moreover, the application also tracks the duration from the moment the user initiates the test to the point where the fifteenth black dot is tapped. The test starts when the user taps the green 'START' button (shown in the second section of Figure 3).

2.3. The Archimedean Spiral Drawing Test

The Archimedean Spiral Drawing Test (ASD) is the second test within the application and is designed to have the user draw a spiral while maintaining the line within specified boundaries. Instructional guidance for this test is provided to the user through an animated tutorial, which demonstrates the correct technique for performing the spiral drawing task. The user's workflow in this test is brought out in Figure 4.

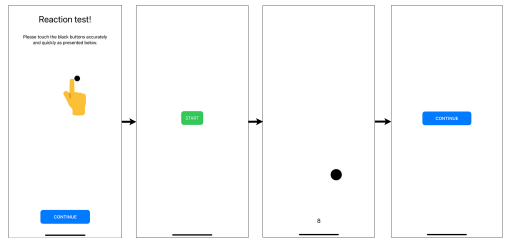


Figure 3: Screen views of the first reaction test (RTS) in the application.

Several key metrics are recorded during this test. These include the height and width of the drawable area on the screen (depending on screen size), the coordinates of each point of the line drawn by the user, and an assessment of whether each point coincides with the pre-defined background line. Additionally, the total duration taken by the user to complete the spiral drawing is measured. Another feature of the test is the real-time calculation of the percentage of the drawing that aligns with the background line, which is incorporated into the resulting data object after the completion of the test.

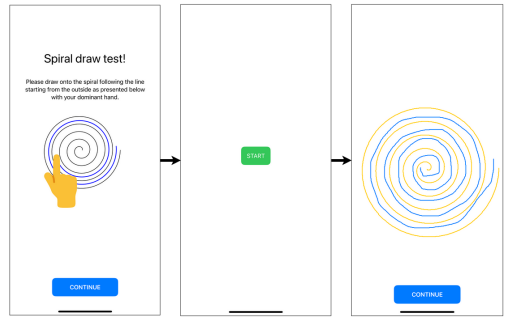


Figure 4: Screen views of the Archimedean spiral test in the application.

2.4. The Reaction Test - Advanced

The Reaction Test Advanced (RTA) is the third test within the application and is designed to challenge users to tap on dots that correspond with a colour indicated at the bottom right of the screen. The dots appear at various locations on the screen in a randomised manner each differing in size and colour. This test features featuring four pre-selected colours - purple, blue, yellow, and black. The user's task is to accurately tap on a dot when its colour matches the indicated colour. An animated tutorial is provided to instruct users on the proper

execution of this test. The user's workflow in this test is brought out in Figure 5.

This test records a variety of metrics: the height of the screen, the coordinates of each tap, the accuracy of tapping on the correct dot, the elapsed time since the last tap, and the time elapsed since the first appearance of a correctly coloured dot. Additionally, the total duration taken by the user to complete the test is also captured. The test starts when the user taps the green 'START' button (shown in the second section of Figure 5) and finishes when the last correct dot is tapped.

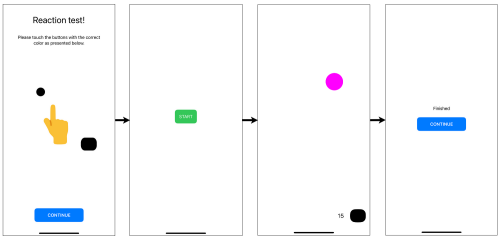


Figure 5: Screen views of the advanced reaction test (RTA) in the application.

2.5. *The Tremor Test*

The Tremor Test is the last test within the application and is designed to measure the hand tremors of the user. The users are expected to extend one hand outward while initiating the test by pressing the start button on the screen with their other hand. This test is repeated with both hands. Instructional guidance for this test is conveyed through an image, which demonstrates the correct method for conducting the tremor test. The user's workflow in this test is brought out in Figure 6.

During this test, the smartphone's accelerometer sensors actively measure the hand's movements in all directions over 10 seconds. The test is to be conducted identically with both hands to ensure consistent data collection starting with the left hand. The first half of the test starts with left-hand measurements when the user taps the green 'START LEFT HAND' button (shown in the second section of Figure 6) and finishes when 10 seconds have passed (timer shown in the third section of Figure 6). The second part of the test for the right hand is identical to that of the left hand as seen from Figure 6.

2.6. *Last application view and feedback*

Upon completing the tests for the second time, the application's back-end processes the data from both sessions and displays the results to the user, as depicted in

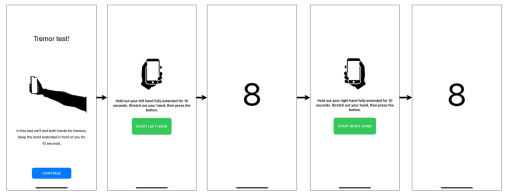


Figure 6: Screen views of the Tremor test in the application.

Figure 7 on the right. Results for each test are presented individually. For the RTS, the feedback shows the difference in the number of mistakes and the change in test duration. The RTA similarly provides feedback on variations in the number of errors and duration. In the case of ASD, feedback not only includes changes in mistakes and duration but also other metrics, marked with a green upward arrow for improvements and a red downward arrow for declines. Likewise, the Tremor test feedback displays changes in hand asymmetry using green upward or red downward arrows, paralleling the feedback format of ASD.

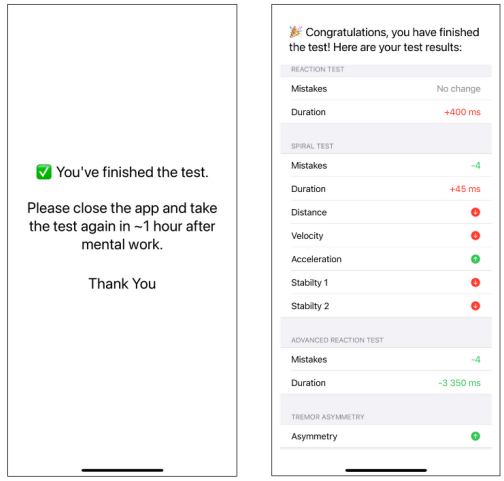


Figure 7: Success message displays upon test completion: initial (left) and subsequent (right).

2.7. *Data collection and analysis*

The principal methodology for data acquisition in this study involved a collaborative arrangement with universities to facilitate data collection during academic lessons. The process commenced with a preliminary

presentation to the students, outlining the research objectives and introducing the functionalities of the application. Subsequently, students were encouraged to download the application, fill out the questionnaire, and perform the tasks. A follow-up session was scheduled post-lesson to prompt students to complete the application tasks a second time, ensuring an inter-test interval of approximately 1.5 hours.

Additionally, as a secondary approach to data collection, comprehensive information documents were disseminated to various educational institutions. These documents explicitly detailed the test completion procedures and articulated the specific types of data being collected, along with the underlying reasons for their collection.

A tertiary method involved circulating informational documents within personal and professional networks. Participants in this group were instructed to initially undertake the application's tasks, engage in a mentally strenuous activity comparable to an academic lesson or a cognitively demanding professional meeting, and subsequently revisit the application's tasks for a second assessment.

The decision number 12 by the Tallinn University Board of Ethics, dated May 12, 2021, established guidelines for the process of data collection.

2.7.1. Experimental setting

In the context of this research, the primary experimental protocol necessitated a two-phase engagement with the application. Initially, participants were obliged to answer a series of foundational questions and execute four tasks within the application, each designed to assess fine motor skills. After this preliminary interaction, participants were involved in activities designed to induce cognitive fatigue for a duration of no less than one hour, optimally extending to ninety minutes. These activities varied, encompassing academic lessons, cognitively demanding non-physical work, or professional meetings, to simulate real-world scenarios that could increase mental fatigue.

Upon completion of these cognitively demanding activities, participants were asked to return to the application for a second session of questionnaires and task performance. This follow-up interaction was especially important for evaluating potential changes in fine motor skills, which are hypostudyed to be indicative of fatigue.

The data collection process was meticulously structured to detect subtle changes in motor skill performance related to cognitive fatigue. Each participant was subsequently provided with personalised feedback, based on a comparative analysis of their performance metrics across the two test sessions. This approach aligns with the study's aim of deploying ML models to identify fatigue by analysing shifts in fine motor skills as measured through smartphone application usage.

2.8. Feature extraction and engineering

Time series data, such as finger positions (represented by $x-$ and $y-$coordinates) and timestamps, are utilised to derive a variety of features. This section, along with Table 1, outlines the engineered feature set developed from this data.

Kinematic Features: For any given timestamp, the velocity of the position vector $\vec{r} = [p_i, p_{i+1}]$ can be calculated. Essentially, velocity measures how quickly the position vector's displacement changes over time. Similarly, acceleration is determined as the rate of change of velocity, and jerk as the rate of change of acceleration, with respect to time. This analysis extends to the sixth time derivative of the position vector. While there are no universally accepted terms for the fourth and higher time derivatives of displacement, the terms snap, crackle, and pop are commonly used in literature for the fourth, fifth, and sixth derivatives, respectively [30]. These higher-order derivative features, which reflect micro-changes in movement acceleration, have been applied in studies such as [31] for Parkinson's disease diagnostics. Figure 8 visually depicts these differential-type features.

