TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Lilian Väli 203984IAPM

# EXPLORING THE EFFICACY OF SMARTPHONE SENSORS IN MENTAL FATIGUE DETECTION: A MACHINE LEARNING APPROACH TO ANALYSING FINE MOTOR SKILLS

Master's Thesis

Supervisor: Elli Valla
MSc

Co-supervisor: Sven Nõmm
PhD

Tallinn 2024

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Lilian Väli 203984IAPM

# VAIMSE VÄSIMUSE HINDAMINE NUTITELEFONI SENSORITE JA PEENMOTOORSETE OSKUSTE MÕÕTMISE ABIL KASUTADES MASINÕPPE LÄHENEMIST

Magistritöö

Juhendaja: Elli Valla
MSc

Kaasjuhendaja: Sven Nõmm
PhD

Tallinn 2024

# Author's Declaration of Originality

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

Author: Lilian Väli

05.01.2024

# Acknowledgements

The author extends sincere gratitude and appreciation to several individuals who played pivotal roles in the successful completion of the master's thesis.

First and foremost, the author acknowledges the main supervisor, Elli Valla, for providing unwavering support, guidance, and dedication throughout the research journey. The active involvement of Elli Valla during data collection greatly enriched the quality of the research.

The author also expresses appreciation for the contributions of the other supervisor, Sven Nõmm, whose expert guidance, constructive feedback, and continuous mentorship were instrumental in shaping the direction of the research. Sven Nõmm's insights and oversight ensured that the author remained on the right path while maintaining the highest standards of academic rigor.

Furthermore, the author extends deep appreciation to Professor Aaro Toomela, a distinguished psychology professor at Tallinn University. Professor Toomela's significant contributions to the research group, including his profound understanding of the subject matter and ability to inspire innovative ideas, greatly enhanced the quality and depth of the research. The author is also grateful for the invaluable feedback and guidance provided by Professor Toomela throughout the research process.

The author acknowledges the indispensable support, expertise, and encouragement received from these individuals, emphasizing their crucial roles in the successful completion of the master's thesis.

# Abstract

This thesis explores the development and application of a smartphone-based system for detecting mental fatigue. Tailored for both iOS and Android platforms, the system includes a suite of tests to evaluate fine motor skills, accompanied by a detailed questionnaire to enrich the collected data. The primary aim is to use this data for training machine learning models capable of determining mental fatigue in users.

A two-phase data collection approach was employed, where participants interacted with the application before and after activities likely to induce cognitive fatigue. This methodology was used for capturing changes in fine motor skills, which are indicative of mental fatigue. In total, 166 unique devices were involved in completing the tests using the developed mobile application, resulting in 347 sessions recorded.

The core of this research lies in the application of various machine learning algorithms, rigorously evaluated through nested cross-validation techniques. The analysis led to an essential finding: self-reported tiredness and mental work hours are reliable indicators for labelling mental fatigue. The models developed in this study achieved high performance, with the best-performing model reaching scores in the higher eighties range. This level of accuracy highlights the potential efficacy of integrating subjective assessments with objective performance metrics in fatigue detection.

The implications of this research are broad, offering potential applications in workplace safety, educational settings, and healthcare. Moreover, the comprehensive dataset generated provides a valuable resource for further exploration into cognitive and motor functions.

The thesis is written in English and is 65 pages long, including 8 chapters, 38 figures and 9 tables.

# Annotatsioon

## Vaimse Väsimuse Hindamine Nutitelefoni Sensorite ja Peenmotoorsete Oskuste Mõõtmise Abil Kasutades Masinõppe Lähenemist

Käesolevas lõputöös uuritakse nutitelefonil põhineva süsteemi arendamist ja rakendamist vaimse väsimuse tuvastamiseks. Süsteem on kohandatud nii iOSi kui ka Androidi platvormidele ja sisaldab peenmotoorika hindamise testide komplekti, millele on lisatud üksikasjalik küsimustik kogutud andmete rikastamiseks. Peamine eesmärk on kasutada neid andmeid masinõppe mudelite treenimiseks, mis võimaldavad määrata kasutajate vaimset väsimust.

Kasutati kaheetapilist andmekogumist, kus osalejad läbisid rakenduses esitatud ülesandeid enne ja pärast tegevusi, mis tõenäoliselt põhjustavad kognitiivset väsimust. Metoodikat kasutati peenmotoorika muutuste registreerimiseks, mis viitavad vaimsele väsimusele. Kokku osales testide läbiviimisel 166 unikaalset seadet, kasutades välja töötatud mobiilirakendust, mille tulemusel registreeriti 347 seanssi.

Uurimistöö tuum seisneb erinevate masinõppe algoritmide rakendamises, mida hinnati rangelt ristvalideerimise *(cross-validation)* meetodite abil. Analüüsi tulemusena saadi oluline järeldus: raporteeritud väsimus ja vaimse töö tegemise aeg tundides on usaldusväärsed näitajad vaimse väsimuse märgistamiseks. Selles uuringus välja töötatud mudelid saavutasid kõrgeid tulemusi, parima mudeli headuse parameetrid jäid kaheksakümnendate kõrgemasse vahemikku. Selline täpsuse tase rõhutab subjektiivsete hinnangute ja objektiivsete töövõime näitajate integreerimise võimalikku tõhusust väsimuse tuvastamisel.

Selle uuringu mõju on laialdane, avades uusi võimalusi tööohutuse, hariduse ja tervishoiu valdkondades. Peale selle loob laiaulatuslik andmekogu olulise aluse kognitiivsete ja motoorsete funktsioonide süvendatud uurimiseks.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 65 leheküljel, 8 peatükki, 38 joonist, 9 tabelit.

# List of Abbreviations and Terms

| | |
|---|---|
| AdaBoost | Adaptive Boosting |
| API | Application Programming Interface |
| EEG | Electroencephalogram |
| K-NN | K-Nearest Neighbour |
| ML | Machine Learning |
| PCA | Principal Component Analysis |
| PERCLOS | Percentage of Eye Closure |
| RFE | Recurrent Feature Elimination |
| SVM | Support Vector Machine |
| SFM | SelectFromModel |
| PD | Parkinson's disease |
| MS | Multiple Sclerosis |
| BTHS | Barth Syndrome |
| TalTech | Tallinn University of Technology |
| DBScan | Density-Based Spatial Clustering of Applications with Noise |
| EM | Expectation-Maximization algorithm |
| GMM | Gaussian Mixture Model |
| WCSS | Within-cluster Sum of Squares |
| PCA | Principal Component Analysis |
| LogReg | Logistic Regression |
| RF | Random Forest |
| DT | Decision Tree |
| SQL | Structured Query Language |

# Table of Contents

# List of Figures

# List of Tables

# 1. Introduction

The concept of fatigue, complex and often intangible, is defined in a variety of ways.

In research, fatigue is recognised as a complex and multidimensional concept, with varying definitions across various domains and studies. An article by S. Skau, K. Sundberg, and H. Kuhn [1] aimed to synthesise a set of unifying definitions that are useful in all areas of fatigue research. Their method used four desiderata: broadness, precision, neutrality, and phenomenon-focus which led to multiple definitions. One definition emphasises the role of fatigue in causing a decrement in performance improvement during a task. Another posits that fatigue results in a mismatch between the effort expended and the actual performance achieved. Additionally, the sensation of fatigue is characterised as the feeling of needing rest, underscoring the subjective experience of this state.

The nuanced effects of fatigue on human cognitive and physical performance have been the subject of extensive research. A series of studies have shed light on this phenomenon, revealing the diverse ways in which fatigue manifests and impacts efficiency.

A seminar study on the effects of prolonged visual attention tasks over 3 hours demonstrated marked increases in reaction times and errors, highlighting fatigue's detrimental impact on tasks requiring sustained concentration [2]. This finding is complemented by another investigation, where mental fatigue from cognitive tasks was shown to significantly impair physical performance, as evidenced in a cycling task following a period of intense cognitive engagement [3].

Further exploring the realm of attention, a substantial study with 228 participants revealed that mental fatigue progressively diminishes the capacity to maintain focus [4]. This was evidenced through an arrow direction reporting task, underscoring the pervasive influence of fatigue on attentional control [4].

Delving deeper into cognitive task performance, another study examined the fatigue effects of various cognitive tasks [5]. The first task was distinguishing odd and even numbers paired with recognising letters from the screen. These were alternated in sequence. The second task was watching a documentary for a total of 90 minutes. The third task was pressing the space bar based on relevant visuals. The last task was the same as the first one, but the presentation time of the numbers and letters was individualised. The first, third,

and last tasks all induced mental fatigue, but the last one produced the highest levels of mental fatigue, suggesting a link between task complexity and fatigue severity.

In a real-world context, a study conducted in the United Kingdom involving work-related cognitive tasks found that post-work cognitive performance significantly declined [6]. This was evidenced by slower response times and reduced accuracy in tasks performed after a workday, reinforcing the idea that occupational fatigue substantially impairs cognitive functioning [6].

The relationship between sustained mental effort and fatigue was further examined through a mental arithmetic study with 20 participants [7]. Here, five EEG signals were measured and statistical analysis was carried out on the results among different brain regions. The signals showed mental fatigue after performing the task which proved that mental arithmetic problems can successfully be used to induce mental fatigue.

Lastly, a study done with 18 participants in 2010 analysed the effect of fatigue on speech over 24 hours [8]. Every 4 hours, speech samples were acquired from the subjects. The subjects performed speech tasks, a sustained vowel, and also read a passage. Notable changes in speech patterns, including slower speech and increased pitch variation, were observed, offering a unique perspective on the physiological dimensions of fatigue.

## 1.1 Related Works

Recent studies have made significant strides in utilising machine learning (ML) and smartphones to measure fatigue, offering innovative and accessible methods for fatigue detection. This section presents the prominent studies and papers done on this topic.

A comprehensive analysis of 67 articles on fatigue detection using ML and mathematical modelling highlighted the effectiveness of various approaches [9]. While EEG-based methods and facial behaviour analysis showed promising results, the cost and complexity of EEG were noted as limitations. In contrast, camera-based detection offers a less expensive and real-time alternative. The study concluded that a combination of biological and physical features yields the highest accuracy in fatigue detection.

Another study compares 48 papers on drowsiness detection techniques using ML to arrive at a recommendation for a strategy based on the findings [10]. The papers were divided into three main categories depending on the features analysed: vehicle features, behavioural features, and physiological features. Vehicular features are commonly extracted from the steering wheel. Physiological features are for example heart rate, pulse, and body tem-

perature. Behavioural features usually depend on image processing to detect movements. Using vehicular features the highest accuracy of 98.1% was achieved with the Supporting Vector Machine (SVM) classifier. Using physiological features the highest accuracy of 98.6% was found using the k-nearest neighbour (K-NN) classifier with synchronisation likelihood and minimum spanning tree together as a feature selection algorithm. Using behavioural features the highest accuracy of 98.0% was achieved through using an SVM classifier with mouth region features. A hybrid feature-based technique was also taken into consideration and with this, the accuracy of 98.6% was achieved through calculating PERCLOS and using voice as a feature with the SVM classifier. The author concludes that physiological features give better results than the other two, but recommends using the hybrid technique. Although they show high performance, the experimental setting is not attainable in most practices. Monitoring the movement of a steering wheel and additionally the physiological features is very expensive and resource heavy.

In the context of Parkinson's disease (PD), a study on 37 PD patients and 38 healthy individuals used handwriting tasks to diagnose PD [11]. ML techniques like SVM, AdaBoost, and K-NN algorithms were employed to analyse kinematic and pressure features in handwriting, with the SVM showing the best results. This study underscores the potential of using digitised motorised skill tests in neurological disorder diagnostics.

In a study on Multiple Sclerosis (MS), the researchers developed a mobile app to assess fatigue and mood symptoms to improve understanding of MS-related fatigue [12]. The app enabled patients to frequently record their fatigue levels, depression, anxiety, and pain using visual analogue scales, supplemented by one-time questionnaires. This method facilitated real-time symptom monitoring, a novel approach in MS research. The study's notable contribution is its focus on the circadian patterns of fatigue and mood symptoms, offering new insights into their daily fluctuations in MS patients. High patient compliance indicated the app's effectiveness as a user-friendly tool for fatigue assessment.

In a randomised controlled trial, the efficacy of the Untire mHealth app was evaluated for improving fatigue and quality of life in cancer patients and survivors [13]. Participants were divided into an intervention group with immediate app access and a control group with delayed access. The app's impact on fatigue severity, interference, and quality of life was measured over 12 weeks. The results indicated significant improvements in fatigue and overall quality of life for the intervention group, particularly among those with medium to high app usage. The app's effectiveness was consistent across various demographic and health factors. The study concluded that the Untire app is an effective tool for managing fatigue in cancer patients and survivors, offering a scalable and accessible treatment option.

In the context of Barth syndrome (BTHS), a phone app was developed to measure fatigue in real-time [14]. The study involved 36 participants, half with BTHS, who reported fatigue levels using the app six times daily while wearing an activity tracker for two weeks. The study aimed to determine if the app could distinguish between BTHS and non-BTHS individuals based on fatigue levels and correlate these with actual energy expenditure. It was found that the app successfully recorded fatigue. However, the main factor differentiating between the BTHS and control participants was "crashes" (person falling) that were recorded using an activity tracker.

The aforementioned studies focused on analysing qualitative data; in contrast, this thesis will incorporate both qualitative insights and quantitative data sourced from smartphone sensors.

A study utilising an Android application to assess fine motor abilities in 41 subjects demonstrated the potential of ML in predicting fatigue [15]. By analysing self-assessed tiredness levels and task data, the model achieved 78.8% accuracy and 96.0% sensitivity in fatigue prediction, with a low specificity of 25.0%. However, the study noted the need for larger data sets and more precise ground truth for enhanced performance of ML models.

Another study employed a spiral drawing test with 14 subjects to detect cognitive fatigue [16]. The test involved drawing a spiral on a tablet three times during a workday. The spiral was also split into sectors because the behaviour of the drawing process differs within the spiral. All the parameters were computed for each segment. Two sets of models were trained. First with only temporal features and the second with non-temporal features. For both sets four ML classifiers were trained: logistic regression, K-NN, decision tree, and the SVM. The resulting solution achieved up to 89.4% accuracy. This suggests that the presence of fatigue is reflected in the precision and smoothness of movement and motor skills and that tasks like spiral drawing are reliable methods for collecting data to detect fatigue. The presentation of these findings is limited, as it addresses only the accuracy aspect.

The presented works demonstrate collectively that the application of ML in fatigue detection holds promise for various applications, including medical diagnostics and safety monitoring. Moreover, the realisation of accurate ML models by the use of fine-motor skill tests, coupled with kinematic parameters to detect PD shows promise that these tests can be used to train accurate models. Also, the promising outlooks for fatigue measurement using smartphones and ML provide an appealing avenue for cost reduction and availability increase. However, improvements to existing solutions are needed to overcome mentioned limitations.

## 1.2 Problem Statement

The principal objective of this research is the identification of mental fatigue using a smartphone-based application as a tool for data acquisition. Integral to the thesis's progression, the fundamental research question that will be investigated is:

- Is it possible to detect mental fatigue using smartphone application-based fine motor skill tests?

The goal of this thesis is to widen the availability of smartphone applications meant for completing tasks that measure fine motor skills and use ML models to detect mental fatigue. Furthermore, to improve the ground truth by detecting the changes in fine motor skills before and after completing a mental fatigue-inducing task and improving the questionnaire presented to the user to widen the dataset, to see if these changes can be useful in training an ML model to achieve higher detection results.

In the earlier works, an Android mobile application was created [15]. The first focus of this thesis is to analyse the existing application and create an updated iOS application so the number of subjects available to complete the tests would be larger. The iOS application will be based on the Android application mentioned previously, but it will contain some updates. The application will still have 6 main parts: initial questionnaire, reaction test, spiral drawing test, reaction time test with colours, tremor test for the right hand, and tremor test for the left hand. In the questionnaire section, questions about daily activity type, level of education, how challenging or boring the previously performed task was for the subject, how difficult the previously performed task was and the level of current anxiety will be added. Additionally, feedback after completing the test will be added for the subject. The application will use a native iOS language called Swift to ensure the highest quality of data available for measuring from the application.

The second focus of this thesis would be to bring the previous Android application and the applications related to their work to use Tallinn University of Technology's servers and accounts so the work done previously could be continued by other researchers.

The next step would be to analyse the data received from the tests and prepare it. This includes feature elimination and feature selection. The final step is to find the best combination of features and classifiers to create a high-performance ML model.

This thesis is organised into five main parts. It starts with an overview of the methodologies used to attain the findings. Chapter 3 presents the development of a novel iOS application,

together with a description of necessary modifications to the back-end application and the Android application. Following this, Chapter 4 of the thesis introduces a new dataset and delves into its analysis, including feature engineering and fatigue categorisation. The next chapter focuses on data exploration and unsupervised ML techniques. Chapter 6 then shifts to explore supervised ML, highlighting the most effective models discovered. The thesis concludes with a discussion of the topic and a conclusion in Chapters 7 and 8 respectively.

