POLITECNICO DI TORINO

MASTER's Degree in ICT FOR SMART SOCIETIES

MASTER's Degree Thesis

Quantitative and Qualitative Insights into Telerehabilitation for Parkinson's Disease: A Data-Driven Study of Game Metrics, Patient Feedback, and UPDRS Scores.

Supervisors

Prof. GABRIELLA OLMO

Dott. GUIDO COPPO

Candidate

FABIO CALLERIO

July 2025

Abstract

Due to the progressive aging of the population and the resulting rise in individuals with neurological diseases and chronic disabilities, such as Parkinson's Disease (PD), there is a growing need for personalized care solutions that promote patient autonomy while maintaining the sustainability of healthcare services. Telemedicine, including remote monitoring and rehabilitation, represents an Information and Communication Technology (ICT)-based approach aimed at complementing traditional care and alleviating pressure on healthcare systems.

Exergames—video games that utilize motion sensors, cameras or other input devices to translate player movements into gameplay—are a promising application within telemedicine. They promote physical activity by integrating exercise into an engaging game-based experience.

Five exergames with adjustable difficulty levels, developed by Synarea Consultants and the National Research Council of Italy – Institute of Electronics, Computer and Telecommunication Engineering (CNR-IEIIT) using Unity and Google's Mediapipe, were tested in a clinical study involving 14 PD patients from the Associazione Amici Parkinsoniani Piemonte (AAPP-AIP).

Data was collected in 5 to 10 one-hour sessions per participant. During each session, patients repeated the games when possible, with difficulty levels increased after maximum scores. Following gameplay, participants completed a 5-star-based survey capturing data on medication timing (Levodopa-based), self-assessment, satisfaction, willingness to repeat, perceived difficulty and fatigue. Additionally, clinical evaluations were conducted using the Unified Parkinson's Disease Rating Scale (UPDRS), provided by an expert neurologist from Department of Neuroscience of University of Turin.

The primary objective of this study was to conduct both qualitative and quantitative analyses of the collected data, aiming to track the evolution of patient performance over sessions and assess the alignment between game metrics and survey responses. Furthermore, a correlation analysis was performed to identify relationships between gameplay performance, UPDRS scores, and survey responses—globally, per game, per session and per patient.

The results indicated a strong correlation between UPDRS motor scores and in-game performance metrics, suggesting that patients' responses over time to the exercises were influenced by their clinical condition, which also appeared to affect their self-reported experience in the surveys. Promising usability outcomes emerged both from the analysis of the maximum levels reached in each exergame and from the participants' questionnaire responses.

Acknowledgements

I wish to express my sincere gratitude to all those who have supported and guided me throughout the development of this thesis.
First, I would like to thank my supervisors, Gabriella Olmo and Guido Coppo, for their invaluable guidance, constructive feedback and constant support during every stage of this work.
I am also deeply grateful to all the members of SynArea Consultants s.r.l. for their technical assistance, encouragement and for providing a stimulating and collaborative environment throughout this project. In particular, I wish to thank Partner and Managing Director Danilo Soprani, as well as Cristiano, Francesca and Simone, for their dedication and support.
In addition, I extend my sincere thanks to Claudia Ferraris and Gianluca Amprimo from CNR-IEIIT, whose advice was instrumental in defining the objectives and guiding the analysis phase. I am also thankful to Alessandro Mauro and Lorenzo Priano from IRCCS Istituto Auxologico Italiano for their valuable contribution to the clinical evaluations and the planning of the data collection.
I am especially grateful to Ubaldo Pilotto, Marilisa Vetrò and Roberta Grasso from the Associazione Amici Parkinsoniani Piemonte (AAPP-AIP) for their commitment and to all its members, who generously participated in the study and made the data collection possible. Their dedication was essential to the success of this research.
I would also like to express my heartfelt thanks to my family, friends and colleagues for their unconditioned support, patience and understanding throughout this journey.
Finally, I wish to express my deepest gratitude to my grandmother Eliana, who has been a devoted volunteer with the Parkinson's Disease association for over twenty years, and to my late grandfather Paolo, whose life and experience inspired this project. Many of the participants in this study were his friends and fellow patients, making this work especially meaningful to me.

Table of Contents

List of Tables

List of Figures

Acronyms

MP

Mediapipe

CNR-IEIIT

Consiglio Nazionale Delle Ricerche - Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni

AAPP-AIP

Associazione Amici Parkinsoniani Piemonte

UPDRS

Unified Parkinson's Disease Rating Scale

H&Y Scale

Hoehn & Yahr Scale

PWDP

People with Parkinson's Disease

std

Standard Deviation

RCT

Randomized Controlled Trial

DBSCAN

Density-Based Spatial Clustering of Applications with Noise

Chapter 1

Introduction

In recent years, alarming demographic shifts-such as aging of global population [1] (Figure 1.1)-have increased the sense of urgency for finding and developing new solutions to cope with the consequent increase of burden on healthcare system. In particular, resources are being invested in quests for accessible and scalable "panacee". Due to the complexity and heterogeneity of the healthcare sector, some chronic, age-related and progressive diseases with motor and non-motor symptoms and complications like Parkinson's Disease (PD) are being particularly attentioned. Maintaining patients' autonomy guaranteeing an high-level cure to the individual while ensuring a sustainable usage of structures and personnel has thus become a priority. In this picture, the avant-garde solution of telemedicine, together with its conjugations like remote rehabilitation and exergames, have emerged as promising supplementations or alternatives to traditional healthcare strategies.