Angular Features: The angle α of a position vector can be derived from its slope k. Considering N observation points, where (x_i, y_i) are the coordinates of point p_i for $i \in \{1, 2, \ldots, N\}$, the slope (k) of the line and the corresponding angle α are calculated as follows:

$$k = \frac{y_i - y_{i-1}}{x_i - x_{i-1}}, \tag{1}$$

$$\alpha = \arctan k, \tag{2}$$

Figure 8 depicts all the angles that are considered in the current research:

$$\phi_i = \pi + \alpha_{i-1} - \alpha_i \tag{3}$$

$$\gamma_i = \alpha_i - \alpha_{i-1} \tag{4}$$

Yaw (γ) represents the change in the direction that the point vector points toward. The set of angular features was expanded to include up to the third derivative of yaw over time.

A study by [32] demonstrated that integral-like features, derived from kinematic parameters and pressure measurements, have significant discriminating power

to distinguish between Parkinson's disease patients and healthy control subjects. Furthermore, research in [27] showed that these features could potentially enable ML techniques to detect mental fatigue; thus, they have been incorporated into this study. To ensure the self-sufficiency of this paper, the computational methods for these motion parameters are detailed below.

Motion Mass Parameters: Introduced by [33], motion mass parameters quantify the amount and smoothness of motion across a limb or group of joints. These parameters are calculated as the sum of the absolute values of each kinematic and geometric parameter that varies during the test. Let N represent the number of observation points in the test (or a segment of the test). If v_k denotes the velocity along the directional vector of the stylus movement at observation point k, where $k \in \{1, \ldots, N\}$, then the *velocity mass* is defined by the following equation:

$$V_N = \sum_{k=1}^{N} |v_k| \tag{5}$$

2.9. Database for Fatigue Assessment Through Digital Fine-Motor Skill Tests (SmartPhoneFatigueV2)

The tests were completed a total of 347 times. A significant portion of the dataset, comprising 94 devices, completed the tests precisely twice, while 35 devices registered a single test completion. In total, 166 unique devices were involved in completing the tests using mobile applications. Appendix 1 offers insights into the demographics and characteristics of the participants. Out of the total data entries, a notable distribution was observed in terms of the operating systems used for recording the data with 51.2% using Android and the remainder, a close 48.8%, opting for iOS. This highlights the importance of having a cross-operating system application, ensuring the data collection process is inclusive and representative of both major mobile platforms.

The collected data includes a comprehensive set of features calculated from the users' interactions with the application. Section 2.7 describes these computed features, including metrics such as Euclidean distance, jerk, angular velocity, and cumulative slope angles, which provide detailed insights into the dynamics of user movements. These features are critical for assessing fine motor skills and detecting fatigue.

In refining the dataset for improved analytical accuracy, a careful approach was adopted for the Tremor Test data. Analysis of user interactions during university lessons revealed a common deviation from the instructed procedure. Notably, many participants tended

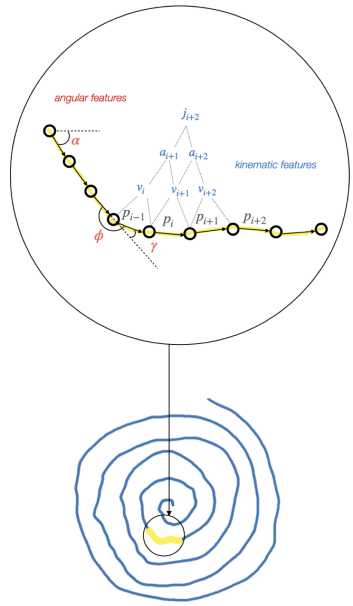


Figure 8: Visual representation of the differential-type (a) and angular-type (b) features

.

to reverse the recommended sequence of actions: instead of extending their hand before initiating the Tremor Test via the screen button, they pressed the button before extending their hand. This pattern is illustrated in Figure 9, showing the button press preceding hand extension. To ensure data integrity, the initial 15% portion of time in each tremor test dataset was systematically excluded from the analysis.

In addition to removing all instances with missing values, entries showcasing the smallest distance in the spiral drawing task were rigorously examined through visual inspection using the front-end application. This step was necessary to verify the accuracy of both the length and shape of the spiral drawings. Following the purification procedure, the dataset was reduced to a total of 343 records.

After segmenting the dataset following the completion of tests within the prescribed timeframe, the dataset diminished to a count of 218 records. These were evenly split into two groups: 109 entries in the 'before' group and 109 in the 'after' group. This partitioning was es-

Table 1: A subset of engineered features derived from fine motor skill test data. See Figure 8 for visual representation.

Test name	Feature set	Description
ASD	distance	$d_i = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}$ (Euclidean distance)
	jerk	The rate of change in acceleration with respect to time, defined as the third time derivative of distance. It quantifies how quickly acceleration is increasing or decreasing, providing insights into the dynamics of motion.
	ϕ_velocity	The rate of change of the angle ϕ, measured in radians per unit time, capturing the angular velocity throughout the test.
	slope_mass	The cumulative sum of the slope angles measured across all observation points during the test. It is indicative of the smoothness of fine motor skills, reflecting the overall consistency and fluidity of the movements recorded.
RTS and RTA	wasHitOnTarget	Boolean values True if the area of the touch overlaps with at least one pixel of the rendered circle.
	timeFromLastTouch	Time between touches
	timeFromFirstCorrect-ColorRender	The difference in time between two matching color renders
Tremor	x, y, z	Acceleration along $x-, y-, z-$axis
	absolute acceleration	$abs = \sqrt{x^2 + y^2 + z^2}$

sential for subsequent ML analysis, ensuring a balanced and precise dataset that accurately reflects the test sessions.

3. Machine Learning Based Analysis for Fatigue Detection

In the context of detecting mental fatigue using ML algorithms, it is imperative to categorise the data into two distinct labels: 'fatigued' and 'non-fatigued'.

3.1. Fatigue categorisation

Initially, the differentiation between 'non-fatigued' and 'fatigued' states was determined through the completion of mental tasks in two sequential sessions, with the presumption that the first session represents a 'non-fatigued' state and the subsequent session signifies a 'fatigued' state. Furthermore, the extent of mental exertion encountered over the course of a day was considered as a criterion for labelling. This was followed by incorporating the duration of sleep attained as a parameter for label assignment. Finally, self-assessment of fatigue levels was also integrated into the labelling process, providing a subjective measure to the classification scheme. The criteria for the fatigue categories are systematically outlined in Table 2.

3.1.1. "Fatigue inducing tasks" as a label

In the process of the first data categorisation, the labels were determined based on the timing of the test relative to the lesson. Consequently, this resulted in the formation of two distinct groups: Group 1, comprising 109 data entries for tests conducted before the fatigue inducing task (lesson, lecture, exam, meeting, etc), and Group 2, also consisting of 109 entries, conducted after. It is important to note that certain features - specifically those pertaining to effort, interest, anxiety, and self-assessed fatigue - were not recorded during the initial completion of the tests.

Table 2: Fatigue categories used for ML-based classification.

Category	Threshold	Count / Label
Performing a fatigue-inducing task	Before the lesson After the lesson	109 (non-fatigued) 109 (fatigued)
Mental work performed in hours v1	> 1	103 (non-fatigued) 115 (fatigued)
Mental work performed in hours v2	> 2	140 (non-fatigued) 78 (fatigued)
Sleep hours v1	< 6	136 (non-fatigued) 30 (fatigued)
Sleep hours v2	< 7	104 (non-fatigued) 62 (fatigued)
Sleep hours v3	< 8	42 (non-fatigued) 124 (fatigued)
Self-assessed fatigue 1	≤ 3 ≥ 6	69 (non-fatigued) 44 (fatigued)
Self-assessed fatigue 2	≤ 3 ≥ 7	69 (non-fatigued) 24 (fatigued)
Self-assessed fatigue 3	≤ 2 ≥ 8	51 (non-fatigued) 14 (fatigued)
Self-assessed fatigue 4	= 1 ≥ 6	40 (non-fatigued) 44 (fatigued)

Furthermore, to standardise the data, the values for physical work hours recorded during the first test completion were inferred from the corresponding values of the second completion. In addition, the mental work hours for the first test completion were adjusted to be one hour less than those recorded in the second completion. This adjustment was made to account for the time elapsed between the two test completions. The sleep hour data was excluded from the dataset due to a prevalence of zero values, which indicated a lack of variability and reliability in this particular measure.

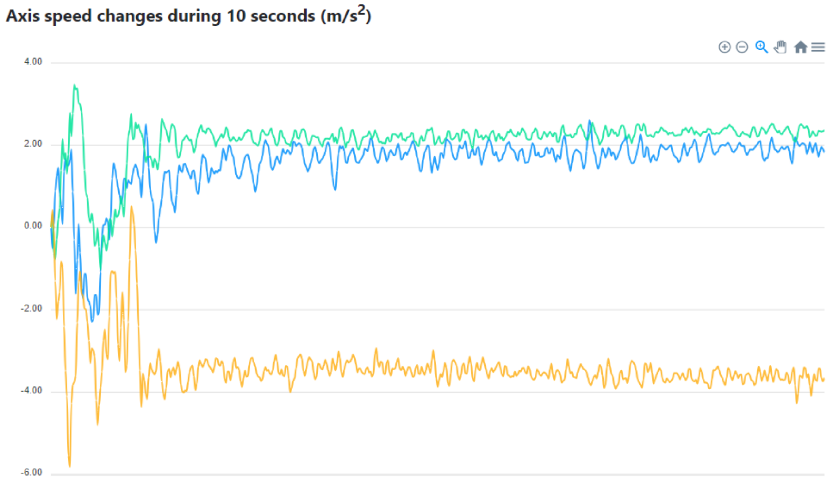


Figure 9: Example output of right-hand tremor measurements.