# 2. Methodology

In this chapter, a general overview of the methods used to achieve the results is given.

## 2.1 Server migration

To further continue the research initiatives of Valla, Nõmm, and Toomela in the Department of Software Science at Tallinn University of Technology, a comprehensive analysis of their preceding work was conducted. As documented in [15], their prior work encompassed the development of an Android application, alongside both back-end and front-end applications. To facilitate the continuity and furtherance of this research by other scholars, it was deemed essential to transition the operational management of these applications from the original developers' accounts and servers to those managed by Tallinn University of Technology. Such a transition was important for enabling ongoing maintenance, modifications, and enhancements of the code base and back-end infrastructure, thereby contributing to the sustained development and empirical evaluation of the solutions derived from this work.

Specifically, for the Android application, a Google Play application transfer request was utilised to ensure its alignment under the official account of Tallinn University of Technology. This process also necessitated the resetting and recreation of the application signing key to enable the upload of new application releases.

Furthermore, the existing architecture, comprising a front-end Svelte application and a Kotlin back-end application with an associated database, was initially distributed across two distinct servers. These components have now been consolidated and are operational within a single server infrastructure at Tallinn University of Technology, accessible via fatiguetest.cs.taltech.ee. Within this server, the '/api' endpoints are configured to route to the back-end application, thereby facilitating seamless interaction with both the mobile and front-end applications. Additionally, the terms of service document has been integrated and is accessible at the '/tos' endpoint, ensuring compliance with relevant legal and ethical standards.

## 2.2 Mobile Application

In the field of ML, where the quantity of data significantly impacts model effectiveness, the creation of an iOS application to complement the existing Android app was essential

for gathering a comprehensive dataset. This initiative was key to obtaining a wide-ranging and extensive set of data. The Android application was first examined, revealing features such as instructional guides, an initial questionnaire, and four different tests. After these tasks, users could access a website to review the data they contributed.

The iOS app was crafted using Swift, a language specifically designed by Apple for its devices [17]. Apple's XCode and App Store Connect were used for writing and distributing the app, respectively, under the management of Tallinn University of Technology.

The first version of the iOS app aimed to closely match the Android app's design and was tested using Apple TestFlight by the thesis supervisor. Feedback from the supervisory team, including Professor Aaro Toomela from Tallinn University, led to several suggestions for enhancing the accuracy of the results.

These suggestions involved more rigorous task completion procedures, requiring users to complete tests twice with about an hour's interval, and providing feedback on their performance. The questionnaire was expanded to include questions about the user's education, the nature of their daily activities, their anxiety levels, their interest in and the mental demand of their recent task.

A major change involved making the instructional content more straightforward to understand. The longer, detailed tutorial guides were replaced with clearer illustrations and animations. These animations, created using Adobe After Effects, Apple Emojis, and Lottie Animations, demonstrated each task before users attempted them.

Following user tests, the questionnaire format was altered. Initially, only basic information was collected, and in the second session, users responded to more detailed questions. After starting to analyse the collected data, a question about the users' own assessment of tiredness was added back to the first session, as this aspect had shown to be particularly effective in earlier research [15].

These changes in the iOS app were also applied to the Android and back-end applications. A significant change enables the provision of feedback to the user based on the differences between the first and second test results.

## 2.3 Data acquisition

The principal methodology for data acquisition in this study involved a collaborative arrangement with university professors 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 brought out in Appendix 3, 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 networks, encompassing family, friends, and peers. 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 data collection phase was initiated on the 16th of November, 2023 and concluded on the 16th of December. Throughout this period, a total of four instructional sessions at Tallinn University of Technology were utilised for the purposes of data gathering. The decision number 12 by the Tallinn University Board of Ethics, dated May 12, 2021, established guidelines for the process of data collection.

### 2.3.1 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 hypothesised 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.

# 3. Fatigue Detection Application

The research outlined in [15] 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 was initiated from scratch.

Enhancements to the back-end application and its associated database were imperative to accommodate additional fields introduced in the basic data block, as well as the new data fields introduced into the questionnaire. 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 ranging between 25 to 100 minutes 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.

Regarding the delivery of user feedback, the updated back-end now includes calculations for various test types. For reaction tests, it quantifies changes in test completion time and error frequency. The spiral test analysis encompasses an evaluation of error variation, along with assessments of changes in duration, spiral length, drawing velocity, acceleration, and stability. For hand tremor tests, the system calculates the variance in asymmetry between the two hands.

The infrastructure hosting these applications underwent a transition, with both the front-end and back-end components being migrated to the servers of Tallinn University of Technology. This move has rendered the applications accessible via https://fatiguetest.cs.taltech.ee/ for the front end and https://fatiguetest.cs.taltech.ee/api for the back end. 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.

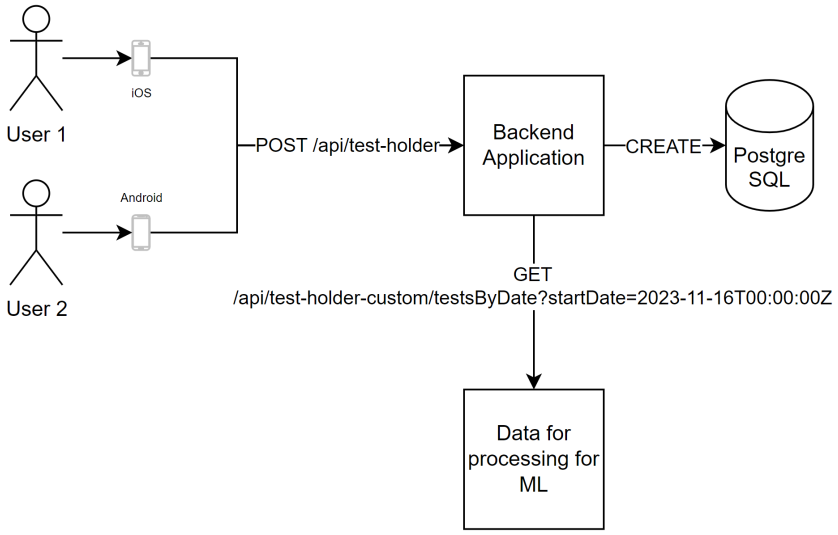



Figure 1. Applications high-level flow chart

## 3.1 Mobile applications

On May 29th, 2023, the inaugural version (1.0) of the iOS application was deployed to Apple's TestFlight platform, initiating its testing phase. This preliminary iteration underwent significant refinements, evidenced by the submission of 17 subsequent updates aimed at enhancing its functionality and user experience. Marking a milestone in its development, the application with the name "Fatigue Test TalTech" was officially released on the App Store on November 7th, 2023, as documented in Figure 2. The version history indicates a series of five updates post-launch, primarily focused on incremental improvements. These enhancements included modifications to tutorial texts, informed by student feedback, minor design changes, expansion of the application's geographical availability, and the reinstatement of a previously omitted question in the initial test sequence. It is important to note that the application collects device IDs from its users, a practice that is explicitly acknowledged in the app's privacy policy under the category 'data not linked to you'.

The Terms of Service document of the mobile applications has undergone comprehensive revision and can be seen in this thesis in Appendix 4. In conjunction with this, a specialised website was developed utilising Google Sites to articulate the privacy policy, a mandate for submission to the Google Play Store and the App Store. Concurrently, a support web page was also established through Google Sites, fulfilling the prerequisites for the App Store's public distribution of the application. The website detailing the privacy policy is accessible to the public at https://sites.google.com/view/fatigue-test-taltech/home, while the support page is available at https://sites.google.com/view/fatigue-test-taltech-help/home.

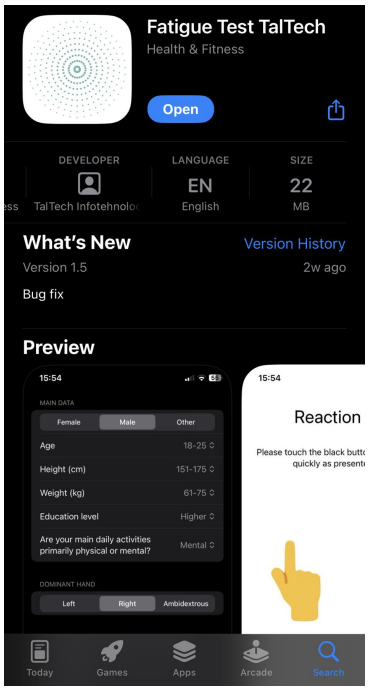



Figure 2. The application in the Apple App Store

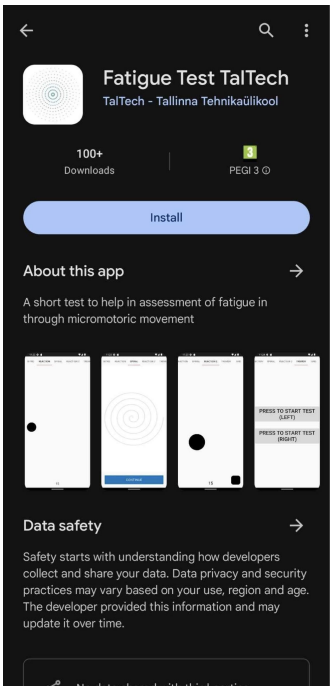



Figure 3. The application in the Google Play Store

The Android variant of the application with the name "Fatigue Test TalTech" was successfully deployed to the Google Play Store on November 6th, 2023. A screenshot of the Android application on the Google Play Store is brought out in Figure 3. After this initial release, there have been four additional version releases. Notably, these subsequent releases have been consistent with the updates made to the iOS application. A parallel development strategy was employed for both platforms.

The testing framework for the iOS application was conceptualised and structured based on the pre-existing Android application architecture. 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 screenshots taken 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 4. After accepting these terms, users are given general instructions for performing the tests. The first interface of the application, together with the general instruction, is brought out in Figure 4. The screenshots used to illustrate the workflow of the application in this section were taken on an iPhone.

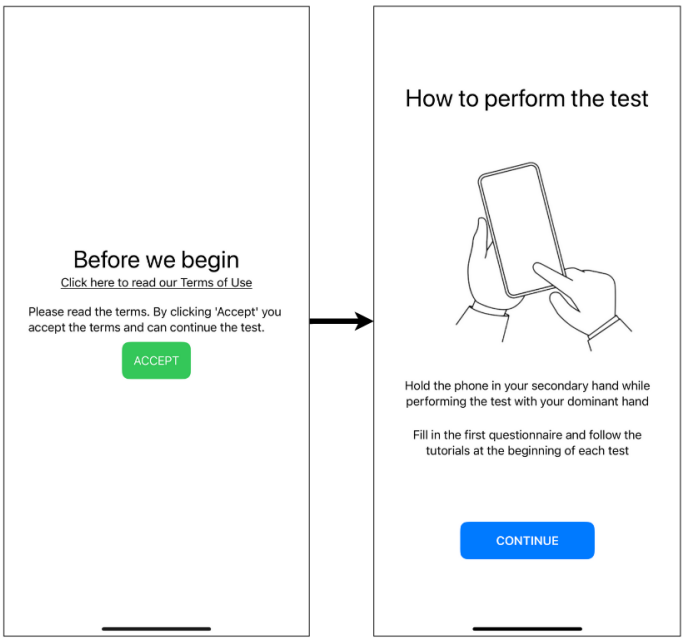



Figure 4. First views in the iOS application.

Initially, upon the first completion of the test, both the iOS and Android applications were programmed to solicit merely basic user data. However, a comprehensive update was made on December 6th, preceding the final session of lessons for data collection. This update was necessary as it modified the applications to include queries about the user's self-assessed level of tiredness during their initial interaction.

The questionnaire designed to collect basic data from the users before the first completion of the test is depicted on the left side of Figure 5. The collection of basic user data includes the user's gender, age, height, weight, education level, dominant hand, self-evaluation of fatigue, and assessment of the nature of daily activities. During the second interaction with the application, both the iOS and Android versions are structured to pose more detailed and specific questions to the user. This is illustrated on the right side of Figure 5. This second data collection includes the user's level of interest in their most recent task, assessment of the mental demands of their most recent task, anxiety level, and exhaustion level. The application allows these figures to range from 0 to 10. Additionally, information regarding the number of hours that the user spent on physical and mental activities during the day and sleeping last night is collected. The application allows these figures to range from 0 to 12.

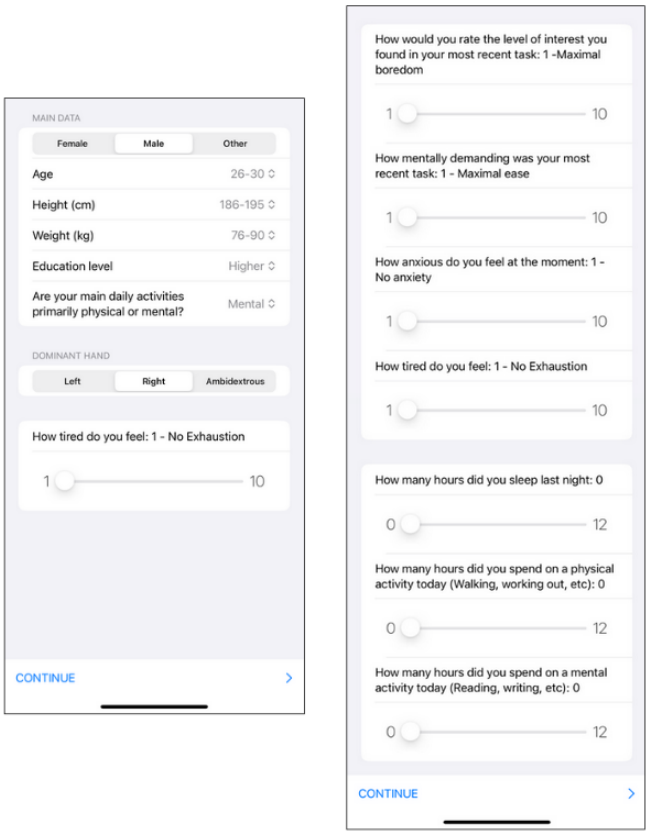



Figure 5. Questionnaires used in the application with the first questionnaire shown on the left.

### 3.1.1 The Simple Reaction Test

The Simple Reaction Test 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 user on how to execute the Simple Reaction Test 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 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. The test starts when the user taps the green 'START' button (shown in the second section of Figure 6).

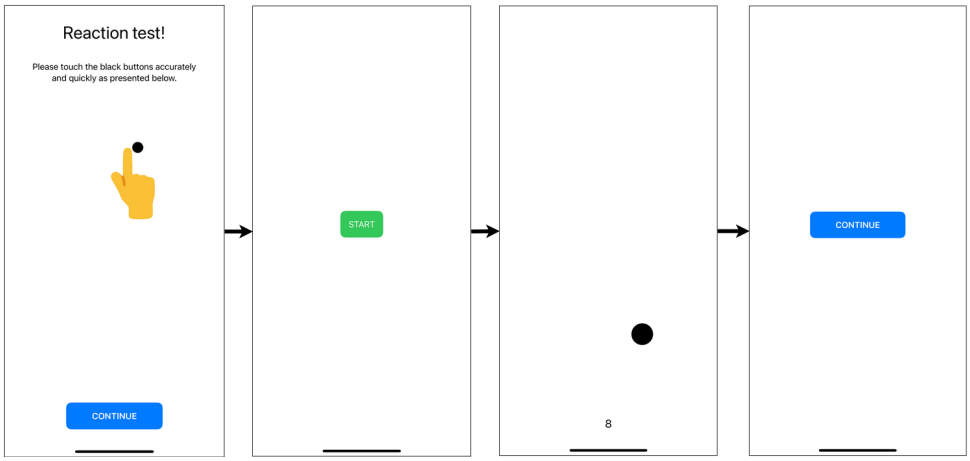



Figure 6. Simple Reaction Test views in the application.

### 3.1.2 The Spiral Drawing Test

The Spiral Drawing Test 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 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. 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. The test starts when the user taps the green 'START' button (shown in the second section of Figure 7).

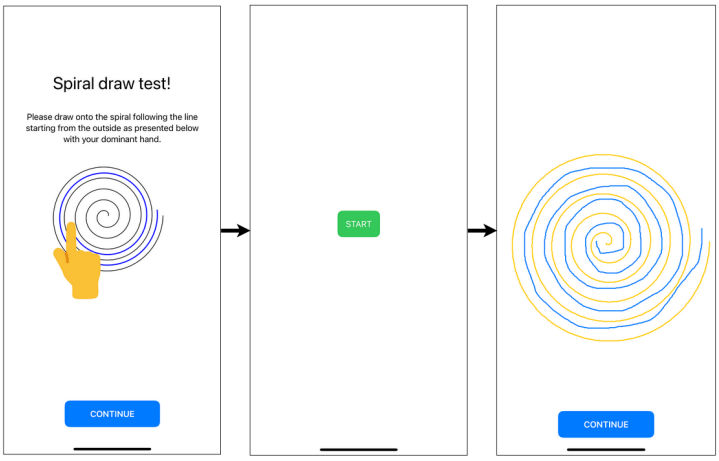



Figure 7. Spiral test views in the application.