1.1 Project Context

This thesis is conducted within the scope of ELEVATOR [4] (Ecological Exercises in Virtual Reality to Stimulate Upper Limb Function and Monitor Motor and Cognitive Aspects at Home), a project proposed under NODES[1] (Nord Ovest Digitale e Sostenibile) ecosystem [5], one of Italy's main initiatives funded through the PNRR program NextGenerationEU. Project is coordinated by SynArea Consultants S.r.l. (Figure 1.2) in collaboration with the Department of Neuroscience at the University of Turin, the IRCCS Istituto Auxologico Italiano and the CNR-IEIIT (Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, part of Italy's National Research Council). ELEVATOR's goal is to develop a set

[1] multi-institutional innovation network to accelerate the digital and ecological transition of key industrial sectors like Health

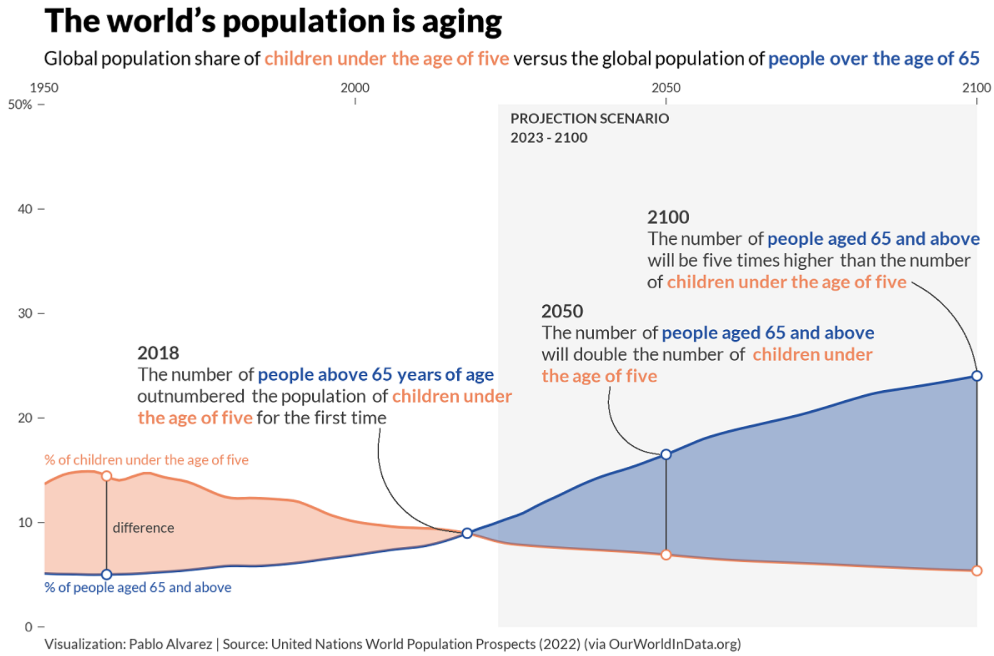

Figure 1.1: Global aging trends from 1950 to 2100. Adapted from Visual Capitalist (2023)[2], based on United Nations World Population Prospects 2022 [3].

of VR exergames, relying on budget HW & SW solutions like webcams and markerless motion trackers, in order to support motor and cognitive rehabilitation or conservation of elders and neurologically-compromised individuals, including patients with PD. Exergames have been developed, with the help of CNR-IEIIT, using Unity platform, with focus on stimulating real-life experiences like flight or ski.

SynArea Consultants [6], the company hosting this thesis project, is a Turin-based ICT firm with over 30 years of experience in the design of custom software solutions for a wide range of sectors, including manufacturing, aerospace, healthcare and education. Regarding health sector, company has invested in the development of virtual and augmented reality platforms for cognitive and motor rehabilitation. As a follow-up to its work on the REHOME [7][8] platform, it now leads the technical development and effectiveness/usability testing of exergames within ELEVATOR. The study presented in this thesis contributes to data analysis and evaluation phase. In particular, involves data collection on game data and on survey feedback obtained from up to 14 PD patients in 10 sessions and the relative qualitative and quantitative inspection to investigate on usability and effectiveness. Gameplay results are correlated with medical PD-specific evaluation scales, such as UPDRS or Hoehn and Yahr, in order to find meaningful connections between performance

Figure 1.2: SynArea Consultants logo.

and clinical conditions. Finally, on a bigger picture, this work serves to validate the potential of webcam-based motion tracking as a reliable, home-deployable alternative to costlier solutions like Microsoft Kinect, with the broader aim of promoting accessible and engaging remote care technologies.