3.1.2. "Hours of mental work performed" as a label

In the analysis focusing on the hours of mental work performed, several columns were excluded from the dataset due to the prevalence of zero values or because they were not recorded during the initial test completion. Specifically, the columns representing sleep scale, effort, interest, anxiety, and self-assessed fatigue were omitted from consideration. To maintain consistency across test completions, the values for physical work hours recorded during the first test were inferred from their counterparts in the second test. Additionally, the mental work hours for the first test were adjusted to be one hour less than those for the second test, acknowledging the passage of time between the two sessions.

A further breakdown of the data reveals that when more than two hours of mental work were performed, the participants were classified as 'fatigued' in 78 instances and 'non-fatigued' in 140 instances. Similarly, when the mental work exceeded one hour, there were 115 instances classified as 'fatigued' and 103 as 'non-fatigued'.

3.1.3. "Hours of sleep" as a label

In the third phase of the analysis, we focused on sleep hours. The classification of the 'non-fatigued' group was based on varying thresholds of sleep duration. For instance, defining 'non-fatigued' as individuals who slept more than 5 hours resulted in 136 individuals in the 'non-fatigued' category and 30 in the 'fatigued' category. Altering this threshold to more than 6 hours of sleep reclassified the groups, resulting in 104 individuals in the 'non-fatigued' category and 62 in the 'fatigued' group.

Further adjustment of the threshold to over 7 hours of sleep caused a notable shift in group sizes, with 42 individuals categorised as 'non-fatigued' and 124 as 'fatigued'. These varying group sizes based on sleep duration thresholds are likely to influence the sensitivity and specificity of the model. Sensitivity, or the true positive rate, could be affected by the smaller size of the 'non-fatigued' group at certain thresholds, potentially leading to a higher rate of false negatives. Similarly, specificity, or the true negative rate, might be impacted by the larger 'fatigued' group sizes, influencing the model's ability to correctly identify true negatives.

3.1.4. "Self-assessed fatigue level" as a label

In the analysis, self-assessed fatigue was a focal point. A meticulous data sorting process resulted in 155 relevant data rows for analysis. We tested different value ranges for labeling fatigue levels. Initially, entries with self-assessed fatigue rated at levels 4 or 5 were excluded, resulting in 113 rows. For classification, ratings above 5 were labeled as 'fatigued', yielding 69 'non-fatigued' and 44 'fatigued' instances. To further refine the dataset, we excluded the value 6. Ratings of 7-10 were categorised as 'fatigued', and 1-3 as 'non-fatigued', effectively eliminating moderate fatigue levels. This adjustment produced 69 'non-fatigued' and 24 'fatigued' instances, enhancing the separation between

the two classes. Further refinement involved excluding values 3-7, resulting in 51 'non-fatigued' and 14 'fatigued' instances. In the most stringent classification, using a value of 1 to indicate 'non-fatigued' and values over 5 to indicate 'fatigued', we achieved a distribution of 40 'non-fatigued' and 44 'fatigued' instances.

These partitioning strategies aimed to increase the decision boundary separation for ML models, thereby improving their ability to accurately classify and predict fatigue levels.

3.2. Machine Learning Pipeline

A total of 60 features were engineered from the raw signals. Most discriminative predictors were selected to reduce dimensionality using wrapper-type feature selection procedures. Feature selection was systematically conducted, choosing subsets of 1, 2, 3, 4, 5, and 10 features, utilising methods such as Recurrent Feature Elimination (RFE) [34] and SelectFromModel (SFM). Given the inherent diversity in the algorithmic nature of ML models, six distinct algorithms were selected for this study.

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Decision Tree (DT)
- Random Forest (RF)
- AdaBoost (AB)

These were trained and cross-validated in a nested k-fold loop. It means that supervised feature selection strategies are nested within the cross-validation iterations so that the most discriminating features are chosen based only on the training set, while the test set is kept only for validation [35]. This way, we avoid possible bias that may lead to model overfitting. For each iteration within this process, the dataset was standardised using the StandardScaler. Critical evaluation metrics including accuracy, precision, sensitivity, specificity, and F1 score were meticulously measured to gauge model efficacy. The overall workflow is visualized in Figure 10.

4. Results

A nested cross-validation was utilised to determine the most effective models, feature selectors, and labelling strategies. This method was essential for identifying the optimal setup. The fatigue classification categories, critical to the model's functionality, are thoroughly outlined in Table 2.

4.1. Best performing models for fatigue detection

The models were trained using the insights gained from cross-validation and applied to the entire dataset, which was divided with a split of 1/3 (test) and 2/3 (train). Table 3 presents the top models that showed the best performance. The highest accuracy was observed in a model using the RF classifier with 10 features, achieving an accuracy of 0.85. This high accuracy can be attributed to a combination of features from the spiral test, tremor tests, and the simple reaction test, along with self-assessed features.

Further analysis of these features shows that a combination of the calculated *slope_mass* with self-assessed effort and anxiety, as well as hours of previous physical work, also resulted in a high accuracy of 0.84, even with a reduced set of only four features. Notably, removing the feature related to physical work hours decreased the accuracy to 0.80, highlighting its importance in the effective detection of self-assessed fatigue.

5. Discussion

The results obtained from ML models have yielded valuable insights into fatigue detection. The expanded dataset, coupled with the inclusion of additional questions in the questionnaire, contributed to achieving superior results. In a previous study that exclusively utilised an Android application, the peak accuracy recorded was 0.79 [24]. However, in this current research, a significantly higher accuracy of 0.85 was achieved, employing self-assessed fatigue and hours of mental work as labels. This finding underscores the feasibility of using self-reported fatigue levels and mental workload as possible fatigue detection indicators. Analysis revealed that feature selectors prominently identified the anxiety and effort scales, along with calculated features, as key contributors to these robust detection results. It was observed that the trajectory angle of the spiral drawing, *slope_mass*, emerged as a particularly vital component in training the models. It has been documented that these angular type features (α, ϕ, etc., see Figure 8) are describing some forms of micro-changes in handwriting and can be linked to hand tremors [36]. These features capture nuanced fluctuations in fine motor movement, providing valuable insights into the effects of fatigue on motor performance. This highlights the capability of ML algorithms to discern between

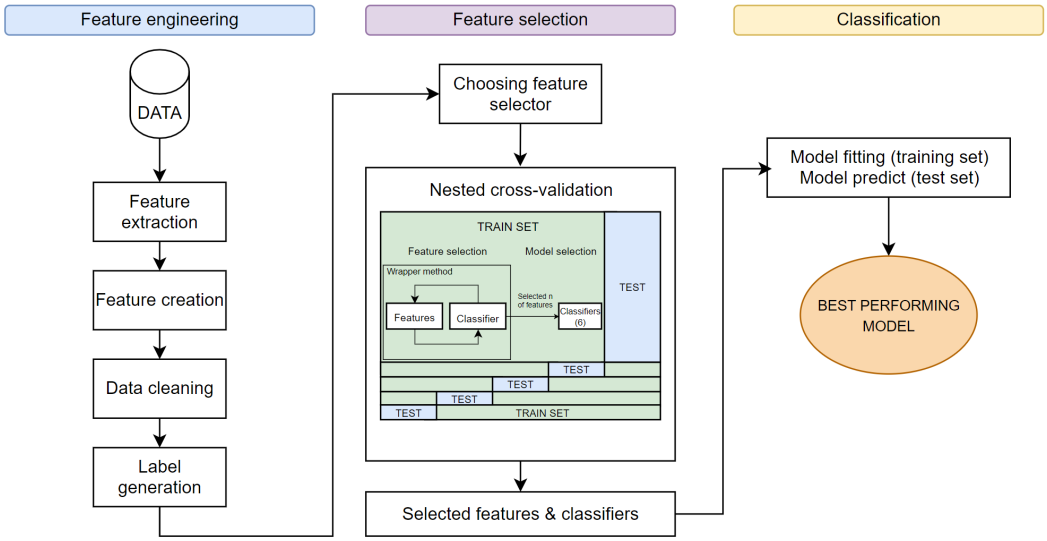


Figure 10: A wrokflow for the machine learning based classification.

these two states based on the study's utilised features, which, although subtle and imperceptible to the naked eye, possess informative value for classification. Furthermore, the most effective model integrated kinematic and tremor features with self-assessed categories and the reaction test, enhancing its performance. These findings have significant implications for developing fatigue detection systems, emphasising the importance of subjective self-assessment and specific psychometric scales in enhancing system accuracy.

In the medical field, the implications are substantial. Healthcare professionals could benefit from a reliable fatigue assessment tool that extends beyond self-reported measures. Patients with chronic conditions, neurological disorders, or undergoing medical treatments could use this tool for objective fatigue level monitoring.

The model's applicability extends beyond the medical realm to industries where human performance is critical, such as aviation, transportation, and manufacturing. Implementing fatigue detection systems could enhance safety and productivity. Professionals in demanding environments, like pilots, truck drivers, or shift workers, could benefit from real-time fatigue assessments for informed work and rest decisions.

In education, this technology could assess and manage student fatigue during exams or academic activities. Identifying fatigue patterns allows educators to adjust curriculum and schedules, optimising learning outcomes.

The dataset collected offers potential for diverse applications beyond fatigue detection, including reaction tests, spiral drawing tests, and hand tremor assessments. This opens up novel research and practical application avenues in various domains.

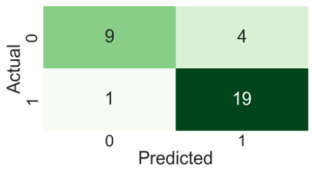
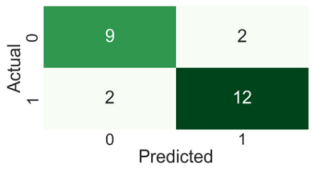
The reaction test data, indicative of cognitive processing speed and motor function, could be leveraged for applications in cognitive neuroscience and motor control studies. This extensive dataset could offer insights into cognitive performance, reaction time variability, and motor coordination, valuable for studying cognitive impairments or evaluating cognitive-enhancing interventions.