### 3.1.3 The Advanced Reaction Test

The Advanced Reaction Test 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 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. 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 8) and finishes when the last correct dot is tapped.

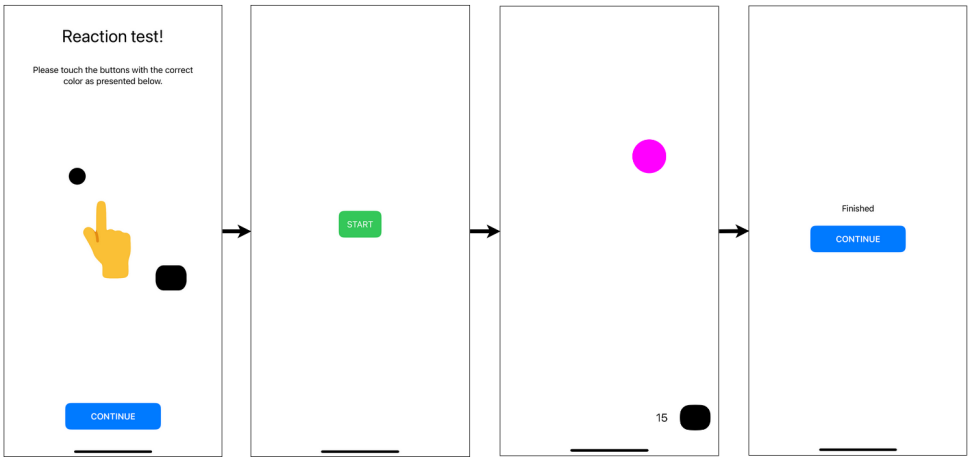



Figure 8. Advanced reaction test views in the application.

### 3.1.4 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 9.

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 9) and finishes when 10 seconds have passed (timer shown in the third section of Figure 9). The second part of the test for the right hand is identical to that of the left hand as seen from Figure 9.

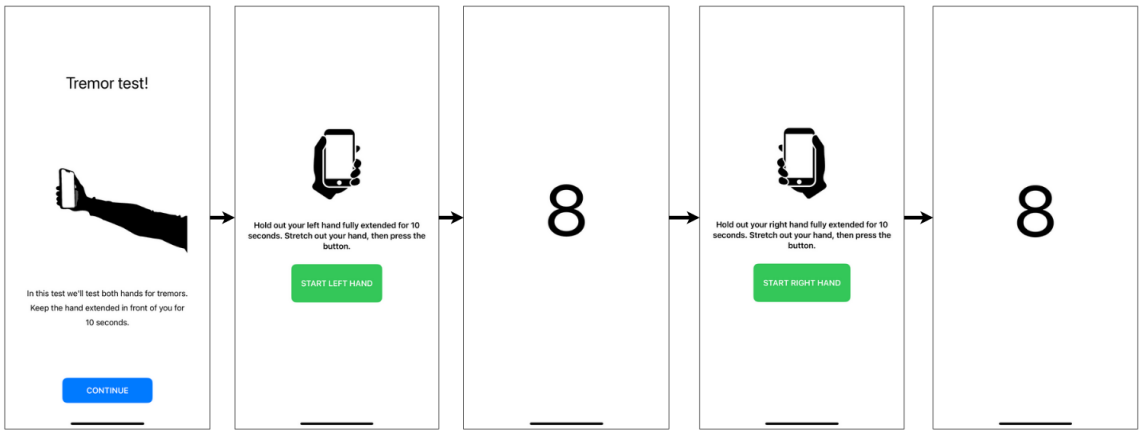



Figure 9. Tremor test views in the application.

### 3.1.5 Last Application View and Feedback

After the last test is completed for the first time, users are presented with a directive, communicated via on-screen text, to close the application and engage in a mentally taxing activity for approximately one hour. This instruction is visually represented in the left section of Figure 10.

After the users complete the tests for the second time, the application's back-end processes the accumulated data from both attempts and presents the computed results to the user. This display of results is depicted in Figure 10 on the right. The data is shown separately for each test. For the Reaction Test, the feedback consists of the difference in the number of mistakes, and the change in duration is shown to the user. Similarly, for the feedback for the Advanced Reaction Test, the difference in the number of mistakes and the change in duration is shown to the user. For the Spiral Test however, in addition to the changes in the number of mistakes and test duration, other metrics are shown to the user with a green upwards arrow (improvement) or a red downwards arrow (worsening). Lastly, for the tremor test, similar to the Spiral Test, the change in asymmetry between the hands is indicated to the user with green upwards or red downwards arrows.

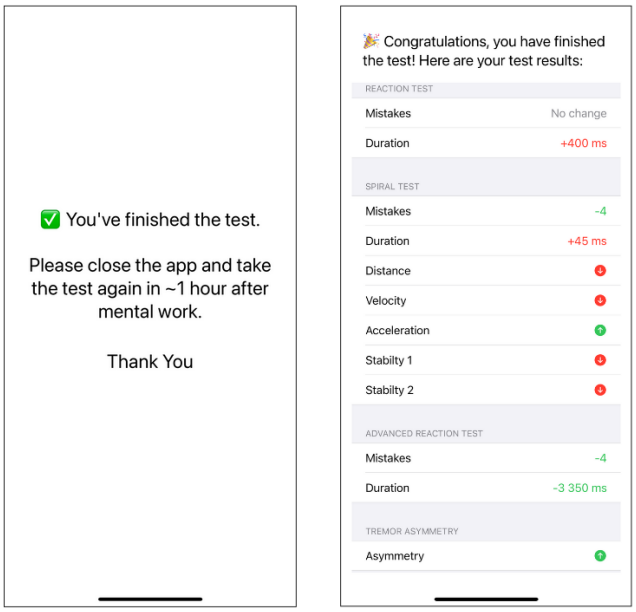



Figure 10. App views after completing the tests (first completion on the left).

# 4. Novel Smartphone Based Fatigue Dataset (SPFA-TIGUE2)

In the period extending from the 16th of November to the 16th of December, the tests were completed a total of 347 times. An analysis of device usage revealed a varied frequency in test completions: a single device recorded six completions, ten devices completed the tests four times each, and 26 devices achieved three completions. 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. It's important to highlight that the distribution of operating systems among the users was nearly even, with 51.2% using Android and the remainder, a close 48.8%, opting for iOS.

The most frequently occurring responses, derived from the initial questionnaire administered to users, are detailed in Table 1. For an in-depth view of the data collected, Appendix 2 contains a comprehensive compilation of the information gathered during the study. This detailed appendix offers useful insights into the observed patterns in the test participants.

Table 1. The most frequently occurring responses in the first completion questionnaire.

| **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% |

## 4.1 Feature Engineering

In the domain of ML, the integrity and purity of data are essential for effective pattern recognition. For this reason, it is essential to conduct rigorous data preparation and data processing. The primary step entailed refining the dataset to include only those instances where participants had completed the designated tasks within the application twice, with a special emphasis on maintaining an appropriate duration between the two test completions.

Necessary to this process was the calculation of specific features, tailored to enhance the performance of the ML algorithms. The methodologies and computations presented in [15] were systematically analysed and thereafter employed in the derivation of features, drawing on the data obtained from spiral drawing tests. This is illustrated in Figure 11. A selection of these features and their descriptions are presented in Table 2. Motion mass parameters are a set of measurements that quantify the dynamics of movement, such as velocity, acceleration, jerk, and pressure. They are calculated by summing the absolute values of these characteristics at each observation point [18]. These parameters are necessary because they provide a detailed and quantifiable analysis of the amount and smoothness of motion, which is vital for understanding complex movement patterns found through the Spiral Test [18]. Similarly, in the Tremor test, the 'absolute acceleration' measurements can be used to calculate its motion mass value. For the Simple Reaction Test, the emphasis was on calculating the mean values for metrics such as 'Was Hit On Target Sum' and 'Time From Last Touch'. This procedure was replicated for the Advanced Reaction Test, where the mean 'Time From First Correct Color Render' was also computed.

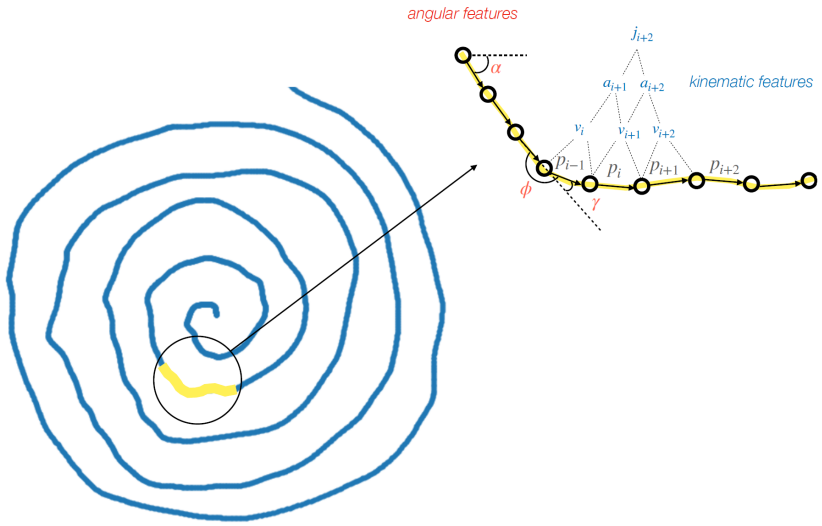



Figure 11. Visual representation of the differential-type and angular-type features [15].

Within each uuid group, the mean values for the aforementioned metrics were calculated,

Table 2. Subset of computed features for each test. Kinematic features computed from spiral tests were developed in [18]

| Test name | Feature set | Description |
|---|---|---|
| Spiral Test | distance | $d_i = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}$ (Euclidean distance) |
| | velocity_mass | Velocity mass of the point vector $[p_1, p_2, \ldots, p_k, \ldots, p_N]$ |
| | acceleration | Rate of change in velocity with respect to time. The second time derivative of the distance. |
| | $\phi$_angle_mass | Mass of the angle $\phi$ (in radians) |
| | $x$_jerk_mass | Mass of the rate of change in x-directional acceleration |
| | crackle_mass | Mass of the fifth time derivative of the distance |
| | x_acceleration_mass | Mass of the x-directional rate of change in velocity |
| | snap_mass | Mass of the fifth time derivative of displacement. |
| Reaction Tests | 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 Test | $x, y, z$ | Acceleration along $x-, y-, z-$axis |
| | absolute acceleration | $abs = \sqrt{x^2 + y^2 + z^2}$ |

ensuring that each group's aggregated data provided a comprehensive representation of the participant's performance. This nuanced approach to data processing and feature engineering was integral in preparing a robust dataset, thereby facilitating the accuracy and efficacy of the ML models in detecting the study's targeted patterns.

## 4.2 Data Cleaning

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 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 12, showing the button press preceding hand extension. For this reason, to ensure data integrity, the initial 15% portion of time in each tremor test dataset was systematically excluded from the analysis.

Further scrutiny of the dataset revealed that 26 participants had recorded their sleep hours as zero. Considering the potential impact of this inconsistency on the study's results, the sleep hours feature was excluded from the majority of cross-validation analyses.

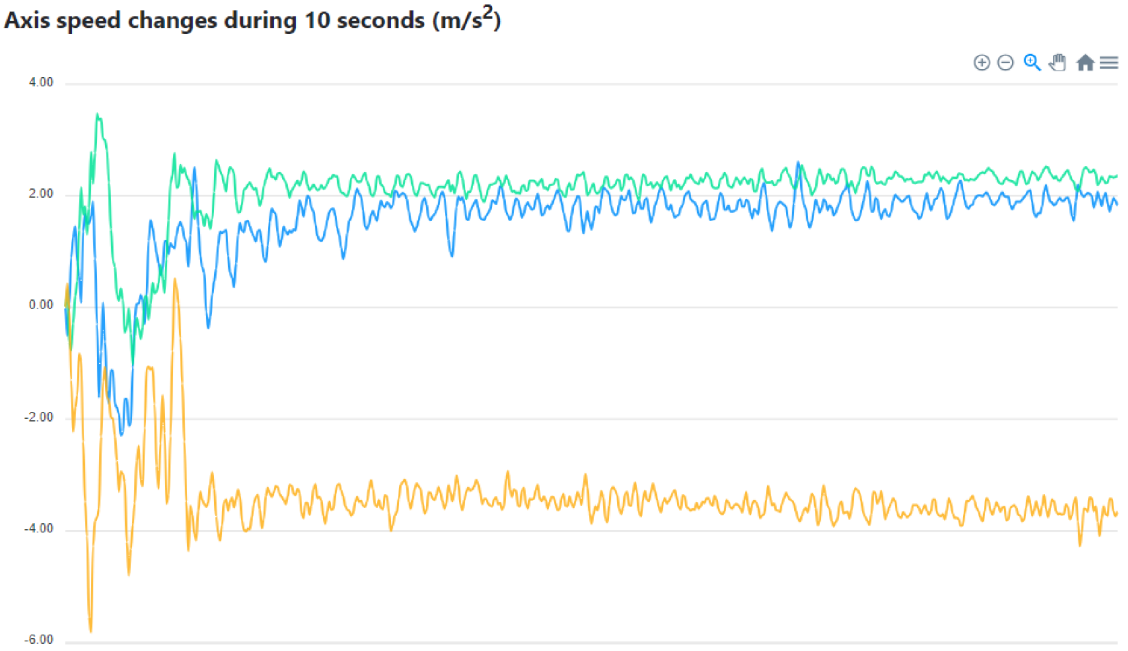



Figure 12. Example of the output of right-hand tremor measurements.

In addition to removing all instances with missing values (NaNs), 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 essential for subsequent ML analysis, ensuring a balanced and precise dataset that accurately reflects the test sessions.

## 4.3 Fatigue Categorisation

Supervised ML operates on the foundation of labelled data, which facilitates the algorithm's ability to discern patterns and subsequently develop models [19]. 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'. These classifications are contingent upon a variety of parameters that are indicative of mental fatigue. The criteria for these classifications are systematically outlined in Table 3.

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.

Table 3. Categories used for classification

| **Category** | **Threshold** | **Label** |
|---|---|---|
| Performing a fatigue-inducing task | Before the lesson<br>After the lesson | 109 (non-tired)<br>109 (tired) |
| Mental work performed in hours 1 | $> 1$ | 103 (non-tired)<br>115 (tired) |
| Mental work performed in hours 2 | $> 2$ | 140 (non-tired)<br>78 (tired) |
| Sleep hours 1 | $< 6$ | 136 (non-tired)<br>30 (tired) |
| Sleep hours 2 | $< 7$ | 104 (non-tired)<br>62 (tired) |
| Sleep hours 3 | $< 8$ | 42 (non-tired)<br>124 (tired) |
| Self-assessed tiredness 1 | $\leq 3$<br>$\geq 6$ | 69 (non-tired)<br>44 (tired) |
| Self-assessed tiredness 2 | $\leq 3$<br>$\geq 7$ | 69 (non-tired)<br>24 (tired) |
| Self-assessed tiredness 3 | $\leq 2$<br>$\geq 8$ | 51 (non-tired)<br>14 (tired) |
| Self-assessed tiredness 4 | $= 1$<br>$\geq 6$ | 40 (non-tired)<br>44 (tired) |

# 5. Data Exploration

The examination of patterns inherent in the gathered dataset necessitated the application of clustering techniques, a fundamental method in ML for categorising similar instances within the data [20]. Clustering aids in identifying analogous entities, thereby offering useful insights for feature selection in the development of ML models [20]. In this study, clustering was specifically applied to the features extracted from the tests conducted using the application.

To quantitatively assess the effectiveness of the clustering approach, the Cluster Purity was calculated. This metric facilitates a comparison of the derived clusters against established 'absolute truth' labels [21]. The divided dataset, each encompassing 109 rows, represented the test sessions conducted before and after the lessons. This division served as the basis for establishing 'absolute truth' labels within the study. Furthermore, the duration of mental work performed was utilised as an additional criterion for these truth labels: sessions, where mental work exceeded one hour, were classified as 'tired', whereas those with less than one hour were labelled as 'non-tired'. This classification approach provided a foundational framework for assessing the validity and accuracy of the clustering results. An overview of the employed clustering techniques will be presented in the following section.