1.2 Introduction to Parkinson's Disease

Parkinson's Disease (PD) is a chronic, progressive neurodegenerative disorder and the second most common condition of its kind. It affects approximately 0.3% of the global population, with prevalence rising to 3% among individuals over 65 years of age. PD is primarily characterized by the degeneration of dopamine[2]-producing neurons in the substantia nigra. The symptoms of this disease are classified as motor and non-motor. Core motor symptoms include bradykinesia[3], rigidity, tremor and postural instability [9]. Non-motor symptoms may involve sleep disturbances, depression, hallucinations, dementia and gastrointestinal dysfunction. Given the diversity and quantity, to manage them a combination of pharmacological and non-pharmacological intervention is employed. Current treatment strategies predominantly rely on the administration of Levodopa[4], which is the most effective pharmacological intervention to date [10]. A single oral dose can substantially alleviate motor symptoms, improving fatigue, gait and speech during a temporary phase known as the "on" time. However, Levodopa neither cures nor halts disease progression [11] and may lead to dyskinesias[5] with long-term use.
Several rating scales are available to assess PD severity. The two most widely used are the Hoehn and Yahr (H&Y) scale and the Unified Parkinson's Disease Rating Scale (UPDRS) [12].

[2] A neurotransmitter and hormone crucial to the brain's reward system, affecting motivation, pleasure, movement and several physiological functions.

[3] Slowness of movement, difficulty initiating movement and reduced amplitude.

[4] A dopamine precursor that the body converts into dopamine.

[5] Involuntary, erratic movements of the limbs or trunk.

1.2.1 Hoehn and Yahr Scale

The Hoehn and Yahr scale, named after its authors, was published in 1967 and was the first rating scale to describe the progression of Parkinson's Disease (PD)[12]. It outlines five stages of PD progression:

- **Stage One**: The earliest stage, with mild symptoms such as unilateral tremor, rigidity, or bradykinesia and minimal or no functional impairment.
- **Stage Two**: Symptoms affect both sides of the body and may include loss of facial expression and speech abnormalities.
- **Stage Three**: A mid-stage marked by loss of balance and slowness of movement. Falls are common, but patients remain fully independent in daily activities such as dressing, hygiene and eating.
- **Stage Four**: PD progresses to a severely disabling stage. Although patients may still walk and stand unassisted, they are significantly impaired and often require a walker. At this point, they are no longer able to live independently and need help with daily tasks.
- **Stage Five**: The most advanced stage, characterized by confinement to a bed or wheelchair. Patients may be unable to rise without assistance, are prone to falling and may freeze or stumble when attempting to walk.

1.2.2 Unified Parkinson's Disease Rating Scale

The Unified Parkinson's Disease Rating Scale (UPDRS), later revised by the Movement Disorder Society (MDS) into the MDS-UPDRS [13], is a comprehensive tool that incorporates input from both patients and caregivers. It consists of four parts, each rated on a scale from 0 (normal) to 4 (severe problems):

- **Part 1 – Non-Motor Experiences of Daily Living**: Covers intellectual function, mood and behavior, including anxiety, delusions, depression, apathy, sleep disturbances, pain, fatigue, constipation and urinary issues.
- **Part 2 – Motor Experiences of Daily Living**: Assesses speech, salivation, chewing, swallowing, handwriting, walking, balance, falls, freezing and the ability to perform routine tasks.
- **Part 3 – Motor Examination**: Examines motor functions such as speech clarity, facial expressions, rigidity, finger tapping, hand movements, toe tapping, leg agility, rising from a chair, gait and freezing of gait, postural stability, bradykinesia and resting tremor.

- **Part 4 – Motor Complications**: Evaluates dyskinesia (involuntary movements), the amount of time spent experiencing these complications and the severity and impact of motor fluctuations (such as "off" periods).

1.3 State of the art

In addition to personally-suited drug prescriptions and neurological examinations, PD and other similar chronic and degenerative diseases rehabilitation have traditionally relied on non-pharmacological, therapist-led complementary strategies that aim to manage symptoms and preserve quality of life.

Although effective, all these treatments often require in-person attendance, are resource-intensive and face adherence challenges in long-term care, due to the increasing prevalence of neurological disorders, as outlined in the abstract.

In order to mitigate the costs and the lack of available personnel and spaces, thanks to the improvements in the ICT field, physical and cognitive rehabilitation are moving forward to remotely-guided solutions, gathered in a sector called telemedicine.

Despite being a reality since the start of the 20th century with telephones and telegraph being the main tools for remote healthcare, with the development of modern cameras and fast internet connections it became progressively more used, with the definitive coronation as a new standard with the COVID-19 pandemic. Telerehabilitation is a specialized branch of telemedicine focused on delivering rehabilitation services remotely over telecommunication networks and the internet. It offers flexibility, cost-effectiveness and convenience, eliminating the need for travel.

Relevant is also the opportunity to receive support for late-stage diseased or individuals with disabilities that are not able to seek in-presence aid. Recent advancements in technology have introduced immersive systems like Virtual Reality (VR) and Augmented Reality (AR), enabling the development of exergames: interactive games designed for rehabilitation. These systems provide more engaging and effective experiences and are showing promising therapeutic outcomes. This gamified approach relies on softwares such as MediaPipe (MP) Pose LandMarker and Azure Kinect, which allow motion tracking and data collection, essential for monitoring patient performance remotely.