Spiral drawing tests provide opportunities for exploring fine motor skills and coordination. The dataset's detailed information on drawing patterns and stability parameters could aid research in motor skill development, therapy impact assessment, or digital art and design applications.

Hand tremor tests offer unique insights into tremor patterns and potential links to health conditions. The dataset could be invaluable for tremor assessment, aiding in early detection of conditions like essential tremor or Parkinson's disease, and analysing tremor characteristics in relation to demographic and health factors.

Looking ahead, there are several promising directions

Table 3: Best performing ML models for fatigue classification.

Category	Features	Classifier	Accuracy	Precision	Sensitivity	Specificity	f1	Confusion Matrix
Mental work hours < 1	'alpha_jerk', 'phi_acceleration', 'yaw_acceleration', 'jerk', 'y_l', 'x_r', 'physicalWorkScale', 'effortScale', 'interestScale', 'timeFromLast-Touch_rts'	RF	0.85	0.86	0.86	0.82	0.84	
Self-assessed fatigue < 2 & > 5	'slope_mass', 'slope', 'alpha', 'phi_velocity', 'distance', 'Z', 'wasHitOnTarget', 'physicalWorkScale', 'effortScale', 'anxietyScale'	RF	0.84	0.86	0.86	0.82	0.86	

for future research. Expanding the dataset size would be beneficial, as larger datasets can provide more comprehensive training for the models, potentially improving their accuracy and robustness. Additionally, the application of explainable AI techniques would offer valuable insights by elucidating the underlying decision-making processes of the models, thereby enhancing our understanding of their predictive capabilities. The testing suite within the smartphone application has potential for further advancement by incorporating microphone and camera-based tests. This would leverage additional smartphone sensors, enriching the testing capabilities and overall functionality.

6. Conclusion

This study developed a smartphone application to assess mental fatigue using fine motor skill tests and a comprehensive questionnaire. The app, compatible with both iOS and Android, collected data through tasks conducted before and after potentially mentally exhausting activities. A two-phase data collection process was used to capture cognitive changes through these fine motor skills tests. The study then used machine learning techniques to analyse the data and develop predictive models for mental fatigue.

Key findings highlighted the predictive power of self-reported fatigue level and workload, alongside changes in fine motor skills, in assessing fatigue. The application and resulting models offer significant potential for real-time fatigue monitoring in environments such as safety-critical workplaces and educational settings. Moreover, the dataset generated from this research provides a valuable resource for future studies on cognitive and motor functions. Overall, this study contributes to the field by offering a comprehensive tool and dataset that enhances our ability to understand and monitor mental fatigue in various contexts.

Acknowledgements

This work was supported by the Estonian Research Council grant PRG 2100.

Data Availability

The presented dataset will be uploaded upon acceptance to the public repository: UCI Machine Learning Repository

References

[1] K. Martin, R. Meeusen, K. G. Thompson, R. Keegan, B. Rattray, Mental fatigue impairs endurance performance: A physiological explanation, Sports medicine 48 (2018) 2041–2051.

[2] M. M. Lorist, M. A. Boksem, K. R. Ridderinkhof, Impaired cognitive control and reduced cingulate activity during mental fatigue, Cognitive Brain Research 24 (2) (2005) 199–205.

[3] V. Rozand, F. Lebon, P. J. Stapley, C. Papaxanthis, R. Lepers, A prolonged motor imagery session alter imagined and actual movement durations: Potential implications for neurorehabilitation, Behavioural Brain Research SreeTestContent1 297 (2016) 67–75.

[4] V. J. Gawron, J. French, D. Funke, An overview of fatigue, Stress, workload, and fatigue (2000) 581–595.

[5] A. G. Bills, Blocking: A new principle of mental fatigue, The American Journal of Psychology 43 (2) (1931) 230–245.

[6] M. A. Boksem, T. F. Meijman, M. M. Lorist, Effects of mental fatigue on attention: An erp study, Cognitive Brain Research 25 (1) (2005) 107–116. doi:https://doi.org/10.1016/j.cogbrainres.2005.04.011.
URL https://www.sciencedirect.com/science/article/pii/S0926641005001187

[7] S. M. Marcora, W. Staiano, V. Manning, Mental fatigue impairs physical performance in humans, Journal of applied physiology (2009).

[8] R. Holtzer, M. Shuman, J. R. Mahoney, R. Lipton, J. Verghese, Cognitive fatigue defined in the context of attention networks, Aging, Neuropsychology, and Cognition 18 (1) (2010) 108–128.

[9] K. O'Keeffe, S. Hodder, A. Lloyd, A comparison of methods used for inducing mental fatigue in performance research: Individualised, dual-task and short duration cognitive tests are most effective, Ergonomics 63 (1) (2020) 1–12.

[10] J. Fan, A. P. Smith, Effects of occupational fatigue on cognitive performance of staff from a train operating company: A field study, Frontiers in Psychology 11 (2020). doi:10.3389/fpsyg.2020.558520.
URL https://www.frontiersin.org/articles/10.3389/fpsyg.2020.558520

[11] A. P. Vogel, J. Fletcher, P. Maruff, Acoustic analysis of the effects of sustained wakefulness on speech, The Journal of the Acoustical Society of America 128 (6) (2010) 3747–3756.

[12] R. Guidoux, M. Duclos, G. Fleury, P. Lacomme, N. Lamaudière, P.-H. Manenq, L. Paris, L. Ren, S. Rousset, A smartphone-driven methodology for estimating physical activities and energy expenditure in free living conditions, Journal of biomedical informatics 52 (2014) 271–278.

[13] M. Kay, J. Santos, M. Takane, mhealth: New horizons for health through mobile technologies, World Health Organization 64 (7) (2011) 66–71.

[14] A. Z. Antosik-Wójcińska, M. Dominiak, M. Chojnacka, K. Kaczmarek-Majer, K. R. Opara, W. Radziszewska, A. Olwert, Ł. Świkecicki, Smartphone as a monitoring tool for bipolar disorder: a systematic review including data analysis, machine learning algorithms and predictive modelling, International journal of medical informatics 138 (2020) 104131.

[15] M. Kheirkhahan, S. Nair, A. Davoudi, P. Rashidi, A. A. Wanigatunga, D. B. Corbett, T. Mendoza, T. M. Manini, S. Ranka, A smartwatch-based framework for real-time and online assessment and mobility monitoring, Journal of biomedical informatics 89 (2019) 29–40.

[16] Y. Fukazawa, T. Ito, T. Okimura, Y. Yamashita, T. Maeda, J. Ota, Predicting anxiety state using smartphone-based passive sensing, Journal of biomedical informatics 93 (2019) 103151.

[17] V. P. Cornet, R. J. Holden, Systematic review of smartphone-based passive sensing for health and wellbeing, Journal of biomedical informatics 77 (2018) 120–132.

[18] R. Hooda, V. Joshi, M. Shah, A comprehensive review of approaches to detect fatigue using machine learning techniques, Chronic Diseases and Translational Medicine (2021).

[19] S. S. Jasim, A. Hassan, Modern drowsiness detection techniques: A review, International Journal of Electrical and Computer Engineering 12 (3) (2022) 2986.

[20] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, M. Faundez-Zanuy, Evaluation of handwriting kinematics and pressure for differential diagnosis of parkinson's disease, Artificial intelligence in Medicine 67 (2016) 39–46.

[21] M. Palotai, M. Wallack, G. Kujbus, A. Dalnoki, C. Guttmann, Usability of a mobile app for real-time assessment of fatigue and related symptoms in patients with multiple sclerosis: observational study, JMIR mHealth and uHealth 9 (4) (2021) e19564.

[22] S. S. Spahrkäs, A. Looijmans, R. Sanderman, M. Hagedoorn, Beating cancer-related fatigue with the untire mobile app: results from a waiting-list randomized controlled trial, Psycho-Oncology 29 (11) (2020) 1823–1834.

[23] V. W. Chu, S. J. Payne, M. P. Hunter, S. Reynolds, Development of a phone application for assessing fatigue levels in rare disorders: a feasibility and validity study, Journal of Rare Diseases 2 (1) (2023) 17.

[24] E. Valla, A.-J. Toose, S. Nõmm, A. Toomela, Transforming fatigue assessment: Smartphone-based system with digitized motor skill tests, International Journal of Medical Informatics 177 (2023) 105152. doi:https://doi.org/10.1016/j.ijmedinf.2023.105152.
URL https://www.sciencedirect.com/science/article/pii/S1386505623001703

[25] V. Rozand, F. Lebon, C. Papaxanthis, R. Lepers, Effect of mental fatigue on speed–accuracy trade-off, Neuroscience 297 (2015) 219–230.

[26] Y. Le Mansec, B. Pageaux, A. Nordez, S. Dorel, M. Jubeau, Mental fatigue alters the speed and the accuracy of the ball in table tennis, Journal of sports sciences 36 (23) (2018) 2751–2759.

[27] O. Senkiv, S. Nõmm, A. Toomela, Applicability of spiral drawing test for mental fatigue modelling, IFAC-PapersOnLine 51 (34) (2019) 190 – 195, 2nd IFAC Conference on Cyber-Physical and Human Systems CPHS 2018. doi:https://doi.org/10.1016/j.ifacol.2019.01.064.
URL http://www.sciencedirect.com/science/article/pii/S2405896319300679

[28] O. Lippold, The tremor in fatigue, Human muscle fatigue: Physiological mechanisms (1981) 234–248.