K-means clustering is a popular technique in data analysis that groups a collection of items into K-distinct clusters [22]. The objective of this method is to organise these items so that the total of the squared distances from each item to the centroid of its cluster (the average point of all the items in that cluster) is as small as possible [22]. This process ensures that items are grouped with others that are most similar to them, based on their distance to these central points [22]

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is an unsupervised ML technique for identifying clusters of varying shapes in a data set [23]. Density-based clustering in unsupervised ML identifies distinct data clusters based on high point density regions, separated by sparse areas where points are considered noise or outliers [24].

The Expectation-Maximisation (EM) algorithm is a method used for maximum likelihood estimation in the presence of latent (hidden) variables within a dataset [25]. This algorithm

alternates between two steps: the expectation step, which estimates the values of the latent variables, and the maximisation step, which optimises the model parameters based on these estimates [25]. This iterative process continues until convergence [25]. The EM algorithm is commonly applied in density estimation and clustering algorithms like the Gaussian Mixture Model (GMM) [25].

Given the considerable number of features (71) present in the dataset, dimensionality reduction techniques were employed to focus on the most salient features, aligning with the research's specific interests. Initial preprocessing involved the exclusion of features where over 50% of the values were zero. This step was critical in reducing noise and enhancing the dataset's relevance for further analysis. Subsequently, a correlation matrix was constructed to explore potential relationships between the variables. Post-exclusion of features predominantly composed of zero values, 49 features remained. The correlation among these features was quantitatively assessed using Pearson's correlation coefficient. Analysis of the correlation matrix revealed instances of pronounced correlation, particularly noticeable in the central regions of the matrix and the upper left section, as illustrated in Figures 13, 14, and 15.

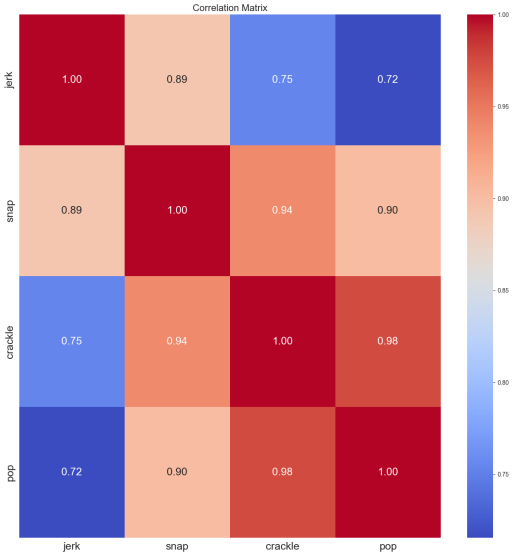



Figure 13. Correlation in the upper middle region.

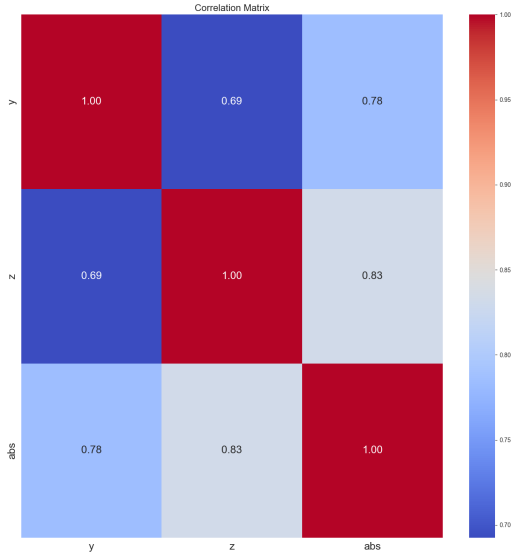



Figure 14. Correlation in the lower middle region.

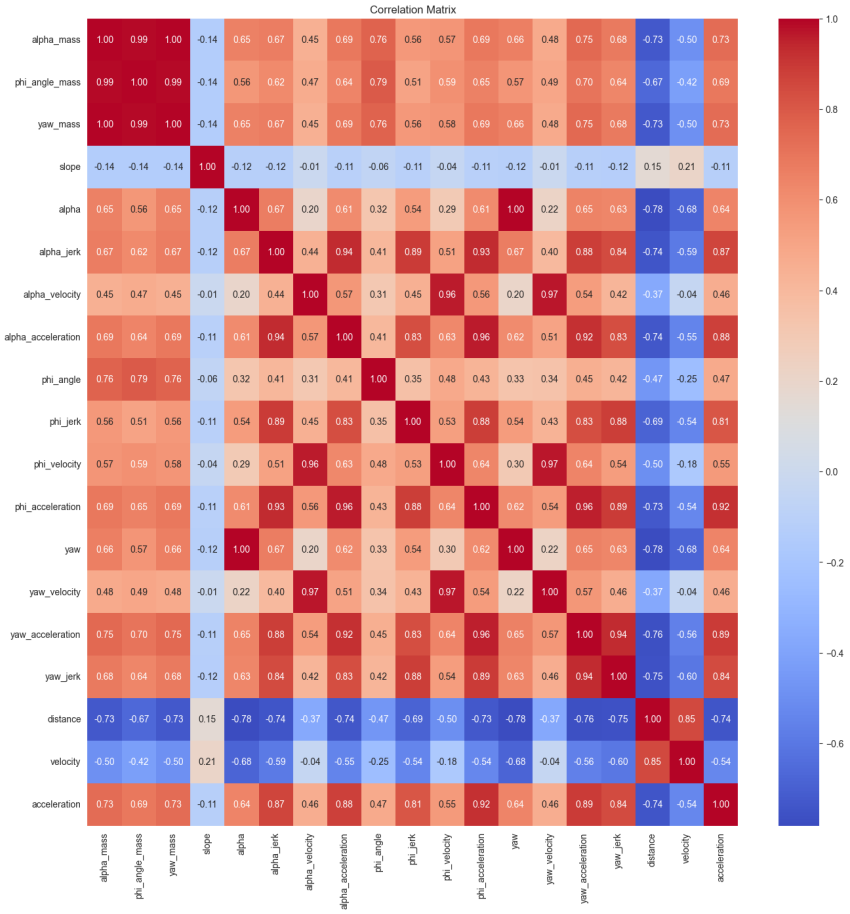



Figure 15. Correlation in the top left region.

In line with established statistical guidelines, which classify correlations ranging from 0.8 to 1.0 as highly significant [26], features exhibiting a correlation greater than 0.8 were identified for exclusion.

The elimination of features was methodically conducted by calculating the mean correlation of each feature with all other variables. This process enabled the retention of features exhibiting the lowest mean correlation, thereby ensuring the preservation of those contributing unique and significant informational value. Following this systematic removal of highly correlated features, the remaining dataset comprised 29 features.

Using the insights gained from the refined dataset, the study proceeded to the application of clustering techniques. An important step in this process was determining the optimal number of clusters (k-value) for the K-Means clustering algorithm. To achieve this, the elbow method was employed, a widely recognised technique for identifying the point at which the within-cluster sum of squares (WCSS) begins to diminish at a diminishing rate [27]. This indicates the optimal cluster count [27]. The graphical representation of

the elbow method, illustrating the relationship between the number of clusters and the corresponding WCSS, is presented in Figure 16.

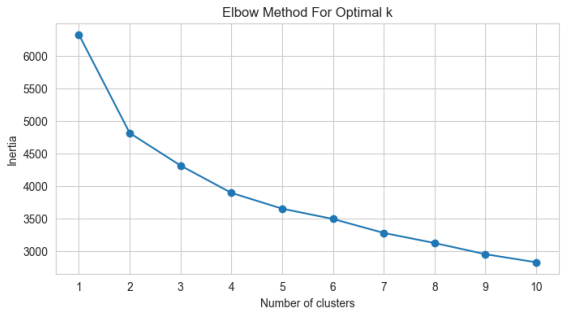



Figure 16. Clustering elbow graph for selecting k.

When employing a k-value of 2 for clustering within Group 1, a total of 110 data points were assigned to the first cluster, while the second cluster contained 108 data points. To facilitate data exploration and analysis, Principal Component Analysis (PCA) was employed. PCA is a fundamental technique aimed at simplifying high-dimensional datasets by transforming the original variables into a new set of variables termed principal components. These principal components are designed to be uncorrelated and adept at capturing the predominant patterns and variances within the data. The ability to reduce data complexity while preserving essential trends is crucial in the realm of statistical analysis [28]. The outcomes of this PCA transformation are presented in Figure 17.

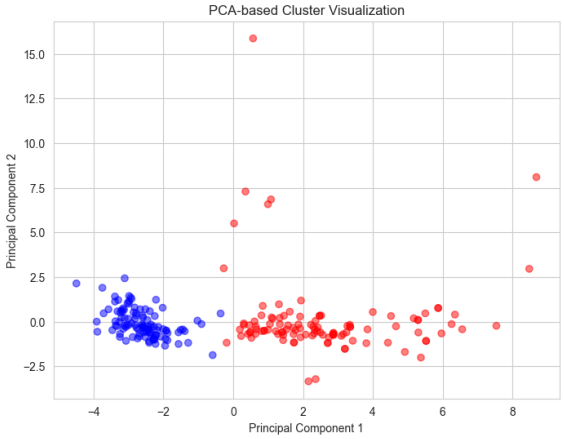



Figure 17. K-Means cluster visualisation using PCA.

The Silhouette score serves as a measure for assessing the effectiveness of clustering outcomes in data clustering [29]. It determines this by evaluating the likeness of each data point to its cluster and contrasting that with its dissimilarity to other clusters [29]. It ranges from -1 to +1 and the closer to 1 the score is the more clearly grouped the data is [29].

In this context, the achieved silhouette score for k=2 was 0.26, indicating a relatively weak clustering performance. To further assess the quality of the clustering results, the cluster purity score was employed, offering a means to gauge the alignment of the clusters with ground truth labels.

The ground truth labels were established based on two criteria: whether the tests were conducted before or after the lesson, and whether individuals engaged in mental work for more than 2 hours earlier in the same day. When comparing the clustering results with the ground truth set linked to the timing of the tests relative to the lesson, a purity score of 0.5 was obtained. Meanwhile, the ground truth set associated with extended mental work analysed through the same clustering approach resulted in a purity score of 0.64. A purity score of 0.64 suggests that approximately 64% of the total data points were accurately assigned to clusters where the most frequent ground truth label aligns with their respective cluster. This finding indicates a moderate level of agreement between the clustering outcomes and the ground truth labels, specifically concerning the influence of mental work.

To gain deeper insights into the formation of these clusters, an examination of the PCA loadings was conducted. The magnitude of each loading conveys the significance or impact of the corresponding feature on the principal component. Larger absolute values indicate a more substantial influence of the feature on that particular principal component. The top 5 features contributing to each cluster can be found in Table 4.

Table 4. Top 5 Features Contributing to PC1 and PC2 with K-Means and their loadings.

| **Feature** | **PC1** | **PC2** |
|---|---|---|
| timeFromLastTouch_rts | 0.289 | - |
| y | 0.252 | - |
| velocity | 0.249 | - |
| phi_angle_mass | 0.247 | - |
| x | 0.244 | - |
| y_acceleration | - | 0.496 |
| y_jerk | - | 0.487 |
| x_jerk | - | 0.458 |
| pop | - | 0.331 |
| y_velocity | - | 0.216 |

In the context of DBSCAN clustering methodology, an essential step involves constructing a k-distance plot. The k-distance plot is a useful tool for determining the optimal epsilon ($\epsilon$) value. This plot is brought out in Figure 18.

The epsilon value, in DBSCAN, serves as a critical parameter, delineating the maximum allowable distance between two data points for them to be categorised as neighbours within the same cluster [23]. In this study, the epsilon value was meticulously adjusted, aiming to mitigate the effects of noise within the data. Moreover, to qualify as a valid cluster, a minimum requirement of five neighbouring data points within the epsilon radius was imposed as a criterion.

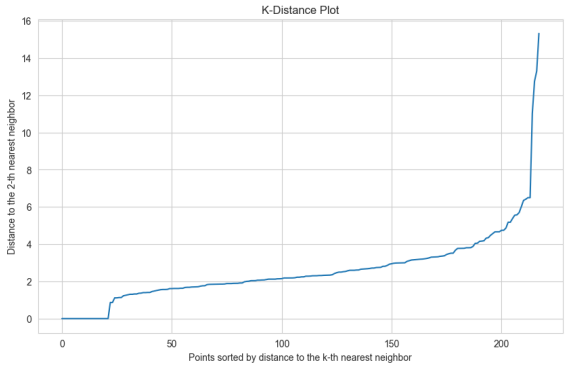



Figure 18. K-Distance plot for DBSCAN.

Based on this plot, the initially selected epsilon value was 2. However, as illustrated in Figure 19, it becomes evident that a significant amount of noise is present, and only one cluster is formed.

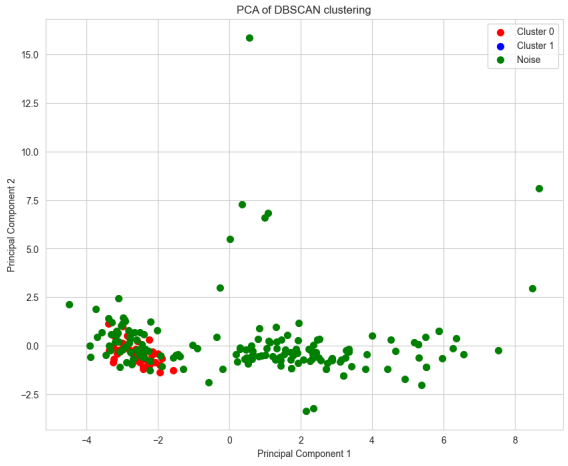



Figure 19. PCA visualisation with epsilon = 2.

After the aforementioned findings, the epsilon parameter was adjusted to a value of 2.5, as illustrated in Figure 20. This adjustment resulted in the formation of two discernible clusters, but a presence of residual noise remains evident within the data.

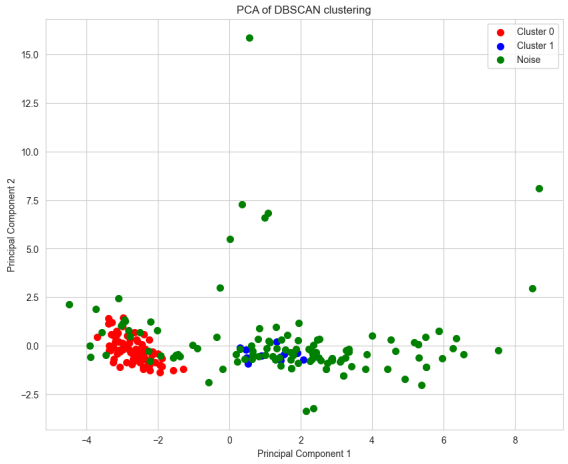



Figure 20. PCA visualisation with epsilon = 2.5.

The third epsilon value selected for the DBSCAN analysis was set to 3, and its impact is visually represented in Figure 21. The outcome reveals that 82 data points are identified as noise, while 98 points are attributed to cluster 1, and an additional 38 points are allocated to cluster 2.

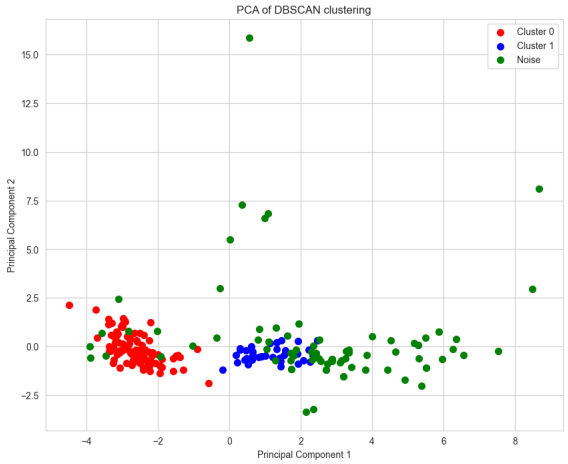



Figure 21. PCA visualisation with epsilon = 3.

The epsilon parameter was subsequently elevated to 3.5. An examination of Figure 22 indicates the absence of the second cluster, which was previously observed.

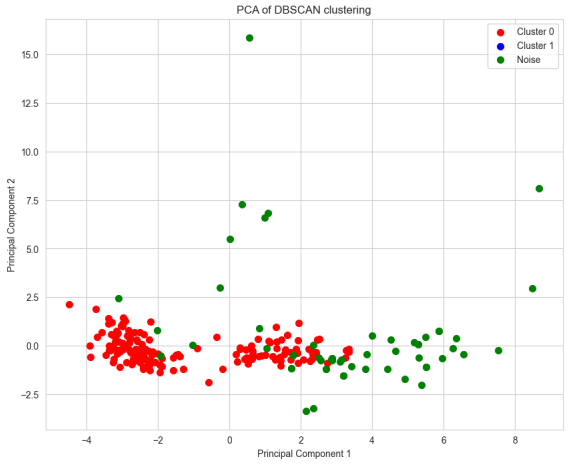



Figure 22. PCA visualisation with epsilon = 3.5.