With these tools, costs are notably cut (no more need of sensors or wearable devices) and a constant health state monitoring is possible: data like survey answers, game results and body joints coordinates is gathered and analysis can be performed to objectively measure outcomes.

A reliable solution to give diagnoses is Machine Learning: Rana et al. [14] conducted a literature analysis of 112 research papers published up until 2022 in order to outline data modalities and artificial intelligence techniques that have been utilized in the analysis and diagnosis of PD. Main ML supervised and un-supervised techniques are Support Vector Machines (SVM), Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), Naïve Bayes, logistic regression, Classification and Regression Trees (CART) and decision tree.

Purpose is to differentiate between healthy subjects and people with PD (PWDP). Symptoms like tremors [15], writing difficulties or speech impairment can be used as attributes for ML algorithms to diagnose the PD severity. Data can be taken from different kind of public dataset, like [16], [17], [18] or [19].

More data is collected, whether from exergames, surveys or in-clinic crowd examinations, more accurate supervised learning classification models can be trained, giving the opportunity in the future to have automatic pseudo-diagnosis and even more auto-personalized non-pharmacological therapies.

1.3.1 Traditional medicine solutions

PD and other similar chronic and degenerative diseases rehabilitation solutions include physical therapy for motor symptoms such as gait and postural instability [20], occupational therapy for daily function [21], speech therapy for voice and swallowing difficulties [22] and cognitive interventions to counteract depression and cognitive decline [23] [24].
Conventional computer-assisted rehabilitation methods typically use basic hardware such as keyboards, mice and monitors. While affordable and accessible, these lack immersion and can result in reduced long-term patient engagement. Moreover, they offer limited support for integrated physical therapy.
This is why other initiatives, more focused on the union of entertainment and exercise, are gaining attention: examples are programs of martial arts like Tai-Chi [25], karate [26] and boxing [27] or initiatives such as dancing [28] and singing [29].

1.3.2 Telemedicine & Telerehabilitation

Telemedicine is a branch of healthcare delivery that employs telecommunications technology to provide clinical services remotely. For diseases like PD, which are

chronic, neurodegenerative and in constant need of monitoring, telemedicine represents a promising strategy to bridge gaps in care.
PD management typically involves regular assessments of motor and non-motor symptoms, medication adjustments, physical and cognitive rehabilitation. Traditional healthcare delivery often falls short in ensuring timely access to such services, especially for patients in rural areas and/or with transportation barriers. Telemedicine enables them to consult with neurologists, physiotherapists and other specialists from home, significantly reducing the physical and logistical burdens of travel.
While traditional care allows for hands-on assessments and is preferred for early diagnosis or cases of acute changes, telemedicine serves an ideal solution for standard scenarios, complementing traditional care by making routine follow-ups, medication management and rehabilitative interventions more convenient and efficient.
Telerehabilitation is one of the main branches of telemedicine. It allows patients to interact with providers remotely and can be used both to make diagnosis and to deliver therapy. Main application are physical therapy, occupational therapy, speech-language pathology, audiology and psychology.
Therapy sessions may be delivered individually or in groups. These are typically supported through webcams, video conferencing platforms, dedicated web portals, or mobile applications.
Dias et al. [30] investigated voice telerehabilitation in Parkinson's disease using videoconferencing and found significant improvements in vocal patterns and reported an high degree of satisfaction and a preference for the remote modality of exercise.
Sekimoto et al. [31] investigated the safety and feasibility of video-based telemedicine system delivered via tablet for PD subjects, which were randomly divided among clinic and online for their sessions. Study proved that medical outcomes for remotely-examined patients were the same as the ones obtained with traditional in-presence therapy.
Due to the visual nature and dependence on internet connections and devices, an important area of this branch is devoted to the investigation on telerehabilitation limits and on the study of its effectiveness via surveys, outcomes analysis and costs. These analyses are mostly based on a randomized controlled trial.[6]
For example, Silva-Batista et al. [32] examined 14 studies to review the current evidence on the effects of telerehabilitation by videoconferencing on balance and gait for patients with PD. Outcome suggested that for these two tasks in-presence solution is better: balance is particular challenging to assess by videoconferencing

[6]RCT is a study in which participants are randomly assigned to either a treatment group (receiving the intervention) or a control group (receiving a placebo or standard treatment)

and patients are generally more confident in clinics, where they can give their best without fear of falling.
Videoconferencing, however, is not the only telerehabilitation solution: thanks to the develop of mark-less motion tracking SWs as Kinect and MP, exergames are becoming for certain specific tasks the main alternative to traditional physical and cognitive re-education.

1.3.3 VR & AR exergames

Virtual Reality games are fully immersive games that place players in a completely virtual environment through which they move using body movement.
Pazzaglia et al. [33] compared a 6-week VR rehabilitation program with a conventional rehabilitation program on 51 PD patients randomly divided. Results proved that VR group showed greater improvements in balance, walking and upper limb motion.