[29] F. Budini, L. Labanca, M. Scholz, A. Macaluso, Tremor, finger and hand dexterity and force steadiness, do not change after mental fatigue in healthy humans, Plos one 17 (8) (2022) e0272033.

[30] R. N. Jazar, Advanced Dynamics. Rigid Body, Multibody, and Aerospace Applications, John Wiley & Sons, Inc, 2007.

[31] E. Valla, S. Nõmm, K. Medijainen, P. Taba, A. Toomela, Tremor-related feature engineering for machine learning based parkinson's disease diagnostics, Biomedical Signal Processing and Control 75 (2022) 103551.

[32] S. Nõmm, K. Bardõš, A. Toomela, K. Medijainen, P. Taba, Detailed analysis of the luria's alternating seriestests for parkinson's disease diagnostics, in: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018, pp. 1347–1352. doi:10.1109/ICMLA.2018.00219.

[33] S. Nõmm, A. Toomela, An alternative approach to measure quantity and smoothness of the human limb motions, Estonian Journal of Engineering 19 (4) (2013) 298–308.

[34] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine learning 46 (1) (2002) 389–422.

[35] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning. 2nd Edition, Springer Series in Statistics, Springer,

2002.

[36] E. Valla, S. Nõmm, K. Medijainen, P. Taba, A. Toomela, Tremor-related feature engineering for machine learning based parkinson's disease diagnostics, Biomedical Signal Processing and Control 75 (2022) 103551. doi:https://doi.org/10.1016/j.bspc.2022.103551.
URL https://www.sciencedirect.com/science/article/pii/S1746809422000738

Appendix 1 - Subject data

Table 4: Distribution of participants' demographics.

Feature name	Most common value	Percentage from total values
Height	151-175	51.8%
Weight	61-75	31.3%
Age	18-25	41.6%
Gender	Male	65%
Received education	Higher	32.5%
Daily Work Type	Mental/Physical combined	44%
Dominant Hand	Right hand	91%

Table 5: Demographic data and characteristics of participants.

ID	height	weight	age	gender	education	dailyWork	mainHand
1	<100	<50	<10	Male	None	Physical	RIGHT
2	<100	<50	<10	Female	None	Physical	RIGHT
3	<100	<50	<10	Male	None	Physical	RIGHT
4	<100	<50	<10	Male	None	Physical	RIGHT
5	<100	<50	<10	Male	None	Physical	RIGHT
6	<100	<50	<10	Male	None	Physical	RIGHT
7	<100	50-60	<10	Female	Higher	Mental	RIGHT
8	101-150	<50	10-13	Male	Primary	50/50	RIGHT
9	151-175	<50	10-13	Male	Primary	50/50	RIGHT
10	151-175	50-60	10-13	Male	Primary	50/50	RIGHT
11	151-175	50-60	10-13	Female	Primary	50/50	RIGHT
12	151-175	50-60	10-13	Male	Primary	50/50	RIGHT
13	151-175	50-60	10-13	Female	Primary	50/50	RIGHT
14	151-175	<50	10-13	Male	Primary	50/50	RIGHT
15	151-175	50-60	10-13	Female	None	50/50	RIGHT
16	151-175	<50	10-13	Male	Primary	50/50	LEFT
17	176-185	76-90	10-13	Female	Primary	50/50	RIGHT
18	151-175	<50	10-13	Female	Primary	50/50	RIGHT
19	151-175	<50	10-13	Male	Primary	50/50	RIGHT
20	151-175	<50	10-13	Male	Primary	50/50	LEFT
21	101-150	<50	10-13	Male	Primary	50/50	RIGHT
22	151-175	<50	10-13	Female	Primary	50/50	RIGHT
23	151-175	50-60	10-13	Female	Primary	50/50	RIGHT
24	151-175	<50	10-13	Other	Higher	Physical	RIGHT
25	151-175	50-60	10-13	Female	Primary	50/50	RIGHT
26	101-150	<50	10-13	Female	Primary	50/50	RIGHT
27	151-175	<50	10-13	Female	Basic	50/50	LEFT
28	151-175	50-60	10-13	Female	Primary	50/50	RIGHT
29	101-150	<50	10-13	Female	Primary	50/50	RIGHT
30	151-175	<50	10-13	Female	Primary	Physical	AMBIDEXTROUS
31	101-150	<50	10-13	Female	Primary	50/50	RIGHT
32	176-185	61-75	14-17	Male	Primary	50/50	RIGHT
33	151-175	61-75	14-17	Male	None	50/50	RIGHT
34	176-185	61-75	14-17	Male	Basic	Mental	RIGHT

35	101-150	61-75	14-17	Female	Secondary	50/50	RIGHT
36	151-175	61-75	14-17	Male	Secondary	50/50	RIGHT
37	101-150	<50	14-17	Female	Primary	50/50	RIGHT
38	151-175	50-60	14-17	Female	Primary	50/50	RIGHT
39	151-175	50-60	14-17	Male	Primary	50/50	RIGHT
40	151-175	<50	14-17	Female	Higher	Physical	RIGHT
41	176-185	76-90	14-17	Other	Primary	50/50	RIGHT
42	176-185	61-75	14-17	Male	Basic	50/50	RIGHT
43	151-175	<50	14-17	Female	None	Physical	RIGHT
44	151-175	<50	14-17	Male	Primary	50/50	RIGHT
45	151-175	76-90	14-17	Male	Primary	50/50	RIGHT
46	176-185	61-75	14-17	Male	Basic	50/50	RIGHT
47	151-175	61-75	14-17	Female	Primary	50/50	RIGHT
48	151-175	50-60	14-17	Female	Basic	50/50	RIGHT
49	191-205	61-75	14-17	Male	Basic	Mental	RIGHT
50	151-175	50-60	14-17	Female	Primary	Mental	RIGHT
51	151-175	50-60	14-17	Male	Secondary	50/50	RIGHT
52	151-175	50-60	14-17	Female	Secondary	Mental	RIGHT
53	186-195	76-90	14-17	Male	None	Physical	RIGHT
54	151-175	50-60	14-17	Female	Secondary	50/50	RIGHT
55	151-175	<50	14-17	Female	None	Physical	RIGHT
56	151-175	50-60	14-17	Male	Secondary	Mental	RIGHT
57	186-190	76-90	14-17	Male	Primary	50/50	RIGHT
58	151-175	<50	14-17	Male	Basic	Physical	RIGHT
59	176-185	61-75	14-17	Male	None	50/50	RIGHT
60	151-175	50-60	14-17	Male	Basic	50/50	RIGHT
61	151-175	50-60	14-17	Male	Basic	50/50	RIGHT
62	176-185	76-90	14-17	Male	Basic	50/50	RIGHT
63	191-205	91-105	14-17	Male	Basic	Physical	RIGHT
64	176-185	76-90	14-17	Male	Basic	Physical	RIGHT
65	151-175	<50	14-17	Male	Basic	Physical	RIGHT
66	151-175	50-60	14-17	Female	Primary	Physical	RIGHT
67	176-185	61-75	18-25	Other	Secondary	Mental	LEFT
68	176-185	61-75	18-25	Male	Secondary	50/50	RIGHT
69	151-175	61-75	18-25	Male	Secondary	50/50	RIGHT
70	151-175	61-75	18-25	Male	Higher	Mental	AMBIDEXTROUS
71	176-185	91-105	18-25	Male	Secondary	Mental	RIGHT
72	186-195	76-90	18-25	Male	Basic	Mental	RIGHT
73	176-185	50-60	18-25	Male	Basic	Physical	RIGHT
74	176-185	<50	18-25	Female	Higher	Mental	RIGHT
75	>205	61-75	18-25	Male	Secondary	50/50	RIGHT
76	176-185	61-75	18-25	Male	Secondary	Mental	RIGHT
77	196-205	76-90	18-25	Male	Higher	50/50	RIGHT
78	151-175	61-75	18-25	Female	Primary	Mental	AMBIDEXTROUS
79	176-185	50-60	18-25	Female	Higher	Mental	RIGHT
80	151-175	61-75	18-25	Female	Basic	Physical	RIGHT
81	176-185	61-75	18-25	Male	Primary	50/50	RIGHT
82	176-185	76-90	18-25	Male	Higher	Mental	LEFT
83	151-175	91-105	18-25	Female	Secondary	Mental	RIGHT
84	176-185	76-90	18-25	Male	Secondary	Mental	RIGHT
85	186-190	61-75	18-25	Male	Higher	Mental	RIGHT
86	176-185	61-75	18-25	Male	Higher	Mental	RIGHT