By configuring the minimum number of samples to 20 and maintaining an epsilon value of 3.5, a small noise reduction was achieved, resulting in the identification of 79 data points classified as noise, alongside 104 data points assigned to cluster 1 and 35 data points to cluster 2. This outcome is depicted in Figure 23.

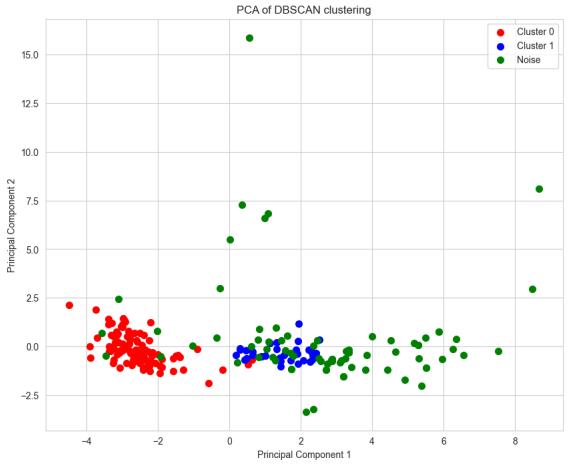



Figure 23. PCA visualisation with epsilon = 3.5 and minimum samples = 20.

Utilising an epsilon value of 3 and setting the minimum number of samples to 5, the clustering purity score was computed with reference to the previously mentioned ground truth labels. After the removal of data points identified as noise, a total of 136 data points remained for analysis.

The purity score, derived from the ground truth categorisation based on whether the tests were conducted prior to or after the lesson, yielded a value of 0.54. Likewise, when assessed against the ground truth labels pertaining to the duration of mental work undertaken, a

purity score of 0.61 was obtained. To identify the features having the most substantial influence on cluster formation, the mean values of each feature within individual clusters were computed and meticulously examined. Features exhibiting the greatest variability across clusters are indicative of their significance in shaping the clustering patterns. Among these, the top 5 features with the most pronounced differences in mean values were identified and are in Table 5.

Table 5. Difference in Feature Means Between DBSCAN Clusters

| **Feature** | **Mean Difference between clusters** |
|---|---|
| timeFromLastTouch_rts | 4236.1 |
| phi_angle_mass | 2572.7 |
| overallTime | 2451.7 |
| overallTime_rts | 1518.4 |
| timeFromFirstCorrectColorRender | 198.5 |

In the final phase of unsupervised analysis, the EM algorithm was employed for the purpose of clustering the dataset, utilising a GMM as its foundation. The 218 data points were categorised, with 111 data points being assigned to Cluster 1 and the remaining 107 to Cluster 2. After the clustering process, PCA was applied to facilitate a more comprehensible visual representation of these clusters. As illustrated in Figure 24, the spatial distribution of the clusters derived from the EM algorithm presents notable similarities to those obtained through the K-Means clustering method. Careful examination reveals subtle variances, with certain data points being allocated to different clusters than those observed in the K-Means results. This observation underscores the nuanced differences in clustering outcomes that are inherent to the distinct methodologies employed by the EM and K-Means algorithms.

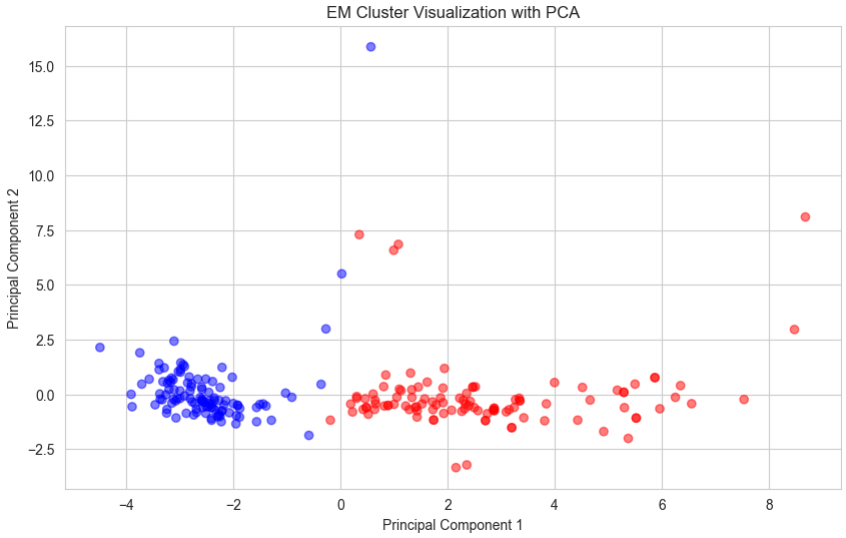



Figure 24. EM Clustering visualisation with PCA.

The purity score of the clusters formed by the EM algorithm was computed, taking into consideration the two distinct label sets previously mentioned. The calculated purity score for the 'before-after lesson' classification yielded a value of 0.51. In contrast, the purity score for the 'mental hours performed' classification was found to be 0.64. These scores provide a quantitative measure of the degree to which each cluster comprises data points from a single class, thus offering insight into the effectiveness of the EM algorithm in segregating the data according to the specified classifications.

Furthermore, an in-depth examination of the feature set was conducted to ascertain the primary factors influencing the formation of the clusters. By analysing the mean values of the features across the clusters, the top five features with the most significant impact were identified. These features are presented in Table 6. This table not only highlights the features but also provides a comparative analysis of their mean values within each cluster, thereby offering a comprehensive understanding of the characteristics defining each cluster.

Table 6. Top 5 Features Influencing Cluster Formation in GMM

| **Feature** | **Mean Difference Between Clusters** |
|---|---|
| timeFromLastTouch_rts | 4269.87 |
| phi_angle_mass | 2756.48 |
| overallTime_rts | 1633.92 |
| overallTime | 1309.17 |
| timeFromFirstCorrectColorRender | 163.75 |

The comparative analysis conducted on various clustering algorithms underscores the efficacy of unsupervised methods in effectively grouping data points that align with the variable of mental work hours. This assertion is substantiated by the observed cluster purity scores, which span a range from 0.61 to 0.64, thereby indicating a high degree of congruence within the clusters concerning the aforementioned variable. Within the scope of this analysis, it has been determined that the dimensions 'timeFromLastTouch_-rts' and 'phi_angle_mass' are instrumental in the formation of these clusters. Their significant influence in determining cluster composition is thus highlighted, underscoring their importance within the overall clustering framework. This insight not only aids in comprehending the dynamics of the clustering process but also in identifying key variables that are important in distinguishing between different groups in the context of mental work hours.

# 6. Machine Learning Based Fatigue Classification

Cross-validation methodology was employed to systematically assess the performance metrics of various ML models. For each iteration within this process, the dataset was standardised using the StandardScaler. Feature selection was systematically conducted, choosing subsets of 1, 2, 3, 4, 5, and 10 features, utilising methods such as Recurrent Feature Elimination (RFE) and SelectFromModel (SFM). Critical evaluation metrics including accuracy, precision, sensitivity, specificity, and F1 score were meticulously measured to gauge model efficacy. Given the inherent diversity in the algorithmic nature of ML models, six distinct algorithms were selected for this study. A comprehensive summary of each algorithm, highlighting their unique characteristics and functionalities, is presented in the subsequent section.

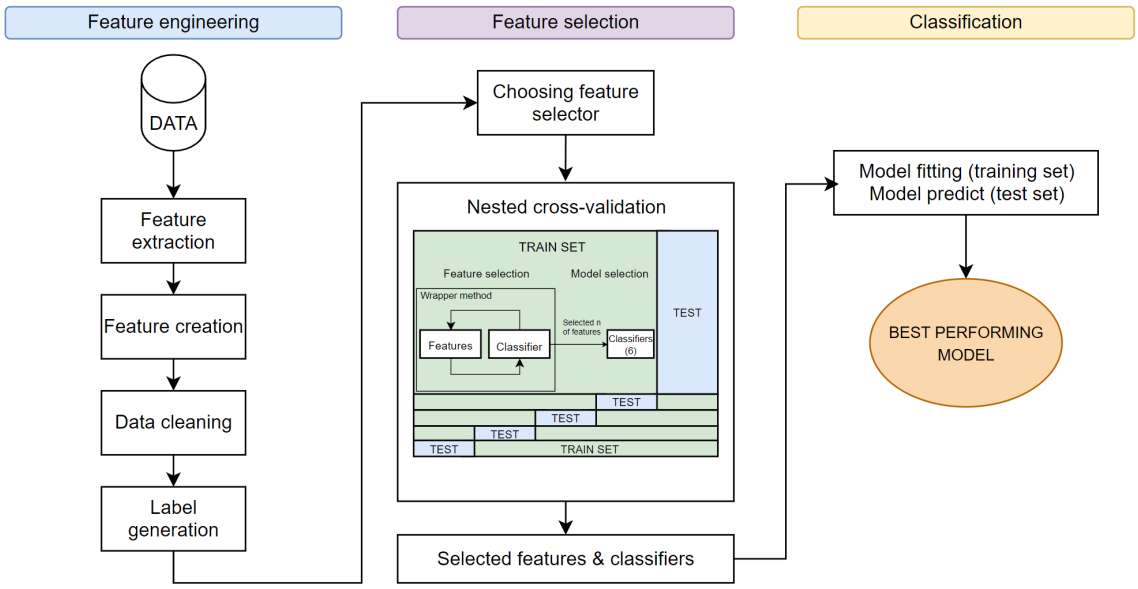



Figure 25. ML pipeline. Nested cross-validation is described more in-depth in 6.1.

### Logistic Regression

Logistic Regression (LogReg) is an ML model that is used to predict if a data point belongs to a specific category. The input of the model is converted into a probability of occurrence. The output is a binary value, whether a variable belongs in a category or not. The model uses different criteria, variables, and history to calculate the probability of falling into a category [30, 31].

### Random Forest Model

The Random Forest (RF) model is based on multiple decision trees that work together. All of the trees give a class prediction and the most common prediction is what the model outputs as the result. For this to work the predictions and errors from the individual trees have to have low correlations with other trees [32].

### K-nearest Neighbors

The dataset provided for K-NN has the data points categorised. The input data does not have a category but starts determining it by looking at the nearest K amount of neighbours in the dataset. The Euclidian distance is calculated between the input sample and the categorised chosen training samples. The category with the maximum number of data points having the least distance from the input sample is chosen. The value for K can be different for each run of the classifier to see which K gives the best results [33].

### Support Vector Machines

The Support Vector Machine (SVM) model creates an ideal separating line between two classes, the ideal separating line is the one that maximises the distance from the nearest element in both groups [34].

### Decision Tree

A decision tree (DT) is a tree-based technique widely used in ML and data mining. It follows a path from the root through a sequence of data separations, leading to a Boolean outcome at the leaf node. This hierarchical representation of knowledge includes nodes and connections, where each node signifies a decision point. Decision trees excel in classification and grouping tasks, known for their simplicity and effectiveness across diverse data types [35].

### AdaBoost

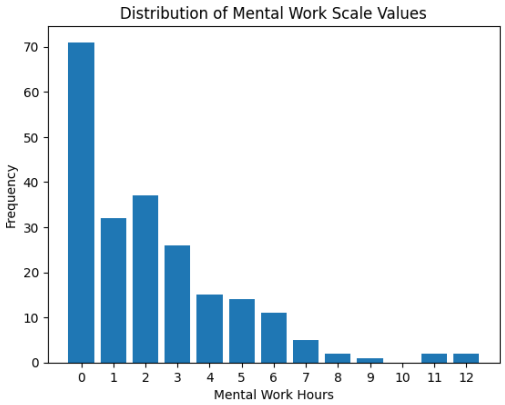
The AdaBoost model is also comprised of smaller models called stumps which are built one after the other, so the accuracy of one model's predictions influences the training of the next models. Because of this, the order is important and the first stump selected should have the lowest impurity and should show the best results in predicting the outputs. After all of the stumps have performed their classification, the categorised outputs are summed up and the category with the highest sum is the output of the whole model [36, 37].

## 6.1 Nested Cross-Validation

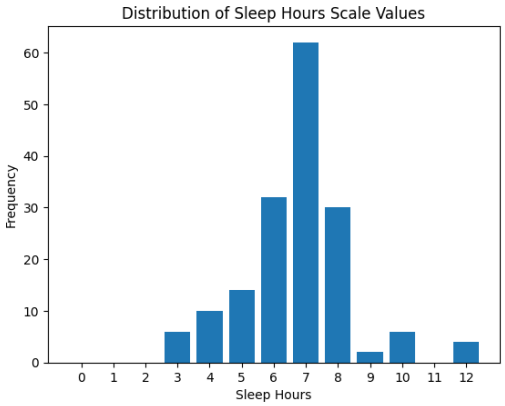
To train and evaluate an ML model, cross-validation was initially utilised to determine the most effective models, feature selectors, and labelling strategies. This method was essential for identifying the optimal setup.

The fatigue classification categories, critical to the model's functionality, are thoroughly outlined in Table 3. These categories are essential for the model's ability to accurately assess the occurrence of fatigue.

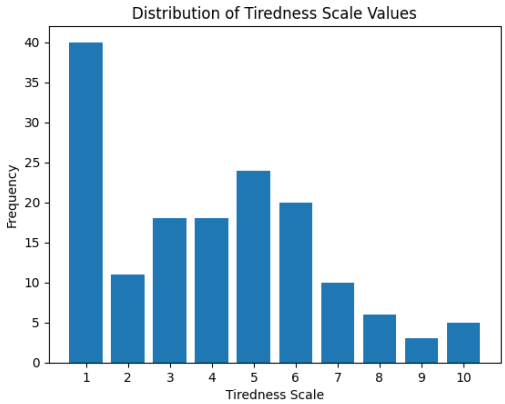
Moreover, the distribution of the features incorporated into the model is visually represented in Figure 26. This representation is key in providing an in-depth understanding of the data characteristics that the model works with, offering a clear overview of the feature set used in the study.




(a) Mental Work Hours




(b) Sleep Hours




(c) Self-assessed Tiredness

Figure 26. Distribution plots.

### Fatigue inducing tasks

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 lesson, and Group 2, also consisting of 109 entries.

Out of the total 218 data entries, a notable distribution was observed in terms of the operating systems used for recording the data: 54 recordings were completed using iOS devices, while 55 were conducted on Android devices. It is important to note that certain columns - specifically those pertaining to effort, interest, anxiety, and self-assessed tiredness - were excluded from the analysis. This exclusion was necessitated by the fact that these variables were not recorded during the initial completion of the tests.

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.

The three most exemplary outcomes derived from cross-validation utilising the Recursive Feature Elimination (RFE) feature selector are presented in Table 7. It is observed that the accuracy of these models ranges approximately between 54% and 57.2%. This range, while indicative of some predictive capability, falls short of being considered highly accurate, suggesting room for further refinement and improvement in the model's performance.

Table 7. Fatigue Inducing Tasks Cross-validation with RFE feature selector.

| Features amount | Features selected | model | accuracy_mean | precision_mean | sensitivity_mean | specificity_mean | f1_mean |
|---|---|---|---|---|---|---|---|
| 10 | Index(['alpha_velocity_mass', 'alpha_jerk_mass', 'phi_velocity_mass', | AdaBoost | 0,572370766 | 0,578646617 | 0,541176471 | 0,570220588 | 0,550760234 |
| 5 | Index(['alpha_jerk_mass', 'slope', 'distance', 'jerk', 'snap'], dtype='object') | KNN | 0,549019608 | 0,553318611 | 0,483088235 | 0,548897059 | 0,51125448 |
| 5 | Index(['alpha_jerk_mass', 'slope', 'distance', 'jerk', 'snap'], dtype='object') | AdaBoost | 0,542245989 | 0,548959276 | 0,541911765 | 0,540441176 | 0,543467324 |

The four most notable results obtained using the SelectFromModel feature selector are illustrated in Table 8. These results demonstrate an accuracy range from 57.3% to 60.8%. While this performance is an improvement over the previously mentioned RFE feature selector, it remains moderately effective, indicating potential areas for enhancement in the model's accuracy.