Thumm et al. [34] created a system based on a treadmill and a depth camera (Microsoft Kinect) to track the movement of the participant's feet during treadmill walking and project it on the screen, providing real-time feedback on performance. The virtual environment consists of different obstacle-lined pathways, requiring the user to modulate his or her gait to negotiate the obstacles projected on the screen (e.g., take a longer step or increase foot clearance). At the end of the trial, outcomes showed an increase in gait speed and walk duration.

Gandolfi et al. [35] study examined patients following traditional postural stability rehabilitation or an innovative VR version based on Nintendo Wii Fit system and teleconferences. In this second case, participants played some exergames rather than practicing some standard exercise. Outcomes suggested that same results were achieved in the two modalities, but with a lower cost in the telemedicine solution.
Augmented Reality games are interactive games that overlay digital content onto the real world, allowing players to remain aware of their physical environment while engaging with virtual elements.
Hardeman et al. [36] created Reality DTx®, a digital therapeutic software platform for AR glasses that enables a home-based gait-and-balance exergaming intervention specifically designed for people with Parkinson's disease. Results showed, following an RTC program, that solution is a safe, well-accepted and potentially effective intervention in PD, with no recorded occurences of falls and just four near falls registered among 24 total participants.

Ferraris et al. [8] presented the REHOME solution, a telerehabilitation platform integrating motor and cognitive exergames for neurological disorders, including Parkinson's disease. The system leverages Kinect-based tracking and gamified tasks to support remote monitoring and rehabilitation. Preliminary usability results from 70 subjects showed high acceptance and engagement, suggesting the feasibility of home-based exergaming protocols for PD patients.

Amprimo et al. [37] proposed an integrated Azure Kinect-based system for remote motor assessment and exergaming in PD. Their approach combines standardized motor tasks with virtual exergames, enabling both evaluation and rehabilitation. In trials involving 20 PD patients and 15 healthy controls, the system demonstrated statistically significant improvements in motor performance and arm swing symmetry, validating its dual-purpose design.
A recent study (2024) by Gandolfi et al. [38] explored the use of immersive VR exergames for PD rehabilitation. The system incorporated obstacle negotiation and gait modulation tasks, with real-time feedback provided via depth cameras and treadmill-based environments. Results showed enhanced motor engagement and adaptability, reinforcing the potential of immersive VR for personalized gait training.

1.3.4 Markerless Motion Tracking Tools: Mediapipe and Kinect

In order to support AR & VR exergames in the context of telerehabilitation, cost-effective and user-friendly motion recognition solutions are needed. As summarized in 1.1, main options are Azure Kinect [39] and Mediapipe Pose Landmarker [40]: the first is more precise in capturing the depth, but the latter does not require any further HW and is very flexible, actually giving access to gamified therapy to a wider pool of people with neurodegenerative diseases. However, these tools can be used not only in exergames, but also in more generic supervised rehabilitation programs, where important parameters such as flexions, adductions, extensions, inclinations, rotations, symmetries and repetition counts can be extracted and evaluated to keep track on patients' evolution. Some studies tried to assess the quality of measures retrieved by using MP: Latreche et al. [41] used this technology for measuring four active shoulder ranges of motion, giving proof of the reliability of mediapipe results by comparing its measures with the ones gathered via Kinect and a clinical goniometer, which respected a 95% Limit Of Agreement (LOA)[7].
Francia et al. [42] confronted inertial measurement unit (IMU)-based sensors with

[7]A statistical concept used to quantify the agreement between a new measurement method and a gold-standard reference, focusing on how much their results differ for the same observations.

Table 1.1: Comparison of markerless motion capture systems with gold-standard solutions

Feature	Azure Kinect	MediaPipe	Gold Standard (Vicon, Xsens)
Input Devices	RGB camera, depth sensor, IMU	Webcam only	Wearable IMUs, optical markers
Output	3D skeletal tracking (32 joints)	2D/3D landmarks (33 points)	High-precision full-body kinematics
Hardware Requirements	Azure Kinect camera PC	Any PC built-in webcam	Multiple sensors/cameras, high-end PC
Cost	Medium–High	Low (free, open source)	Very High
Integration	Unity, Python, C++	Unity, Python, JS	Specialized SDKs (often proprietary)
Use Case Suitability	Home/telehealth	Home/telehealth	Clinical/research lab
Tracking Precision	High	Moderate	Very high (research-grade)
Setup Complexity	Moderate (hardware, drivers)	Low (no sensors)	High (sensor placement, calibration)
Best For	Remote rehab with depth sensing	Accessible and scalable rehab	Biomechanical studies, clinical validation

an MP algorithm in reproducing the angular trends over time for four gestures. Outcomes showed how MP, despite being more imprecise and with quality differences between acquisitions among upper (16% RMSE) and lower joints (20% RMSE), could be a valid tool for tele-rehabilitation.
Mustar et al. [43] investigated with an RCT on monitoring the rehabilitation of a patient's Range of Motion (ROM) using the MediaPipe library. The developed detection system measures angles concerning the patient's body ROM patterns. The results showed an overall system error of 27%. The minimum system error was 11%.
Roggio et al. [44] used MP to evaluate applicability and reliability of a machine