87	176-185	76-90	18-25	Male	Secondary	Physical	RIGHT
88	176-185	61-75	18-25	Male	Basic	Physical	RIGHT
89	151-175	61-75	18-25	Male	Secondary	50/50	AMBIDEXTROUS
90	151-175	61-75	18-25	Male	Basic	Mental	RIGHT
91	191-205	61-75	18-25	Male	Basic	50/50	RIGHT
92	151-175	<50	18-25	Female	Secondary	50/50	RIGHT
93	151-175	50-60	18-25	Male	Secondary	Mental	RIGHT
94	151-175	<50	18-25	Female	Secondary	Mental	RIGHT
95	176-185	61-75	18-25	Male	Secondary	Mental	RIGHT
96	151-175	50-60	18-25	Female	Higher	Mental	RIGHT
97	151-175	<50	18-25	Female	Secondary	50/50	RIGHT
98	176-185	91-105	18-25	Male	Secondary	Mental	RIGHT
99	151-175	<50	18-25	Male	Basic	Mental	RIGHT
100	151-175	50-60	18-25	Male	Higher	50/50	RIGHT
101	151-175	61-75	18-25	Female	Higher	50/50	RIGHT
102	176-185	76-90	18-25	Male	Secondary	Mental	RIGHT
103	151-175	50-60	18-25	Female	Secondary	50/50	RIGHT
104	176-185	76-90	18-25	Male	Secondary	Mental	RIGHT
105	176-185	76-90	18-25	Male	Higher	50/50	RIGHT
106	191-205	91-105	18-25	Male	Secondary	Mental	RIGHT
107	186-195	91-105	18-25	Male	Secondary	Mental	RIGHT
108	186-190	76-90	18-25	Male	Higher	Mental	RIGHT
109	176-185	61-75	18-25	Male	Secondary	50/50	RIGHT
110	151-175	76-90	18-25	Female	Higher	Mental	RIGHT
111	151-175	61-75	18-25	Female	Higher	Mental	LEFT
112	151-175	50-60	18-25	Female	Secondary	Mental	RIGHT
113	191-205	76-90	18-25	Male	Basic	Mental	RIGHT
114	191-205	76-90	18-25	Male	Secondary	Physical	RIGHT
115	191-205	76-90	18-25	Male	Secondary	50/50	RIGHT
116	186-190	76-90	18-25	Male	Higher	Mental	RIGHT
117	151-175	61-75	18-25	Female	Higher	50/50	RIGHT
118	186-190	91-105	18-25	Male	Secondary	50/50	RIGHT
119	151-175	76-90	18-25	Male	Secondary	Physical	RIGHT
120	176-185	76-90	18-25	Male	Basic	50/50	RIGHT
121	176-185	61-75	18-25	Male	Secondary	50/50	RIGHT
122	176-185	50-60	18-25	Female	None	Physical	RIGHT
123	191-205	76-90	18-25	Male	Secondary	Mental	RIGHT
124	151-175	50-60	18-25	Male	Higher	Mental	RIGHT
125	176-185	50-60	18-25	Female	Secondary	Mental	RIGHT
126	151-175	61-75	18-25	Female	Secondary	50/50	RIGHT
127	176-185	91-105	18-25	Male	Primary	Mental	RIGHT
128	186-190	76-90	18-25	Male	Higher	50/50	RIGHT
129	151-175	61-75	18-25	Female	Higher	50/50	LEFT
130	151-175	61-75	18-25	Male	Basic	50/50	RIGHT
131	151-175	61-75	18-25	Male	Higher	50/50	RIGHT
132	151-175	61-75	18-25	Male	Secondary	Physical	RIGHT
133	151-175	61-75	18-25	Male	Higher	Physical	RIGHT
134	151-175	>120	18-25	Female	Secondary	Mental	LEFT
135	151-175	61-75	18-25	Male	Higher	Mental	RIGHT
136	186-195	76-90	26-30	Male	Higher	Mental	RIGHT
137	151-175	61-75	26-30	Male	Higher	Mental	RIGHT
138	151-175	61-75	26-30	Male	Higher	Mental	RIGHT

139	176-185	91-105	26-30	Male	Higher	50/50	RIGHT
140	151-175	76-90	26-30	Male	Higher	Mental	LEFT
141	176-185	76-90	26-30	Male	Higher	50/50	RIGHT
142	151-175	50-60	26-30	Female	Higher	Mental	RIGHT
143	<100	106-120	26-30	Female	Higher	Physical	RIGHT
144	151-175	<50	26-30	Female	Secondary	Mental	RIGHT
145	151-175	61-75	26-30	Female	Higher	Mental	LEFT
146	176-185	61-75	31-35	Male	Higher	50/50	RIGHT
147	191-205	91-105	31-35	Male	Higher	Mental	LEFT
148	151-175	61-75	31-35	Female	Higher	Mental	RIGHT
149	151-175	91-105	31-35	Male	Higher	Mental	RIGHT
150	151-175	61-75	31-35	Male	Higher	Mental	RIGHT
151	176-185	61-75	31-35	Male	Higher	50/50	RIGHT
152	176-185	76-90	36-45	Male	Higher	50/50	RIGHT
153	186-195	76-90	36-45	Male	None	Mental	RIGHT
154	186-195	>120	36-45	Male	Higher	Mental	RIGHT
155	151-175	61-75	36-45	Male	Higher	Mental	RIGHT
156	151-175	61-75	36-45	Male	Higher	Mental	RIGHT
157	176-185	91-105	36-45	Male	Higher	Mental	RIGHT
158	176-185	76-90	36-45	Male	Higher	Mental	RIGHT
159	151-175	61-75	36-45	Male	Higher	Mental	RIGHT
160	151-175	61-75	36-45	Male	Higher	Mental	RIGHT
161	176-185	76-90	36-45	Male	Higher	Mental	RIGHT
162	151-175	61-75	36-45	Male	Higher	50/50	RIGHT
163	151-175	76-90	46-55	Male	Higher	Mental	RIGHT
164	186-195	91-105	46-55	Male	Higher	Mental	RIGHT
165	151-175	61-75	46-55	Female	Higher	Mental	RIGHT
166	151-175	61-75	56-65	Female	Higher	Mental	RIGHT

Appendix 2 - Fatigue Test Terms of Use

RESEARCH PARTICIPATION INFORMATION SHEET

Welcome to the Fatigue Test Application Terms of Use agreement. For purposes of this agreement, “App” refers to our mobile application in which users are asked to complete the questionnaire and three fine-motor skill related tests. The terms “we,” “us,” and “our” refer to the Fatigue Test App. “You” refers to you, as a participant in this research.

The following Terms of Use apply when you use the App on your mobile device.

Please review the following terms carefully and signify your agreement to these Terms of Use at the bottom by clicking Agree. If you do not agree to be bound by these Terms of Use in their entirety, you may not access or use the App.

I - INTRODUCTION

This research is conducted by researchers at the Tallinn University of Technology Department of Software Science. The main scope of the study is to develop a framework for human motor function and cognitive impairment analysis. Movement and neurological disorders pose a significant burden on the healthcare system.

Our goal is to provide decision support tools to help clinicians with data collection, diagnostics, and treatment processes. The more data we collect, the more accurate and reliable applications we can develop. We are thankful for any contribution. Participation is entirely voluntary, and you can withdraw your data anytime.

II - INFORMATION WE COLLECT

We collect “Non-Personal Information”. Non-Personal Information includes information that cannot be used to personally identify you, such as anonymous usage data, and general demographic information we may collect. The collected data is specified below.

1. Data that we collect through the questionnaire:

    a. gender
    b. age (interval)
    c. height (interval)
    d. weight (interval)
    e. education level

f. type of main daily activities (mental, physical, 50/50)
g. dominant hand
h. interest level in the last task with which the user was engaged with (scale 1-10)
i. mental demand level in the last task with which the user was engaged with (scale 1-10)
j. the current perceived state of anxiety (scale 1-10)
k. the current perceived state of fatigue (scale 1-10)
l. the number of hours slept the previous night (scale 0-12)
m. the number of hours spent on a physical activity (scale 0-12)
n. the number of hours spent on a mental activity (scale 0-12)

2. Data that we collect through tests:

a. reaction time
b. test duration
c. error rate
d. kinematic and dynamic parameters:
    i. screen coordinates
    ii. time
e. axial derivations recorded by the accelerometer

III. HOW WE USE AND SHARE INFORMATION

The collected data will be used as research data by the TalTech University the Department of Software Science to further the knowledge around cognitive impairment and human motor function analysis.

IV. HOW WE STORE AND PROTECT INFORMATION

We further protect your information from potential security breaches by implementing encrypted data transfer over a secure socket layer connection and storing it in a secured database. The data will become accessible over an off-site application programming interface by authorized users. However, these measures do not guarantee that your information will not be accessed, disclosed, altered, or destroyed by a breach of such firewalls and secure server software. By using our App, you acknowledge that you understand and agree to assume these risks.

We keep information for as long as we need it for our research. We decide how long we need information on a case-by-case basis.

V. YOUR RIGHTS REGARDING THE USE OF YOUR DATA

You have the right to erasure. You can request for your data to be deleted from our databases at any time.

VI. CONTACT US

If you have any technical questions and concerns regarding the practices of this App, please contact us by sending an email to elli.valla@taltech.ee.
Last Updated: This Information Sheet was last updated on 30.10.2023.

YOU ACKNOWLEDGE THAT YOU HAVE READ THIS RESEARCH PARTICIPATION INFORMATION SHEET , UNDERSTAND THE TERMS OF USE, AND WILL BE BOUND BY THESE TERMS AND CONDITIONS. YOU FURTHER ACKNOWLEDGE THAT THESE TERMS OF USE REPRESENT THE COMPLETE AND EXCLUSIVE STATEMENT OF THE AGREEMENT BETWEEN US AND THAT IT SUPERSEDES ANY PROPOSAL OR PRIOR AGREEMENT ORAL OR WRITTEN, AND ANY OTHER COMMUNICATIONS BETWEEN US RELATING TO THE SUBJECT MATTER OF THIS AGREEMENT.