Table 8. Fatigue Inducing Tasks cross-validation with SelectFromModel feature selector.

| Features amount | Features selected | model | accuracy_mean | precision_mean | sensitivity_mean | specificity_mean | f1_mean |
|---|---|---|---|---|---|---|---|
| 5 | Index(['alpha_velocity_mass', 'alpha_jerk_mass', 'slope', 'distance', 'acceleration', 'jerk', 'snap', 'pop', 'x_acceleration', 'abs_l'], dtype='object') | AdaBoost | 0,60855615 | 0,605323887 | 0,604411765 | 0,609191176 | 0,601309524 |
| 10 | Index(['alpha_velocity_mass', 'alpha_jerk_mass', 'slope', 'distance', 'acceleration', 'jerk', 'snap', 'pop', 'x_acceleration', 'abs_l'], dtype='object') | AdaBoost | 0,585026738 | 0,578529412 | 0,614705882 | 0,584926471 | 0,595085995 |
| 10 | Index(['alpha_velocity_mass', 'alpha_jerk_mass', 'slope', 'distance', 'acceleration', 'jerk', 'snap', 'pop', 'x_acceleration', 'abs_l'], dtype='object') | LogReg | 0,584491979 | 0,602940383 | 0,531617647 | 0,586029412 | 0,557265512 |
| 10 | Index(['alpha_velocity_mass', 'alpha_jerk_mass', 'slope', 'distance', 'acceleration', 'jerk', 'snap', 'pop', 'x_acceleration', 'abs_l'], dtype='object') | SVM | 0,573083779 | 0,5767507 | 0,397794118 | 0,573529412 | 0,451988796 |

### Hours of mental work performed

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 tiredness 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 'tired' in 78 instances and 'non-tired' in 140 instances. Similarly, when the mental work exceeded one hour, there were 115 instances classified as 'tired' and 103 as 'non-tired'.

The most effective results achieved using the RFE and SFM feature selectors, for both the one-hour and two-hour mental work thresholds, are detailed in Table 9. This figure presents a comparative analysis of the performance metrics associated with these feature selection methods under the specified conditions.

Table 9. Cross-validation results using RFE and SFM for both more than 1-hour and 2-hour mental work results.

| Category | Feature selector | Feature count | Features used | model | accuracy_mean | precision_mean | sensitivity_mean | specificity_mean | f1_mean |
|---|---|---|---|---|---|---|---|---|---|
| > 1 Hour | RFE | 1 | Index(['physicalWorkScale'], dtype='object') | LogReg | 0,756659619 | 0,778080808 | 0,765217391 | 0,756180124 | 0,766724901 |
| > 1 Hour | RFE | 1 | Index(['physicalWorkScale'], dtype='object') | AdaBoost | 0,743023256 | 0,770673401 | 0,739130435 | 0,743136646 | 0,750895114 |
| > 1 Hour | RFE | 2 | Index(['velocity_mass', 'physicalWorkScale'], dtype='object') | SVM | 0,738477801 | 0,751124339 | 0,773913043 | 0,736956522 | 0,753761141 |
| > 1 Hour | SFM | 1 | Index(['physicalWorkScale'], dtype='object') | LogReg | 0,756659619 | 0,778080808 | 0,765217391 | 0,756180124 | 0,766724901 |
| > 1 Hour | SFM | 1 | Index(['physicalWorkScale'], dtype='object') | AdaBoost | 0,743023256 | 0,770673401 | 0,739130435 | 0,743136646 | 0,750895114 |
| > 1 Hour | SFM | 2 | Index(['velocity_mass', 'physicalWorkScale'], dtype='object') | SVM | 0,738477801 | 0,751124339 | 0,773913043 | 0,736956522 | 0,753761141 |
| > 2 Hour | RFE | 5 | Index(['velocity_mass', 'distance', 'x', 'timeFromLastTouch', 'physicalWorkScale'], dtype='object') | RF | 0,78012685 | 0,722640595 | 0,654166667 | 0,752083333 | 0,684665622 |
| > 2 Hour | RFE | 4 | Index(['distance', 'x', 'timeFromLastTouch', 'physicalWorkScale'], dtype='object') | RF | 0,770718816 | 0,695961538 | 0,641666667 | 0,742261905 | 0,665207373 |
| > 2 Hour | RFE | 5 | Index(['velocity_mass', 'distance', 'x', 'timeFromLastTouch', 'physicalWorkScale'], dtype='object') | KNN | 0,761733615 | 0,710374332 | 0,614166667 | 0,728511905 | 0,649417249 |
| > 2 Hour | SFM | 4 | Index(['distance', 'x', 'timeFromLastTouch', 'physicalWorkScale'], dtype='object') | RF | 0,775158562 | 0,698706637 | 0,653333333 | 0,748095238 | 0,672763976 |
| > 2 Hour | SFM | 5 | Index(['velocity_mass', 'distance', 'x', 'timeFromLastTouch', 'physicalWorkScale'], dtype='object') | RF | 0,770718816 | 0,701049784 | 0,640833333 | 0,741845238 | 0,666947891 |
| > 2 Hour | SFM | 5 | Index(['velocity_mass', 'distance', 'x', 'timeFromLastTouch', 'physicalWorkScale'], dtype='object') | KNN | 0,761733615 | 0,710374332 | 0,614166667 | 0,728511905 | 0,649417249 |

Upon closer examination of the results, as illustrated in the aforementioned table, it becomes evident that labelling data with the threshold of more than two hours of mental work yielded higher accuracy, ranging between 0.76 and 0.78. However, it is important to note that within this group - where 78 instances were classified as 'tired' and 140 as 'non-tired' - the average precision was 0.7, with an average sensitivity of 0.63, specificity of 0.74, and an F1 score of 0.66. The comparatively larger number of 'non-tired' instances in this group could be influencing the sensitivity and precision metrics, reflecting the model's ability to correctly identify true positives and negatives within an imbalanced dataset.

In contrast, when the threshold was set at more than one hour of mental work, the accuracy slightly decreased, with a range of 0.738 to 0.756. However, in this scenario - comprising 115 'tired' and 103 'non-tired' instances - the model exhibited a higher average precision of 0.76, sensitivity of 0.759, specificity of 0.74, and an F1 score of 0.757. The more

balanced distribution of 'tired' and 'non-tired' instances in this group may contribute to the heightened sensitivity and precision, indicating a more consistent performance by the model in a relatively balanced dataset. This correlation between group size and performance metrics highlights the impact of data distribution on the efficacy of the model.

Utilised solely the data from the second completion of the tests, encompassing a total of 109 instances and including key features such as effort, anxiety, interest in the most recent task, and self-assessed fatigue. The classification criterion was based on the duration of mental work performed: instances, where users engaged in more than one hour of mental work, were labelled as 'tired' (comprising 67 users), while those involving 0 to 1 hour were categorised as 'non-tired' (accounting for 42 users).

For this specific subset of data, the most effective results, employing SFM as the feature selector, are showcased in Table 10. Similarly, when adopting RFE as the feature selector, while maintaining the same classification criteria, the top three outcomes are illustrated in Table 11. These tables provide a comparative insight into the performance of different feature selection methods under the specified conditions, particularly highlighting the impact of including psychological and self-assessment parameters in the model.

Table 10. Cross-validation results for more than 1-hour mental work using SFM feature selector including extra features.

| Category | Features amount | Features | model | accuracy_mean | precision_mean | sensitivity_mean | specificity_mean | f1_mean |
|---|---|---|---|---|---|---|---|---|
| > 1 hour | 10 | Index(['alpha_jerk', 'phi_acceleration', 'yaw_velocity', 'yaw_acceleration', 'acceleration', 'y_l', 'x_r', 'physicalWorkScale', 'effortScale', 'interestScale'], dtype='object') | RF | 0.852813853 | 0.873489011 | 0.896703297 | 0.841407204 | 0.88289 |
| | | | SVM | 0.798268398 | 0.783005477 | 0.956043956 | 0.757188645 | 0.85485 |
| | | | DT | 0.797835498 | 0.865627706 | 0.805494505 | 0.797191697 | 0.82922 |

Table 11. Cross-validation results for more than 1-hour mental work using RFE feature selector including extra features.

| Category | Features amount | Features | model | accuracy_mean | precision_mean | sensitivity_mean | specificity_mean | f1_mean |
|---|---|---|---|---|---|---|---|---|
| > 1 hour | 10 | Index(['alpha_jerk', 'phi_acceleration', 'yaw_acceleration', 'jerk', 'y_l', 'x_r', 'timeFromLastTouch_rts', 'physicalWorkScale', 'effortScale', 'interestScale'], dtype='object') | RF | 0.852813853 | 0.833838612 | 0.954945055 | 0.82469475 | 0.888311229 |
| | 5 | Index(['phi_acceleration', 'yaw_acceleration', 'timeFromLastTouch_rts', 'physicalWorkScale', 'effortScale'], dtype='object') | RF | 0.852380952 | 0.86220696 | 0.910989011 | 0.836050061 | 0.883976815 |
| | 4 | Index(['phi_acceleration', 'yaw_acceleration', 'physicalWorkScale', 'effortScale'], dtype='object') | RF | 0.842857143 | 0.863930187 | 0.896703297 | 0.830296093 | 0.875816036 |

The analysis of the preceding tables reveals that while the RFE feature selector demonstrates a slightly enhanced performance over SFM, it is important to note that both feature selectors exhibit commendable efficacy. Notably, the SFM feature selector, particularly the model utilising Random Forest (RF) with 10 features, achieves an impressive accuracy

of 0.85. This model also shows high precision at 0.87, along with a sensitivity of 0.89, specificity of 0.84, and an F1 score of 0.88. These metrics collectively indicate a robust performance, highlighting the effectiveness of the SFM feature selector in this specific analytical context, alongside the marginally superior results of the RFE feature selector.

In the cross-validation process, the top three models employing the RFE feature selector, coupled with the Random Forest (RF) model, demonstrate notable accuracy, with values spanning from 0.84 to 0.85. Additionally, the precision of these models is noteworthy, varying between 0.83 and 0.86.

Furthermore, the sensitivity of the models, a measure of correctly identifying true positives, is commendably high, with values spanning from 0.89 to 0.95. The specificity, indicative of the model's ability to correctly identify true negatives, also shows robust performance, ranging from 0.82 to 0.836. Lastly, the F1 scores of these models, a harmonic mean of precision and sensitivity, range from 0.875 to 0.888, approaching the optimal score of 1. It is evident from the analysis that, irrespective of the feature selector employed, both the effort scale and interest scale were identified as significant features in determining fatigue levels in relation to the mental work undertaken earlier in the day. This observation demonstrates the importance that these scales have in the accurate detection of fatigue based on prior mental exertion.

**Sleep hours**

In the third phase of the analysis, sleep hours were looked at. The classification of the 'non-tired' group was based on varying thresholds of sleep duration. For instance, defining 'non-tired' as individuals who slept more than 5 hours resulted in a distribution of 136 individuals in the 'non-tired' category and 30 in the 'tired' category. Altering this threshold to more than 6 hours of sleep reclassified the groups, resulting in 104 individuals in the 'non-tired' category and 62 in the 'tired' group.

Further adjustment of the threshold to over 7 hours of sleep revealed a notable shift in group sizes, with 42 individuals categorised as 'non-tired' and 124 as 'tired'. 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-tired' group in some thresholds, potentially leading to a higher rate of false negatives. Similarly, specificity, or the true negative rate, might be impacted by the larger 'tired' group sizes, influencing the model's ability to correctly identify true negatives.

The results derived from employing RFE and SFM feature selectors under these varying sleep hour thresholds are detailed in Table 12. This table provides a comparative analysis of how different sleep duration thresholds affect the performance metrics of the feature selection methods, particularly in terms of sensitivity and specificity.

Table 12. Cross-validation results using RFE and SFM based on different sleep durations.

| Category | Feature selector | Feature count | Features used | model | accuracy_mean | precision_mean | sensitivity_mean | specificity_mean | f1_mean |
|---|---|---|---|---|---|---|---|---|---|
| < 6 hours | RFE | 3 | Index(['acceleration_mass', 'y_jerk_mass', 'slope'], dtype='object') | RF | 0,837433155 | 0,6 | 0,166666667 | 0,576058201 | 0,257142857 |
| | | 3 | Index(['acceleration_mass', 'y_jerk_mass', 'slope'], dtype='object') | DT | 0,813190731 | 0,556666667 | 0,233333333 | 0,587169312 | 0,307950938 |
| | SFM | 10 | Index(['acceleration_mass', 'y_jerk_mass', 'slope', 'alpha_jerk', 'velocity', 'y_jerk', 'z_l', 'timeFromLastTouch', 'overallTime', 'timeFromLastTouch_rts'], dtype='object') | RF | 0,843315508 | 0,88 | 0,233333333 | 0,605555556 | 0,344155844 |
| | | 2 | Index(['slope', 'velocity'], dtype='object') | RF | 0,831194296 | 0,62 | 0,366666667 | 0,65026455 | 0,416233766 |
| < 7 hours | RFE | 5 | Index(['x_velocity_mass', 'x_acceleration_mass', 'y_acceleration_mass', 'y_jerk_mass', 'crackle'], dtype='object') | LogReg | 0,626559715 | 0,516666667 | 0,129487179 | 0,526172161 | 0,205630252 |
| | SFM | 3 | Index(['velocity_mass', 'y_jerk_mass', 'crackle'], dtype='object') | LogReg | 0,656862745 | 0,7 | 0,112820513 | 0,546886447 | 0,186399483 |
| < 8 hours | RFE | 1 | Index(['snap_mass'], dtype='object') | SVM | 0,747058824 | 0,747058824 | 1 | 0,5 | 0,855153957 |
| | | 10 | Index(['alpha_jerk_mass', 'phi_jerk_mass', 'velocity_mass', 'snap_mass', 'y_acceleration_mass', 'x_velocity', 'y_velocity', 'y', 'abs_l', 'abs_r'], dtype='object') | RF | 0,74688057 | 0,780602256 | 0,92 | 0,578055556 | 0,844105021 |
| | SFM | 10 | Index(['velocity_mass', 'x_acceleration_mass', 'phi_angle', 'snap', 'crackle', 'pop', 'y_velocity', 'y', 'z', 'abs_r'], dtype='object') | RF | 0,77771836 | 0,801174968 | 0,935666667 | 0,624777778 | 0,862831215 |
| | | 5 | Index(['velocity_mass', 'phi_angle', 'snap', 'crackle', 'abs_r'], dtype='object') | SVM | 0,75311943 | 0,751793672 | 1 | 0,5125 | 0,858178761 |

The analysis of the table in question reveals that the highest accuracy, ranging from 0.81 to 0.84, was obtained in scenarios where participants had slept over 5 hours. This was consistent across both feature selectors and the RF and DT models. However, it is important to note that in these instances, the sensitivity was notably low, with values ranging from 0.167 to 0.367. This lower sensitivity can be attributed to the disparity in group sizes, with 136 individuals in the 'non-tired' group and only 30 in the 'tired' group.

In contrast, the most optimal results were observed in the group categorised as having slept for more than 7 hours. This superior performance was evident in models using SVM and RF, across both feature selectors. The peak performance within this category was achieved using SFM with 10 features, combined with the RF model. This configuration resulted in

an accuracy of 0.778, a precision of 0.8, a high sensitivity of 0.936, a specificity of 0.625, and an F1 score of 0.86.

**Self-assessed tiredness**

In the analysis where self-assessed tiredness was a focal point, a meticulous data sorting process was required to ensure that only entries with this field completed by the user were included. This criterion encompassed all second-attempt tests conducted up to the 6th of December and both attempts post the 6th of December. After this data cleansing, a total of 155 relevant data rows remained for analysis.

Initially, RFE was employed as the feature selector, the results of which are depicted in Table 13. Subsequently, the SFM feature selector was utilised, with its corresponding outcomes presented in Table 14.

Further refinement of the dataset involved the exclusion of entries where self-assessed tiredness was rated at levels 4 or 5. This reduction resulted in 113 rows. For classification, any rating above 5 was labelled as 'tired', resulting in 69 instances being classified as 'non-tired' and 44 as 'tired'.

An additional layer of data filtering was conducted by excluding the value 6. Consequently, ratings of 7-10 were categorised as 'tired' and 1-3 as 'non-tired', effectively removing a significant portion of moderate fatigue levels. This adjustment led to a new distribution: 69 instances in the 'non-tired' category and 24 in the 'tired' category.

A further division of the groups, by excluding values 3-7, resulted in 51 instances classified as 'non-tired' and 14 as 'tired'. In an even more strict classification, using only a value of 1 as indicative of 'non-tired' and all values over 5 as 'tired', the group sizes were 40 in 'non-tired' and 44 in 'tired'.