learning (ML) pose estimation model for the human posture assessment, while also exploring the underlying structure of the data through principal component and cluster analyses. Outcomes pointed out gender-specific differences in hip and shoulder adduction angles, giving proof of measurements quality obtaining a ICC (Intraclass Correlation Coefficient) of 0.95 at maximum and showing that this could be potentially a new non-intrusive method for rapid, reliable screening in physical therapy and sports.
Instead, for what concerns Azure Kinect, Bertram et al. [45] confirmed its high accuracy and repeatability in measuring motor function compared to a gold-standard motion capture system. Brambilla et al. [46] assessed Azure Kinect's performance under varied environmental conditions, showing its reliability in upper limb biomechanical tracking, supporting its feasibility for tele-rehabilitation applications.
Finally, as shown in 1.3 and 1.4, the joint landmarks used for pose analysis in both technologies are illustrated. These systems have the ability to consistently reproduce and save three-dimensional coordinates of the detected landmarks, with a frequency determined by a defined frame rate (FPS). Despite differing in accuracy and accessibility, these tools open the door to scalable, home-based rehabilitation by enabling real-time motion capture without the need for expensive infrastructure, required instead by gold-standard solutions.

1.4 Objectives of the Study

The aim of this study is to perform both qualitative and quantitative analyses of data collected from five custom-designed exergames that integrate physical activity within a gamified framework. The broader goal is to assess whether, in the near future, such exergames represent a viable telemedicine solution for individuals with Parkinson's Disease (PD).

The specific objectives are the following:

- To **evaluate** the **clinical usability and effectiveness** of the proposed exergames in a pool of individuals with PD.
- To monitor the **progression of patients' performance over multiple sessions**, with the aim of identifying **collective trends** in game-based metrics.
- To investigate **correlations** between objective **game indicators** and **subjective responses** gathered through participant session surveys.
- To explore **patient-specific** relationship between UPDRS motor assessments and individualized **gameplay trends**.

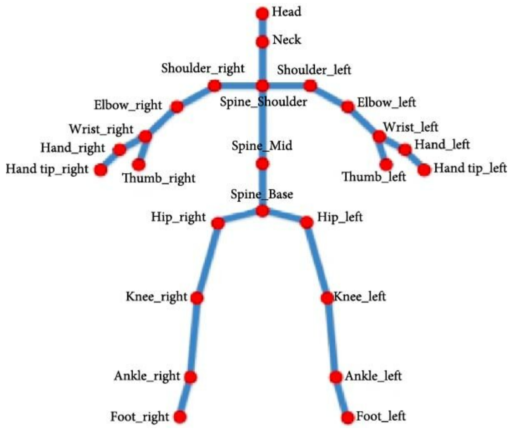

Figure 1.3: Azure Kinect joints landmarks [47]

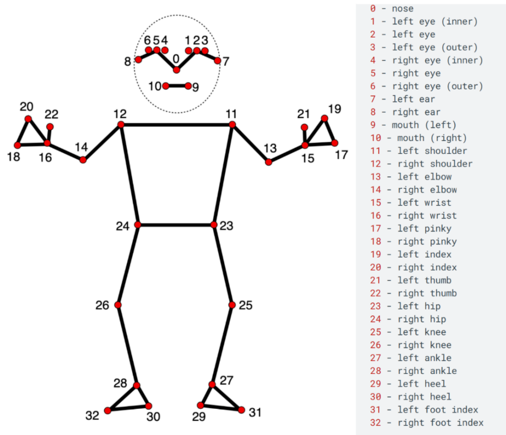

Figure 1.4: Default Mediapipe joints landmarks [40]

- To **compare** the **differential impact of each exergame** on user engagement and performance outcomes.

In order to carry on the study, data collection was an implicit, thus crucial objective.

1.5 Structure of the Thesis

This thesis is structured into five main chapters, each addressing a specific aspect of the study:

- **Chapter 1 - Introduction**: Gives a brief introduction to the context, motivation, parent project, followed by background on Parkinson's Disease, the state of art, objectives and structure of the thesis.

- **Chapter 2 - Methods**: Describes exergames and the data collection process. In addition, displays overall statistics on the acquisition, with a section dedicated to data types summarisation.
- **Chapter 3 - Analysis and Results**: Provides an overview on patients' medical attributes and on gameplay data, aggregated per patient and per session, aiming at tracking their evolution in time and trends. It evaluates alignment between game metrics and survey responses, paying attention in particular to games' max levels. Furthermore, performs correlation analysis to identify relationships among gameplay performance, UPDRS scores and survey responses-at multiple levels (globally, per game, per session and per patient). The chapter concludes with a focused study on three specific patients and presents results from the application of unsupervised learning techniques to cluster sessions.
- **Chapter 4 - Discussion**: Interprets findings in relation to clinical relevance, user experience and technological feasibility.
- **Chapter 5 - Conclusion and Future Work**: Summarizes main outcomes and proposes directions for enhancing the analytical framework and platform usability.