Appendix 3 - Fatigue Test iOS Application

https://apps.apple.com/ee/app/fatigue-test-taltech/id6449683047

Supplementary material

Table 19: Boolean search queries

Database	Search Query	Results
PubMed	((parkinson*[Title] OR fatigue*[Title]) AND (digital tool*[Title/Abstract] OR digital test*[Title/Abstract] OR smartphone*[Title/Abstract] OR tablet*[Title/Abstract] OR mobile application*[Title/Abstract] OR handwriting[Title/Abstract] OR stylus[Title/Abstract] OR touchscreen[Title/Abstract] OR digital assessment[Title/Abstract] OR wearable*[Title/Abstract]) AND (fine motor[Title/Abstract] OR handwriting[Title/Abstract] OR drawing[Title/Abstract] OR "motor function"[Title/Abstract] OR kinematic*[Title/Abstract] OR tremor*[Title/Abstract] OR "motor control"[Title/Abstract] OR dexterity[Title/Abstract] OR "movement disorder"[Title/Abstract]) AND (artificial intelligence[Title/Abstract] OR machine learning[Title/Abstract] OR deep learning[Title/Abstract] OR "feature extraction"[Title/Abstract] OR classification[Title/Abstract] OR diagnostics[Title/Abstract] OR detection[Title/Abstract] OR monitoring[Title/Abstract] OR "pattern recognition"[Title/Abstract] OR predictive[Title/Abstract])) NOT (gait[Title/Abstract] OR speech[Title/Abstract] OR voice[Title/Abstract] OR vocal*[Title/Abstract] OR MRI[Title/Abstract] OR EEG[Title/Abstract]))	155
Google Scholar	(parkinson* OR fatigue*) AND ("fine motor" OR handwriting OR drawing OR "motor function" OR tremor* OR kinematic*) AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "feature extraction" OR classification) AND ("digital tool*" OR smartphone* OR tablet* OR "mobile application*" OR wearable* OR touchscreen*) -gait -speech -voice -vocal -MRI -EEG -"multiple sclerosis" -"alzheimer*" -review -survey	112
Scopus	(TITLE (parkinson* OR fatigue*) AND TITLE-ABS-KEY ("digital tool*" OR "digital test*" OR "smartphone*" OR "tablet*" OR "mobile application*" OR "handwriting" OR stylus OR "touchscreen" OR "digital assessment" OR "wearable") AND TITLE-ABS-KEY ("fine motor" OR handwriting OR drawing OR "motor function" OR "kinematic*" OR "motor performance" OR "motor impairment" OR tremor* OR "motor control" OR dexterity OR "movement disorder") AND TITLE-ABS-KEY ("artificial intelligence" OR "machine learning" OR "deep learning" OR "feature extraction" OR classification OR diagnostics OR detection OR monitoring OR "pattern recognition" OR predictive) AND NOT (gait OR review OR speech OR voice OR vocal* OR mri OR eeg))	73
Web of Science	TS=(parkinson* OR fatigue*) AND TS=("fine motor" OR handwriting OR drawing OR "motor function" OR tremor* OR kinematic*) AND TS=("artificial intelligence" OR "machine learning" OR "deep learning" OR "feature extraction" OR classification) AND TS=("digital tool*" OR smartphone* OR tablet* OR "mobile application*" OR wearable* OR touchscreen*) NOT TS=(gait OR speech OR voice OR vocal* OR MRI OR EEG OR "multiple sclerosis" OR alzheimer* OR review OR survey)	163

Table 20: Inclusion and exclusion criteria

Criteria	Specified Criteria
Inclusion	• Studies that analysed motor functions (e.g., gait analysis, fine motor skills) in the context of diagnosing PD. Classification of PD from healthy controls (HC). • Peer-reviewed articles, conference proceedings, reports, theses, and dissertations. • Studies written in English.
Exclusion	• Studies that did not apply machine learning methods. • Studies that did not use motor function data for analysis (e.g., speech data, brain scan images, EEG signals) unless they are combined with motor data for diagnosing neurological disorders. • Studies focused solely on non-diagnostic applications such as symptom classification, treatment response, disease progression, or severity assessment without a primary focus on diagnosis. • Studies that used non-human subjects (e.g., animal models). • Review articles, including literature reviews, scoping reviews, and overviews. • Studies not written in English. • Non-peer-reviewed articles.

Table 21: Comprehensive overview of PD studies

Category	Sub-category	Description	References
Data acquisition technology	Pen and paper	Traditional paper-based tests including handwriting samples, drawing tasks, or other manual activities that are later digitised and analysed for motion or tremor assessment.	[75][33][76][32][77][78][19][79]
	Wearable sensors	Includes smartwatches, trackers, IMUs, accelerometers, gyroscopes, etc for monitoring motion or tremors.	[80][81][14][82][83][84]
	Non-wearable sensors	External devices such as digital/graphic tablets, an intelligent sensory pen integrated with accelerometers and gyroscopes, keyboard and mouse used in structured web-based tests to assess motor functions like typing and hand movement, or handwriting datasets for motion tracking.	[28][22][85][25][23][26][18][86][87][88][89] [90][91][92][93][14][94][18][29][16][95][96] [97][98][15][24][99][100][101][36][102][103]
	Smart devices	Smartphone and tablet based systems with integrated apps for monitoring.	[104][31][30][105][106][107][20][21][27] [14][108][109][110][111][112][113][114]
Machine learning techniques	Traditional machine learning	Algorithms such as SVM, Decision Trees, Random Forests, and k-means clustering.	[28][30][105][106][25][26][18][75][86][87][33] [76][32][80][81][107][20][21][19][27][91][92] [14][82][108][94][109][18][110][29][16][95][111] [96][97][98][24][99][83][100][36][102]
	Deep learning	Neural network-based techniques such as CNNs, RNNs, and BiGRUs for advanced pattern recognition.	[104][31][22][85][23][33][88][77][78][80][107] [89][90][21][93][14][115][16][96][15][83][112] [113][114][84][101][79][103]
Data types	Time series	Sequential data captured from sensors during tasks like spiral drawing or motor tests. Pose data extracted from camera recordings.	[104][31][30][85][105][106][25][23][26][18][86] [33][87][88][33][76][32][80][81][107] [20][89][21][27][91][92][93][14][82][108][115] [94][109][18][110][111][97][98][24][99][83][112][113] [114][84][100][101][79]
	Image data	Static images such as spirals, lines, or handwriting samples analysed for kinematic/geometric properties.	[28][104][22][85][75][88][86][32][77][78][89] [90][19][14][96][15][102]
	Hybrid data	Combination of different types for multi-modal analysis. Combination of on-surface (contact) and in-air (no-contact) handwriting movements.	[33][107][14][29][16][95][99][100][101][36][103]
Exctracted features	Kinematic and dynamic	Metrics such as velocity, angular velocity, pressure difference, and acceleration derived from handwriting data.	[104][85][105][106][25][23][26][18][86][33][87] [88][76][32][80][81][107][20][89][21][19] [27][91][92][93][14][82][115][94][109][18][29] [16][95][111][97][98][24][99][83][114][84][101] [36][103]
	Geometric	Variability in spiral tracing and character formation patterns during writing tasks. Variability in writing size, spiral precision, and other shape-specific parameters. Analysing drawing shape, or handwriting dynamics and tracing metrics like deviation from centerline and accuracy in predefined patterns.	[28][75][86][33][86][32][20][19][92][14][109] [29][16][96][97][15][100][99][101][102][79][103]
	User interaction patterns	Natural smartdevice interactions analysed for fine motor impairments. Some examples may include keystroke hold time, flight time, Metrics like typing response time, accuracy, and false presses during keyboard tests.	[110][112][113][114][100][103]
	Advanced metrics	Includes features such as entropy measures (e.g., Shannon and Renyi entropy), energy metrics (e.g., Taeger-Kaiser energy), and non-linear dynamics features like complexity and chaos in handwriting signals. Other examples may include Histogram of Oriented Gradients (HOG) for capturing detailed spatial relationships in the drawings, Spectral analysis features like Discrete Wavelet Transform (DWT) and Fast Fourier Transform (FFT) applied to handwriting dynamics. High-dimensional feature vectors automatically extracted by CNN architectures.	[104][31][30][22][85][25][23][88][77][78][80] [81][20][89][90][21][19][27][91][93][14][82] [108][115][94][18][29][96][98][97][36][102]

Curriculum Vitae

1. Personal data

Name	Elli Valla
Date and place of birth	11 February 1987, Georgia
Nationality	Estonian

2. Contact information

Address	Tallinn University of Technology, School of Information Technologies Department of Software Science, Akadeemia tee 15a, 12618 Tallinn, Estonia
Phone	+372 58058878
E-mail	elli.valla@taltech.ee

3. Education

2020–…	Tallinn University of Technology, School of Information Technologies, PhD studies
2011–2014	Tallinn University of Technology, School of Science, Engineering Physics and Applied Mathematics, MSc *cum laude*
2007–2011	Tallinn University of Technology, School of Science, Engineering Physics, BSc

4. Language competence

Estonian	native
English	fluent
Russian	native

5. Professional employment

2020– …	Tallinn University of Technology, Department of Software Science, early stage researcher

6. Computer skills

- **Operating systems:** macOS, Linux, Windows
- **Office Suites:**
  - **Microsoft Office:** Word, Excel, PowerPoint, and Outlook.
  - **Apple iWork:** Pages, Numbers, and Keynote.
  - **Google Workspace:** Docs, Sheets, Slides, and Forms.
- **Document preparation:** Advanced proficiency in LaTeX for high-quality typesetting.
- **Programming languages:** Proficient in Python and MATLAB; familiar with basic concepts of R and Java.