Table 13. Cross-validation results for self-assessed tiredness using RFE.

| Category | Feature selector | Feature count | Features used | model | accuracy_mean | precision_mean | sensitivity_mean | specificity_mean | f1_mean |
|---|---|---|---|---|---|---|---|---|---|
| <= 3 >=6 | RFE | 10 | Index(['alpha_velocity_mass', 'yaw_acceleration_mass', 'x_acceleration_mass', 'alpha_jerk', 'alpha_acceleration', 'phi_velocity', 'distance', 'y_jerk', 'timeFromLastTouch', 'timeFromLastTouch_rts'], dtype='object') | SVM | 0,628458498 | 0,614285714 | 0,277777778 | 0,566910867 | 0,343030303 |
| <= 3 >=7 | RFE | 5 | Index(['phi_velocity_mass', 'x_velocity_mass', 'slope', 'alpha_jerk', 'y'], dtype='object') | SVM | 0,730994152 | 0,266666667 | 0,13 | 0,535879121 | 0,171428571 |
| | | 10 | Index(['phi_velocity_mass', 'x_velocity_mass', 'slope', 'alpha_jerk', 'phi_acceleration', 'yaw', 'yaw_velocity', 'x', 'y', 'wasHitOnTarget'], dtype='object') | SVM | 0,720467836 | 0,2 | 0,04 | 0,498571429 | 0,066666667 |
| | | 5 | Index(['phi_velocity_mass', 'x_velocity_mass', 'slope', 'alpha_jerk', 'y'], dtype='object') | KNN | 0,709356725 | 0,38 | 0,28 | 0,566923077 | 0,30987013 |
| | | 2 | Index(['phi_velocity_mass', 'x_velocity_mass'], dtype='object') | KNN | 0,70877193 | 0,266666667 | 0,16 | 0,527802198 | 0,187012987 |
| <= 2 >=8 | RFE | 2 | Index(['alpha', 'mentalWorkScale'], dtype='object') | KNN | 0,8 | 0,466666667 | 0,266666667 | 0,604242424 | 0,326666667 |
| | | 3 | Index(['alpha', 'wasHitOnTarget', 'mentalWorkScale'], dtype='object') | KNN | 0,8 | 0,533333333 | 0,4 | 0,650909091 | 0,427619048 |
| | | 1 | Index(['mentalWorkScale'], dtype='object') | LogReg | 0,784615385 | 0,266666667 | 0,166666667 | 0,563333333 | 0,2 |
| | | 2 | Index(['alpha', 'mentalWorkScale'], dtype='object') | AdaBoost | 0,769230769 | 0,433333333 | 0,366666667 | 0,625151515 | 0,360952381 |
| = 1 >=6 | RFE | 10 | Index(['phi_angle_mass', 'pop_mass', 'phi_acceleration', 'yaw_velocity', 'y_jerk', 'paintOnTarget', 'z_r', 'wasHitOnTarget', 'mentalWorkScale', 'physicalWorkScale'], dtype='object') | SVM | 0,703676471 | 0,719761905 | 0,75 | 0,7 | 0,729718414 |
| | | 5 | Index(['phi_angle_mass', 'yaw_velocity', 'paintOnTarget', 'mentalWorkScale', 'physicalWorkScale'], dtype='object') | LogReg | 0,701470588 | 0,750714286 | 0,658333333 | 0,704166667 | 0,697368421 |
| | | 3 | Index(['phi_angle_mass', 'yaw_velocity', 'mentalWorkScale'], dtype='object') | LogReg | 0,700735294 | 0,771428571 | 0,611111111 | 0,705555556 | 0,681666667 |

The table presented above reveals noteworthy observations when stratifying the groups unevenly, specifically through the exclusion of subsets of values (i.e., excluding 4-6 and 3-7). In these configurations, the accuracy remains relatively elevated, ranging from 0.7 to 0.8. However, it is essential to note that the precision, sensitivity, and specificity metrics exhibit predominantly lower values within these contexts.

The precision metric ranges from 0.2 to 0.53, indicating a variable degree of correct positive identifications within the group. Similarly, the sensitivity metric, denoting the rate of true positive identifications, ranges from 0.04 to 0.26, reflecting a relatively modest

ability to accurately detect positive cases. Lastly, the specificity metric, which signifies the capacity to correctly identify true negatives, demonstrates values ranging from 0.499 to 0.65, implying a moderate level of precision in identifying negative cases.

These findings emphasise the trade-offs associated with uneven group stratification and highlight the delicate balance between accuracy and other essential performance metrics in the context of fatigue detection based on self-assessed tiredness levels.

The most favourable outcomes across all performance metrics are observed in the scenario where group division is relatively even, characterised by the classification of a 'non-tired' category using a value of 1, and categorising all values exceeding 5 as 'tired.' In this configuration, the group sizes are notably balanced, with 40 instances in the 'non-tired' group and 44 in the 'tired' group.

Although the accuracy in this scenario is slightly reduced, hovering around 0.7 for all three models, the precision metric attains an average of 0.74, indicating a consistent ability to correctly identify positive cases. Additionally, the sensitivity metric achieves an average of 0.673, denoting a reliable capacity to identify true positives, while the specificity metric averages at 0.7, reflecting a commendable ability to identify true negatives.

These results underscore the importance of balanced group division in achieving an equilibrium between accuracy and other critical performance metrics in the context of fatigue detection based on self-assessed tiredness levels.

Table 14. Cross-validation results for self-assessed tiredness using SFM.

| Category | Feature selector | Feature count | Features used | model | accuracy_mean | precision_mean | sensitivity_mean | specificity_mean | f1_mean |
|---|---|---|---|---|---|---|---|---|---|
| <= 3 >=6 | SFM | 10 | Index(['alpha_velocity_mass', 'velocity_mass', 'x_velocity_mass', 'alpha_jerk', 'phi_acceleration', 'velocity', 'x_jerk', 'y_acceleration', 'y_jerk', 'y'], dtype='object') | SVM | 0,627667984 | 0,555 | 0,363888889 | 0,580845543 | 0,430656109 |
| <= 3 >=7 | SFM | 10 | Index(['alpha_jerk', 'phi_velocity', 'phi_acceleration', 'yaw_velocity', 'distance', 'jerk', 'snap', 'y_acceleration', 'x', 'timeFromLastTouch_rts'], dtype='object') | SVM | 0,752631579 | 0,4 | 0,13 | 0,550164835 | 0,19047619 |
| | | 1 | Index(['alpha_jerk'], dtype='object') | AdaBoost | 0,731578947 | 0,466666667 | 0,25 | 0,573901099 | 0,296233766 |
| | | 10 | Index(['alpha_jerk', 'phi_velocity', 'phi_acceleration', 'yaw_velocity', 'distance', 'jerk', 'snap', 'y_acceleration', 'x', 'timeFromLastTouch_rts'], dtype='object') | RF | 0,709356725 | 0,146666667 | 0,13 | 0,521593407 | 0,137142857 |
| | | 10 | Index(['alpha_jerk', 'phi_velocity', 'phi_acceleration', 'yaw_velocity', 'distance', 'jerk', 'snap', 'y_acceleration', 'x', 'timeFromLastTouch_rts'], dtype='object') | KNN | 0,719298246 | 0,516666667 | 0,21 | 0,553901099 | 0,280952381 |
| <= 2 >=8 | SFM | 4 | Index(['slope_mass', 'alpha_jerk', 'acceleration', 'mentalWorkScale'], dtype='object') | KNN | 0,815384615 | 0,566666667 | 0,366666667 | 0,654242424 | 0,426666667 |
| | | 3 | Index(['slope_mass', 'alpha_jerk', 'mentalWorkScale'], dtype='object') | KNN | 0,815384615 | 0,466666667 | 0,266666667 | 0,613333333 | 0,326666667 |
| | | 5 | Index(['slope_mass', 'alpha_jerk', 'acceleration', 'wasHitOnTarget', 'mentalWorkScale'], dtype='object') | KNN | 0,8 | 0,6 | 0,433333333 | 0,667575758 | 0,48 |
| | | 5 | Index(['slope_mass', 'alpha_jerk', 'acceleration', 'wasHitOnTarget', 'mentalWorkScale'], dtype='object') | SVM | 0,8 | 0,4 | 0,2 | 0,58 | 0,26 |
| | | 2 | Index(['slope_mass', 'mentalWorkScale'], dtype='object') | LogReg | 0,8 | 0,3 | 0,133333333 | 0,556666667 | 0,18 |
| = 1 >=6 | SFM | 10 | Index(['slope_mass', 'slope', 'alpha_jerk', 'phi_acceleration', 'y_jerk', 'paintOnTarget', 'wasHitOnTarget', 'timeFromFirstCorrectColorRender', 'timeFromLastTouch_rts', 'mentalWorkScale'], dtype='object') | LogReg | 0,666176471 | 0,660353535 | 0,725 | 0,6625 | 0,689444444 |

The findings obtained with the SFM feature selector mirror those previously discussed for the RFE feature selector. In instances where group division is uneven, the precision, sensitivity, and specificity metrics register reductions in their values. Conversely, when the group sizes are relatively balanced, with 40 and 44 instances, the accuracy metric experiences a decline, while all other performance metrics exhibit an increase.

Exclusively utilising data from second-attempt test completions and incorporating features such as effort, anxiety, interest, and self-assessed fatigue, a classification scheme was

applied based on self-assessed tiredness. Specifically, individuals who self-assessed their tiredness as 1 were classified as 'non-tired' (29 users), while those who rated their tiredness as 6-10 were categorised as 'tired' (35 users). The optimal results obtained using the SFM feature selector under this classification are presented in Table 15. Simultaneously, the employment of the RFE feature selector within the same classification category yielded the best results showcased in Table 16.

Table 15. Cross-validation best results for self-assessed fatigue using SFM feature selector including extra features.

| Category | Feature count | Features used | model | accuracy_mean | precision_mean | sensitivity_mean | specificity_mean | f1_mean |
|---|---|---|---|---|---|---|---|---|
| = 1 >= 6 | 10 | Index(['slope_mass', 'alpha', 'yaw', 'acceleration', 'x_l', 'x_r', 'timeFromFirstCorrectColorRender', 'physicalWorkScale', 'effortScale', 'anxietyScale'], dtype='object') | RF | 0,828205128 | 0,861111111 | 0,857142857 | 0,821904762 | 0,84006993 |
| | 4 | Index(['slope_mass', 'physicalWorkScale', 'effortScale', 'anxietyScale'], dtype='object') | RF | 0,811538462 | 0,83452381 | 0,828571429 | 0,807619048 | 0,82798535 |
| | 3 | Index(['slope_mass', 'effortScale', 'anxietyScale'], dtype='object') | RF | 0,811538462 | 0,841666667 | 0,828571429 | 0,810952381 | 0,82897436 |

Table 16. Cross-validation best results for self-assessed fatigue using RFE feature selector including extra features.

| Category | Features amount | Features | model | accuracy_mean | precision_mean | sensitivity_mean | specificity_mean | f1_mean |
|---|---|---|---|---|---|---|---|---|
| =1 >=6 | 1 | Index(['effortScale'], dtype='object') | KNN | 0.812820513 | 0.926984127 | 0.742857143 | 0.818095238 | 0.796498501 |
| | 10 | Index(['slope_mass', 'slope', 'alpha', 'phi_velocity', 'distance', 'z', 'wasHitOnTarget', 'physicalWorkScale', 'effortScale', 'anxietyScale'], dtype='object') | RF | 0.798717949 | 0.880555556 | 0.771428571 | 0.799047619 | 0.787121212 |
| | | | DT | 0.798717949 | 0.841666667 | 0.828571429 | 0.794285714 | 0.816736597 |

The analysis of the top three models for each feature selector, selected from cross-validation conducted with the inclusion of additional features and a relatively balanced dataset, reveals consistently elevated performance across multiple metrics. The pinnacle of performance is observed in the results obtained with the SFM feature selector, particularly in the configuration employing 10 features and the RF model. In this setup, the achieved accuracy reaches 0.828, with precision at 0.86, sensitivity at 0.857, specificity at 0.82, and an F1 score of 0.84.

Furthermore, an examination of both tables highlights the significance of effort and anxiety recordings, as both feature selectors incorporate these variables into their modelling process. For instance, the RFE feature selector, when combined with the K-NN model and utilising only the effortScale feature, attains an accuracy of 0.81. In this scenario, the precision, sensitivity, specificity, and F1 score also maintain commendably high values.

Self-assessed fatigue was further integrated into the analysis by generating a new dataframe derived from the difference between values recorded during the first and second test attempts. In this context, it is imperative to note that only test instances conducted after the 6th of December were considered for analysis. This selective approach was necessitated by the requirement for users to assess their tiredness levels during both the initial and subsequent test completions. Consequently, the dataset exclusively encompassed test data collected after the 6th of December to ensure alignment with the assessment of self-assessed fatigue levels. This new dataframe was structured to include the four self-assessed fatigue categories as the primary labelling scheme, as shown in Table 17.

Table 17. Data classification distribution based on self-assessed fatigue categories.

| Category | Non-Tired | Tired |
|---|---|---|
| $\leq 3$<br>$\geq 6$ | 50 | 35 |
| $\leq 3$<br>$\geq 7$ | 50 | 19 |
| $\leq 2$<br>$\geq 8$ | 36 | 12 |
| $= 1$<br>$\geq 6$ | 35 | 29 |

As evident from Tables 18 and 19, the utilisation of both SFM and RFE feature selection methods yielded the most favourable outcomes when an evenly balanced dataset was employed. In this configuration, sensitivity, precision, and specificity metrics did not exhibit substantial reductions.

For SFM, the optimal performing model was the K-NN model utilising four selected features, achieving an accuracy of 0.73. In contrast, with the RFE feature selector, employing ten features and the RF model yielded an accuracy of 0.69, indicating commendable performance under these conditions.

Table 18. Best cross-validation results for dataset based on difference in values for each category using SelectFromModel.

| Category | Feature count | Features used | Model | accuracy_mean | precision_mean | sensitivity_mean | specificity_mean | f1_mean |
|---|---|---|---|---|---|---|---|---|
| <=3 >=6 | 3 | Index(['velocity_diff', 'snap_diff', 'x_l_diff'], dtype='object') | KNN | 0.658823529 | 0.62 | 0.457142857 | 0.628571429 | 0.524009 |
| <=3 >=7 | 2 | Index(['distance_diff', 'x_l_diff'], dtype='object') | LogReg | 0.710989011 | 0 | 0 | 0.49 | 0 |
| <=2 >=8 | 1 | Index(['slope_mass_diff'], dtype='object') | RF | 0.788888889 | 0.5 | 0.3 | 0.623214286 | 0.36 |
| <=2 >=8 | 1 | Index(['slope_mass_diff'], dtype='object') | DT | 0.788888889 | 0.5 | 0.3 | 0.623214286 | 0.36 |
| <=2 >=8 | 1 | Index(['slope_mass_diff'], dtype='object') | AdaBoost | 0.788888889 | 0.5 | 0.3 | 0.623214286 | 0.36 |
| =1 >=6 | 4 | Index(['alpha_jerk_mass_diff', 'yaw_acceleration_diff', 'x_l_diff', 'overallTime_diff'], dtype='object') | KNN | 0.734615385 | 0.793333333 | 0.714285714 | 0.737142857 | 0.744872 |

Table 19. Best cross-validation results for dataset based on the difference in values for each category using RFE.

| Category | Feature count | Features used | Model | accuracy_mean | precision_mean | sensitivity_mean | specificity_mean | f1_mean |
|---|---|---|---|---|---|---|---|---|
| <=3 >=6 | 3 | Index(['x_velocity_mass_diff', 'x_acceleration_diff', 'y_velocity_diff'], dtype='object') | DT | 0.635294118 | 0.545 | 0.628571429 | 0.634285714 | 0.576772247 |
| <=3 >=7 | 5 | Index(['alpha_velocity_mass_diff', 'x_jerk_diff', 'x_diff', 'y_diff', 'x_l_diff'], dtype='object') | SVM | 0.73956044 | 0.2 | 0.1 | 0.54 | 0.133333333 |
| <=2 >=8 | 1 | Index(['timeFromFirstCorrectColorRender_diff'], dtype='object') | RF | 0.746666667 | 0.4 | 0.466666667 | 0.649404762 | 0.418095238 |
| =1 >=6 | 10 | Index(['alpha_acceleration_mass_diff', 'alpha_jerk_mass_diff', 'phi_jerk_mass_diff', 'crackle_mass_diff', 'phi_velocity_diff', 'yaw_acceleration_diff', 'y_diff', 'abs_diff', 'x_l_diff', 'overallTime_diff'], dtype='object') | RF | 0.685897436 | 0.736984127 | 0.742857143 | 0.684761905 | 0.715168603 |

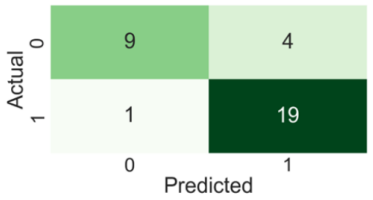
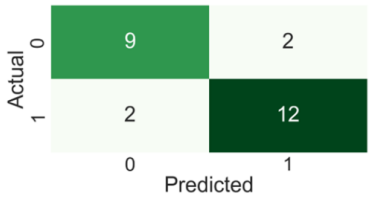
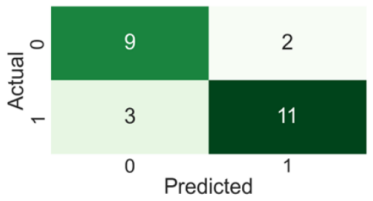
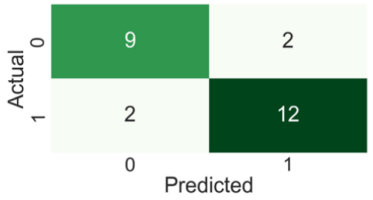
## 6.2 Best Performing Models for Fatigue Detection

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 and 2/3. Table 20 presents the top four models that showed the best performance. Among these, three models utilised the

RF classifier, while one used the K-NN classifier. The highest accuracy was observed in a model using the RF classifier with 10 features, achieving an accuracy of 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 84%, even with a reduced set of only four features. Notably, removing the feature related to physical work hours decreased the accuracy to 80%, highlighting its importance in the effective detection of self-assessed fatigue.