Chapter 2

Methods

In this section ELEVATOR platform, exergames and the collection process are described, together with a summary on total gathering time, sessions and patients. Following, a subsection is dedicated to presentation of different data types obtained. Finally HW used for the experimentation is described.

2.1 ELEVATOR Platform Presentation

As previously mentioned in chapter 1 and in the abstract, thesis' host Synarea Consultants developed four exergames (Airplane, Tile, Skewer and Ski), which, together with CNR-IEIIT's Gym, compose the set of gamified rehabilitation solutions enclosed in the ELEVATOR platform. Platform is developed as a Software as a Service (SaaS) and guarantees a cloud database backup both for game-related (movement and game metrics) and survey data. Once logged in, the operator has the possibility to start a new session with a new/already registered patient or to access the global or per participant results dashboard (Figures 2.1, 2.2).

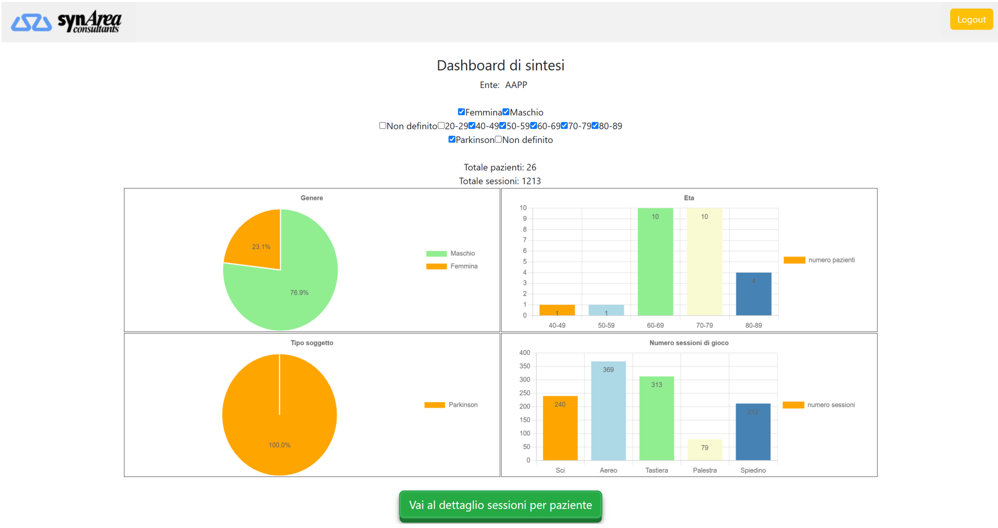

Figure 2.1: Dashboard with global results. On the left side, pie charts for sex and disease. On the right, bar plots for age demographic analysis and total number of attempts per game. Above visualization of data, there are filters for selection of age, sex and disease.

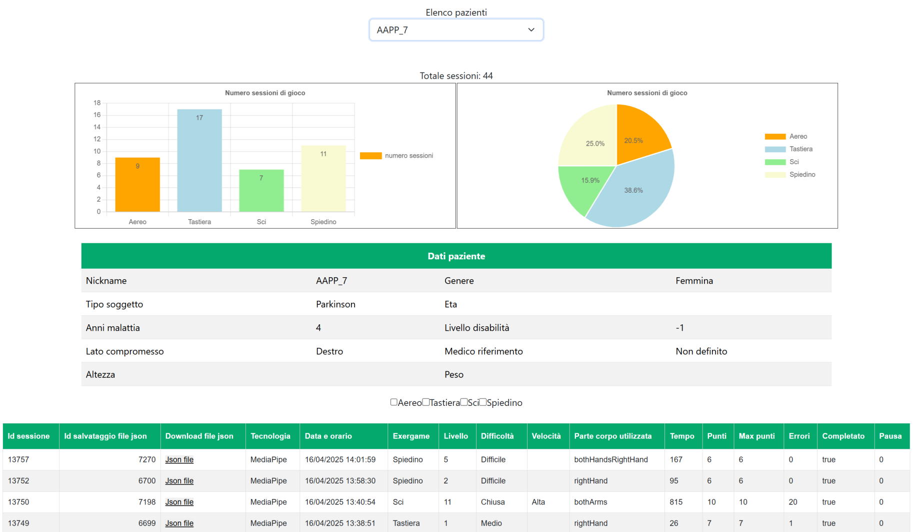

Figure 2.2: Dashboard with patient-specific results and data retrieval. Above, pie chart with fraction on the total of specific exergames attempts and bar plot with total number of tries per game. Below, main patient features are displayed, with possibility to check and download single attempts data.

After patient selection (Figure 2.3), a pre-test questionnaire has to be compiled (Figure 2.4) in order to gain access to the game selection interface (Figure 2.5).