- **Scientific computing packages:**
  - **Python libraries:** Expert in NumPy, pandas, scikit-learn, TensorFlow, PyTorch, among many others.
- **Specialised skills:**
  - **Pose estimation algorithms:** Knowledgeable in computer vision techniques for human pose estimation.
  - **Amazon SageMaker:** Familiar with deploying machine learning models and managing ML workflows.
  - **High-performance computing:** Experienced in utilising virtual computing machines for large-scale computations.
  - **Generative artificial intelligence:** Practical experience with closed-source (e.g., ChatGPT, Gemini, Claude) and open-source (e.g., LLaMA, Mistral, DeepSeek) language models for enhancing productivity in data exploration, educational material creation, technical writing, and software development.

7. Honours and awards

- 2024, **Ustus Agur's scholarship** The Ustus Aguri scholarship, awarded by the Education and Youth Board of Estonia, recognised my contribution to the development of Estonian society through the use of artificial intelligence and smart devices to analyse fine motor skill kinematics for early diagnosis of Parkinson's disease, while scaling healthcare delivery, enhancing accessibility, and enabling potential cross-border services.
- 2024, **Andres Keevallik's scholarship** The Andres Keevallik's scholarship, awarded by TalTech Development Fund, recognised my doctoral work, which focuses on AI-based diagnostics for Parkinson's disease.
- 2014, **Silver Decoration "Fidelis Studiosus"** Tallinn University of Technology (TalTech). Awarded for exemplary service and significant contributions to the TalTech student body. Recognised for founding the TalTech Cheerleaders organisation and serving as a managing member.

8. Defended theses

- 2014, "Inverse problems for a linear model of microstructure with multiple scales", MSc, supervisor Prof. Jaan Janno, Tallinn University of Technology, Institute of Cybernetics

9. Supervised dissertations

1. Erik Dzotsenidze, Master's Degree, 2022, (sup) Elli Valla; Sven Nõmm, Generative Adversarial Networks as a Data Augmentation Tool for CNN-based PD Diagnostics, Tallinn University of Technology School of Information Technologies, Department of Software Science
2. Henry Laur, Master's Degree, 2022, (sup) Sven Nõmm; Elli Valla, Automated Segmentation and Semantic Analysis of Writing and Drawing Tests for PD Diagnostics, Tallinn University of Technology School of Information Technologies, Department of Software Science

3. Ain-Joonas Toose, Master's Degree, 2023, (sup) Elli Valla; Sven Nõmm, A Novel Data Collection Solution Through Digital Fine Motor Skill Assessment for Fatigue Visualization and Analytics, Tallinn University of Technology School of Information Technologies, Department of Software Science

4. Lilian Väli, Master's Degree, 2024, (sup) Elli Valla; Sven Nõmm, Exploring the Efficacy of Smartphone Sensors in Mental Fatigue Detection: A Machine Learning Approach to Analysing Fine Motor Skills, Tallinn University of Technology, School of Information Technologies, Department of Software Science

10. Field of research

- ETIS RESEARCH FIELD: 4. Natural Sciences and Engineering; 4.6. Computer Sciences
- CERCS RESEARCH FIELD: P176 Artificial intelligence

11. Scientific work

1. Elli Valla, Sven Nõmm, Kadri Medijainen, Pille Taba, and Aaro Toomela. Tremor-related feature engineering for machine learning based Parkinson's disease diagnostics. *Biomedical Signal Processing and Control*, 75:103551, 2022

2. Erik Dzotsenidze, Elli Valla, Sven Nõmm, Kadri Medijainen, Pille Taba, and Aaro Toomela. Generative adversarial networks as a data augmentation tool for CNN-based Parkinson's disease diagnostics. volume 55, pages 108–113. Elsevier, 2022

3. Vassili Gorbatsov, Elli Valla, Sven Nõmm, Kadri Medijainen, Pille Taba, and Aaro Toomela. Machine learning based analysis of the upper limb freezing during handwriting in Parkinson's disease patients. volume 55, pages 91–95. Elsevier, 2022

4. Elli Valla, Henry Laur, Sven Nõmm, Kadri Medijainen, Pille Taba, and Aaro Toomela. Deep learning based segmentation of Luria's alternating series test to support diagnostics of Parkinson's disease. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1066–1071. IEEE, 2023

5. Elli Valla, Ain-Joonas Toose, Sven Nõmm, and Aaro Toomela. Transforming fatigue assessment: Smartphone-based system with digitized motor skill tests. *International journal of medical informatics*, 177:105152, 2023

6. Xuechao Wang, Junqing Huang, Marianna Chatzakou, Sven Nõmm, Elli Valla, Kadri Medijainen, Pille Taba, Aaro Toomela, and Michael Ruzhansky. Comparison of one-two-and three-dimensional CNN models for drawing-test-based diagnostics of the Parkinson's disease. *Biomedical Signal Processing and Control*, 87:105436, 2024

7. Elli Valla, Gert Kanter, Sven Nõmm, Anton Osvald Kuusk, Peeter Maran, Karl Mihkel Seenmaa, Killu Mägi, and Aaro Toomela. Enhancing cerebral palsy gait analysis with 3D computer vision: A dual-camera approach. In *2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 1352–1357, 2024

Elulookirjeldus

1. Isikuandmed

Nimi	Elli Valla
Sünniaeg ja -koht	11.02.1987, Gruusia
Kodakondsus	Eesti

2. Kontaktandmed

Aadress	Tallinna Tehnikaülikool, Tarkvarateaduse Instituut, Akadeemia tee 15a, 19086 Tallinn, Estonia
Telefon	+372 58058878
E-post	elli.valla@taltech.ee

3. Haridus

2020–…	Tallinna Tehnikaülikool, Infotehnoloogia teaduskond, doktoriõpe
2007-2011	Tallinna Tehnikaülikool, Matemaatika- ja loodusteaduste teaduskond, Tehniline füüsika ja rakendusmatemaatika, MSc *cum laude*
2007–2011	Tallinna Tehnikaülikool, Matemaatika- ja loodusteaduste teaduskond, Tehniline füüsika, BSc

4. Keelteoskus

eesti keel	emakeel
inglise keel	kõrgtase
vene keel	emakeel

5. Teenistuskäik

2020– …	Tallinna Tehnikaülikool, Tarkvarateaduse instituut nooremteadur

6. Arvutioskused

- **Operatsioonisüsteemid:** macOS, Linux, Windows
- **Kontoritarkvara:**
  - **Microsoft Office:** Word, Excel, PowerPoint ja Outlook.
  - **Apple iWork:** Pages, Numbers ja Keynote.
  - **Google Workspace (endine G-Suite):** Docs, Sheets, Slides ja Forms.
- **Dokumendi ettevalmistus:** Täiustatud oskus LaTeX'is kvaliteetse trükikujunduse jaoks.
- **Programmeerimiskeeled:** Edasijõudnud Pythoni ja MATLABiga; tuttav R-i ja Java põhikonseptsioonidega.
- **Teadustarkvara paketid:**
  - **Pythoni teegid:** Ekspert NumPy, pandas, scikit-learn, TensorFlow, PyTorch, jt teekides.

- **Erialased oskused:**
  - **Poosihindamise algoritmid:** Teadmised inimpoosi hinnangute masinnägemise tehnikatest.
  - **Amazon SageMaker:** Teadmised masinõppe mudelite juurutamise ja ML töövoogude haldamisega.
  - **Kõrgjõudlusega andmetöötlus:** Kogemus virtuaalsete arvutusmasinate kasutamisega suuremahuliste arvutuste jaoks.
  - **Generatiivne tehisintellekt:** Praktiline kogemus kinnise (nt ChatGPT, Gemini, Claude) ja avatud lähtekoodiga (nt LLaMA, Mistral, DeepSeek) keelemudelite kasutamisel andmeanalüüsi toetamisel, õppevahendite loomisel, tehnilises kirjutamises ja tarkvaraarenduses.

7. Autasud

- 2024, **Ustus Aguri stipendium** Ustus Aguri stipendium, mis anti välja Haridus- ja Noorteameti poolt, tunnustas minu panust Eesti ühiskonna arengusse, kasutades tehisintellekti ja nutiseadmeid, et analüüsida inimese peenmotoorika kinemaatikat Parkinsoni tõve varaseks diagnoosimiseks, samal ajal tervishoiuteenuste skaleerimise, ligipääsetavuse suurendamise ja potentsiaalsete piiriüleste teenuste võimaldamisega.
- 2024, **Andres Keevalliku stipendium** Andres Keevalliku stipendium, mis anti välja TalTechi Arengufondi poolt, tunnustas minu doktoritööd, mis keskendub tehisintellekti-põhisele Parkinsoni tõve diagnostikale.
- 2014, **Hõbedane teenetemärk "Fidelis Studiosus"** Tallinna Tehnikaülikool (TalTech). Antud eeskujuka teenistuse ja olulise panuse eest TalTechi üliõpilaskonda. Tunnustatud TalTechi Cheerleaders organisatsiooni asutamise ja juhatuse liikmena teenitud panuse eest.

8. Kaitstud lõputööd

- 2014, "Pöördülesanded lineaarse multiskaalalise mikrostruktuuriga mudeli jaoks", MSc, juhendaja professor Jaan Janno, Tallinna Tehnikaülikool, Küberneetika Instituut, MSc, juhendaja Prof. Jaan Jano, Tallinna Tehnikaülikool, Küberneetika Instituut

9. Teadustöö põhisuunad

- ETIS VALDKOND: 4. Loodusteadused ja tehnika; 4.6. Arvutiteadused
- CERCS VALDKOND: P176 Tehisintellekt

10. Teadustegevus

Teadusartiklite, konverentsiteeside ja konverentsiettekannete loetelu on toodud ingliskeelse elulookirjelduse juures.

ISSN 2585-6901 (PDF)
ISBN 978-9916-80-306-6 (PDF)