Table 20. 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 | 0.84 | 0.84 | **0.838** | **0.84** |   |
| Self-assessed fatigue <2 & >5 | 'slope_mass', 'effortScale', 'anxietyScale' | RF | 0.8 | 0.8 | 0.8 | 0.802 | 0.8 |   |
| Self-assessed fatigue <2 & >5 | 'slope_mass', 'effortScale', 'anxietyScale', 'physicalWorkScale' | K-NN | 0.84 | 0.84 | 0.84 | **0.839** | **0.84** |   |

# 7. Discussion

The results obtained from ML models have yielded valuable insights into fatigue detection. In contrast to previous studies, the volume of data collected in this research was substantially greater. 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 78.8% [15]. However, in this current research, a significantly higher accuracy of 85.0% was achieved, employing self-assessed tiredness and hours of mental work as labels. This finding underscores the effectiveness of using self-reported fatigue levels and mental workload as reliable 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. Delving deeper, it was observed that the kinematic feature 'slope_mass' emerged as a particularly vital component in training the models. It has been documented that these angular type features are describing some forms of micro changes in handwriting and can be linked to hand tremors [18]. 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.

The promising results achieved by the ML algorithms suggest a potential avenue for practical deployment. While the models in this research already demonstrate high accuracy, sensitivity, and specificity, further enhancing their accuracy to approach near 100% could significantly strengthen their potential for live deployment. Deploying such a model could provide users with rapid and accurate fatigue level assessments based on various input parameters.

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 demand-

ing 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 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.

# 8. Conclusion

This thesis centres on the development of a smartphone application designed to collect data to assess mental fatigue. The application, developed for both iOS and Android platforms, is equipped with a variety of tests to measure fine motor skills and a comprehensive questionnaire.

The methodology for data collection involved two phases of interaction with the application. Users performed tasks on the app before and after engaging in activities that could potentially induce mental fatigue. This approach was used for capturing data that reflects changes in fine motor skills due to cognitive exertion.

A significant portion of this thesis is dedicated to the analysis of the collected data. ML techniques were employed to evaluate the data and develop models capable of assessing mental fatigue. The study experimented with various algorithms to determine the most effective approach for fatigue detection.

One of the notable findings of this research is the role of self-reported tiredness and mental workload in predicting fatigue. The ML models showcased a degree of accuracy in identifying fatigue based on these user-reported factors, along with the changes in motor skills.

The developed application and the ensuing ML models could be utilised in contexts where monitoring mental fatigue is necessary, such as in safety-critical workplace environments or educational settings. The approach could offer a tool for real-time assessment of fatigue, providing users and researchers with valuable data on cognitive health.

Additionally, the dataset generated from this study has utility beyond fatigue detection. It provides a rich source of information on cognitive and motor functions, which could be valuable for further technical research in these areas.

In summary, the thesis presents a technical work focused on the development and utilisation of a smartphone application for mental fatigue detection. The application serves as a tool for data collection, which is then analysed using ML models to assess user fatigue. This approach contributes to the field by offering an improved method with a novel dataset and accurate models for monitoring and understanding mental fatigue.

# References


[1] Simon Skau, Kristoffer Sundberg, and Hans-Georg Kuhn. "A Proposal for a Unifying Set of Definitions of Fatigue". In: *Frontiers in Psychology* 12 (2021). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2021.739764. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.739764.

[2] Maarten A.S. Boksem, Theo F. Meijman, and Monicque M. Lorist. "Effects of mental fatigue on attention: An ERP study". In: *Cognitive Brain Research* 25.1 (2005), pp. 107–116. ISSN: 0926-6410. DOI: https://doi.org/10.1016/j.cogbrainres.2005.04.011. URL: https://www.sciencedirect.com/science/article/pii/S0926641005001187.

[3] Samuele M Marcora, Walter Staiano, and Victoria Manning. "Mental fatigue impairs physical performance in humans". In: *Journal of applied physiology* (2009).

[4] Roee Holtzer et al. "Cognitive fatigue defined in the context of attention networks". In: *Aging, Neuropsychology, and Cognition* 18.1 (2010), pp. 108–128.

[5] Kate O'Keeffe, Simon Hodder, and Alex Lloyd. "A comparison of methods used for inducing mental fatigue in performance research: Individualised, dual-task and short duration cognitive tests are most effective". In: *Ergonomics* 63.1 (2020), pp. 1–12.

[6] Jialin Fan and Andrew P. Smith. "Effects of Occupational Fatigue on Cognitive Performance of Staff From a Train Operating Company: A Field Study". In: *Frontiers in Psychology* 11 (2020). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2020.558520. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.558520.

[7] Gang Li et al. "The impact of mental fatigue on brain activity: A comparative study both in resting state and task state using EEG". In: *BMC neuroscience* 21 (2020), pp. 1–9.

[8] Adam P Vogel, Janet Fletcher, and Paul Maruff. "Acoustic analysis of the effects of sustained wakefulness on speech". In: *The Journal of the Acoustical Society of America* 128.6 (2010), pp. 3747–3756.

[9] Rohit Hooda, Vedant Joshi, and Manan Shah. "A comprehensive review of approaches to detect fatigue using machine learning techniques". In: *Chronic Diseases and Translational Medicine* (2021).

[10] Sarah Saadoon Jasim and AK Hassan. “Modern drowsiness detection techniques: A review”. In: *International Journal of Electrical and Computer Engineering* 12.3 (2022), p. 2986.

[11] Peter Drotár et al. “Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson’s disease”. In: *Artificial intelligence in Medicine* 67 (2016), pp. 39–46.

[12] Miklos Palotai et al. “Usability of a mobile app for real-time assessment of fatigue and related symptoms in patients with multiple sclerosis: observational study”. In: *JMIR mHealth and uHealth* 9.4 (2021), e19564.

[13] Simon Sebastian Spahrkäs et al. “Beating cancer-related fatigue with the Untire mobile app: results from a waiting-list randomized controlled trial”. In: *Psycho-Oncology* 29.11 (2020), pp. 1823–1834.

[14] Virginia WT Chu et al. “Development of a phone application for assessing fatigue levels in rare disorders: a feasibility and validity study”. In: *Journal of Rare Diseases* 2.1 (2023), p. 17.

[15] Elli Valla et al. “Transforming fatigue assessment: Smartphone-based system with digitized motor skill tests”. In: *International Journal of Medical Informatics* 177 (2023), p. 105152. ISSN: 1386-5056. DOI: https://doi.org/10.1016/j.ijmedinf.2023.105152. URL: https://www.sciencedirect.com/science/article/pii/S1386505623001703.

[16] Olesja Senkiv, Sven Nõmm, and Aaro Toomela. “Applicability of Spiral Drawing Test for Mental Fatigue Modelling”. In: *IFAC-PapersOnLine* 51.34 (2019). 2nd IFAC Conference on Cyber-Physical and Human Systems CPHS 2018, pp. 190–195. ISSN: 2405-8963. DOI: https://doi.org/10.1016/j.ifacol.2019.01.064. URL: https://www.sciencedirect.com/science/article/pii/S2405896319300679.

[17] Frederic Lardinois. “Apple Launches Swift, A New Programming Language For Writing iOS And OS X Apps”. In: *TechCrunch* (June 2014). URL: https://techcrunch.com/2014/06/02/apple-launches-swift-a-new-programming-language-for-writing-ios-and-os-x-apps/.

[18] Elli Valla et al. “Tremor-related feature engineering for machine learning based Parkinson’s disease diagnostics”. In: *Biomedical Signal Processing and Control* 75 (2022), p. 103551. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2022.103551. URL: https://www.sciencedirect.com/science/article/pii/S1746809422000738.

[19] Vladimir Nasteski. “An overview of the supervised machine learning methods”. In: *Horizons. b* 4 (2017), pp. 51–62.

[20] Patrick Hall. *Clustering in Machine Learning*. URL: https://www.techtarget.com/searchenterpriseai/definition/clustering-in-machine-learning.

[21] Charu C. Aggarwal. *Data Mining: The Textbook*. Cham: Springer, 2015. ISBN: 978-3-319-14141-1. DOI: 10.1007/978-3-319-14142-8.

[22] Pulkit Sharma. *The Ultimate Guide to K-Means Clustering: Definition, Methods and Applications*. URL: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/.

[23] Tara Mullin. *DBSCAN Parameter Estimation Using Python*. URL: https://medium.com/@tarammullin/dbscan-parameter-estimation-ff8330e3a3bd.

[24] Domino. *Density-Based Clustering*. URL: https://domino.ai/data-science-dictionary/density-based-clustering.

[25] Jason Brownlee. *A Gentle Introduction to Expectation-Maximization (EM Algorithm)*. URL: https://machinelearningmastery.com/expectation-maximization-em-algorithm/.

[26] Boston University School of Public Health. *PH717 Module 9 - Correlation and Regression*. URL: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module9-Correlation-Regression/PH717-Module9-Correlation-Regression4.html?.

[27] Basil Saji. *Elbow Method for Finding the Optimal Number of Clusters in K-Means*. Accessed on December 31, 2023. Sept. 2023. URL: https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/.

[28] Zakaria Jaadi. *A Step-by-Step Explanation of Principal Component Analysis (PCA)*. URL: https://builtin.com/data-science/step-step-explanation-principal-component-analysis.

[29] Hazal Gültekin. *What is Silhouette Score?* URL: https://medium.com/@hazallgultekin/what-is-silhouette-score-f428fb39bf9a#:~:text=The%20Silhouette%20score%20is%20a%20metric%20used%20to%20evaluate%20how,it%20is%20from%20other%20clusters.

[30] Tim Lou. *The Meaning Behind Logistic Classification, from Physics*. Jan. 2023. URL: https://towardsdatascience.com/the-meaning-behind-logistic-classification-from-physics-291774fda579.

[31] Xiaonan Zou et al. “Logistic Regression Model Optimization and Case Analysis”. In: (2019), pp. 135–139. DOI: 10.1109/ICCSNT47585.2019.8962457.

[32] Tony Yiu. *Understanding Random Forest*. June 2019. URL: https://www.towardsdatascience.com/understanding-random-forest-58381e0602d2.

[33] Kashvi Taunk et al. “A Brief Review of Nearest Neighbor Algorithm for Learning and Classification”. In: (2019), pp. 1255–1260. DOI: 10.1109/ICCS45141.2019.9065747.

[34] Bhanwar Saini. *Support Vector Machine — Explained*. Jan. 2021. URL: https://medium.com/swlh/support-vector-machine-explained-66b615ba0958.

[35] Bahzad Charbuty and Adnan Abdulazeez. “Classification Based on Decision Tree Algorithm for Machine Learning”. In: *Journal of Applied Science and Technology Trends* 2.01 (Mar. 2021), pp. 20–28. DOI: 10.38094/jastt20165. URL: https://www.jastt.org/index.php/jasttpath/article/view/65.

[36] Rowan Curry. *AdaBoost: Explained!* Jan. 2022. URL: https://medium.com/@curryrowan/adaboost-explained-92408a6713da.

[37] Neelam Tyagi. *Understanding the Gini Index and Information Gain in Decision Trees*. Mar. 2020. URL: https://medium.com/analytics-steps/understanding-the-gini-index-and-information-gain-in-decision-trees-ab4720518ba8.

# Appendix 1 – Non-Exclusive License for Reproduction and Publication of a Graduation Thesis[1]

I Lilian Väli

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis "Exploring the Efficacy of Smartphone Sensors in Mental Fatigue Detection: A Machine Learning Approach to Analysing Fine Motor Skills", supervised by Elli Valla and Sven Nõmm
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

05.01.2024

[1]The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

# Appendix 2 - Data Collection

Table 21. Participant information

| 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 3 - Information Sheet

## Lugupeetud õpetaja!

**Kutsume Teie kooli osalema Tallinna Ülikooli ja Tallinna Tehnikaülikooli uurimisprojektis** ***Kultuuri-, bioloogiliste ja arenguliste tegurite roll kognitiivse reservi mehhanismides ja kognitiivse taandarengu ennetamises*** (uuringu vastutav täitja, Tallinna Ülikooli professor Aaro Toomela).

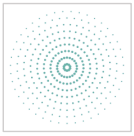
Inimkond vananeb ja vananemisega kaasneb vaimse võimekuse vähenemine kuni selle kadumiseni veel inimese eluajal. Käesoleva uuringu kõige olulisemaks eesmärgiks on vananemise ja ajukahjustusega seotud vaimse taandarengu mehhanismide mõistmine ning selle alusel taandarengu aeglustamist toetavate tegevuste parem planeerimine. Üks meie laiaulatusliku uuringu alateemadest on **inimese peenmotoorsete oskuste ja vaimse väsimuse vahelise seose mõistmine**. Teie koolis soovime õpilastel hinnata peenmotoorseid oskusi. Testid esitatakse digitaalselt iga õpilase personaalse nutitelefoni rakenduses. Testide täitmiseks kulub aega orienteeruvalt 5 minutit.

**Rakenduse nimi:** Fatigue Test TalTech

**App Store:** https://apps.apple.com/us/app/fatigue-test-taltech/id6449683047

**Google Play Store:** https://play.google.com/store/apps/details?id=ee.ainsus.fatiguetest.android

**Rakenduse Privacy Policy:** https://sites.google.com/view/fatigue-test-taltech/home

**Rakenduse ikoon:** 

Rakendus on mõeldud Tallinna Tehnikaülikooli ja Tallinna Ülikooli koostöös tehtava uuringu jaoks.

- Rakenduse küsitakse isikustamata andmeid nagu sugu, pikkus, kaal jms.
- Seejärel tuleb sooritada neli ülesannet (soorituse aeg: max 5 min, erinevad peenmotoorsed testid, nt reaktsioonitest, joonistustest).
- Peale esmast täitmist tuleks teha vähemalt 45 min kuni 1.5 h aega mentaalset pingutust nõudvaid ülesandeid (näiteks koolitund, koduülesannete lahendamine, koosolek).
- Rakenduse teist korda avamisel küsitakse esmalt isikult paar täpsustavat küsimust tema seisundi kohta ning seejärel tuleb samad neli ülesannet sooritada uuesti (soorituse aeg: max 5 min).
- Lõpus kuvatakse kasutajale tagasiside, kus on üldinfo ülesannete soorituste kohta kahel korral.

**Eetilised küsimused**

Uuringu läbiviijad garanteerivad isikuandmete puutumatuse. Kõik kogutavad andmed on anonüümsed. Uuringu läbiviimiseks on luba Tallinna Ülikooli eetikakomiteelt (12. mai 2021 otsus nr 12).

Kontakt:
Elli Valla, doktorant-nooremteadur
+37258058878
elli.valla@taltech.ee
Tarkvarateaduse instituut, Tallinna Tehnikaülikool

---

RUS

## Уважаемый учитель!

Приглашаем вашу школу принять участие в исследовательском проекте Таллиннского университета и Таллиннского Технического университета под названием **"Роль культурных, биологических и развивающихся факторов в механизмах когнитивного резерва и профилактике когнитивной деградации"** (руководитель исследования - профессор Таллиннского университета Ааро Тоомела).

Человечество стареет, и старение сопровождается снижением когнитивных способностей до их полного исчезновения в течение жизни человека. Основной целью данного исследования является понимание механизмов психической деградации, связанных со старением и повреждением мозга, и на основе этого планирование действий, способствующих замедлению процесса деградации. Одним из подзаголовков нашего обширного исследования является **понимание связи между мелкими моторными навыками человека и умственной усталостью**.

Мы хотели бы оценить уровень мелких моторных навыков у учащихся вашей школы. Тесты будут представлены в цифровом виде через персональное мобильное приложение каждого ученика. Для прохождения тестов потребуется примерно 5 минут времени.

**Название приложения:** Fatigue Test TalTech

**App Store:** https://apps.apple.com/us/app/fatigue-test-taltech/id6449683047

**Google Play Store:** https://play.google.com/store/apps/details?id=ee.ainsus.fatiguetest.android

**Privacy Policy:** https://sites.google.com/view/fatigue-test-taltech/home

**Значок приложения:**

# Appendix 4 - Terms of Service

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.