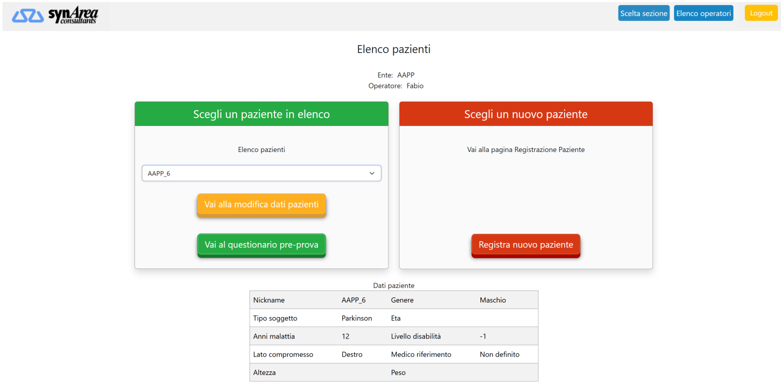

Figure 2.3: Interface for patient selection or registration, accessible only by registered operators. Below selection, patient key features are displayed.

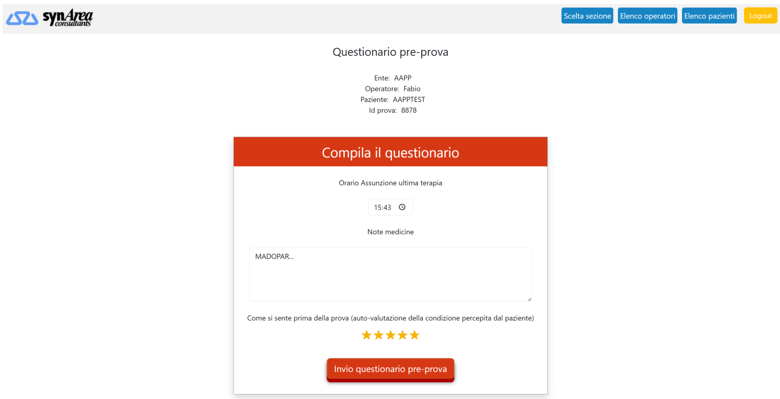

Figure 2.4: Pre-test questionnaire. Required information regard last Levodopa-based medicament assumption, description of the medicine (mg, name, frequency...) and a self-evaluation for the perceived actual condition.

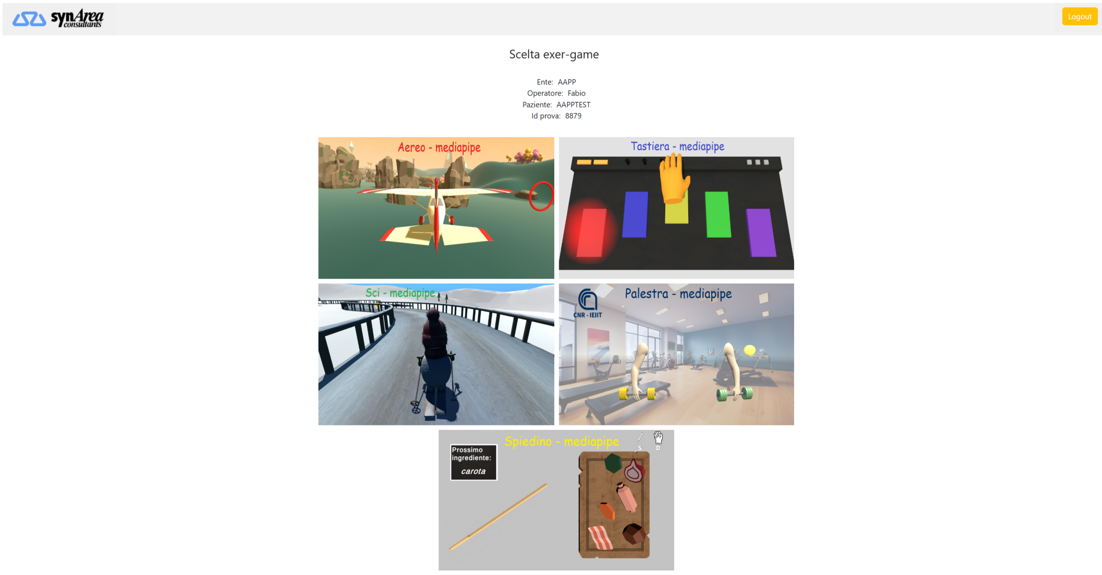

Figure 2.5: Game selection interface with the 5 exergames considered for the study.

At the end of the session, a post-test questionnaire (Figure 2.6) was asked as well to keep track of subjective impressions regarding the games, considering the variety of difficulties proposed across the path.

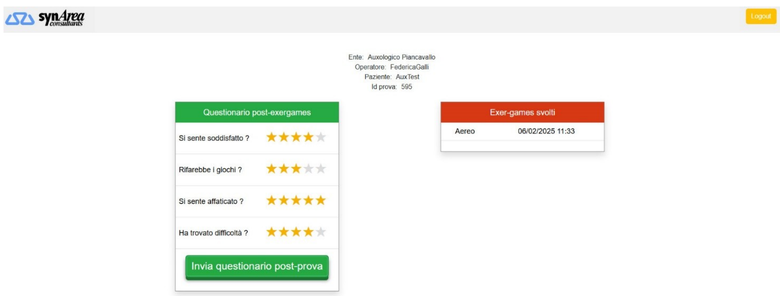

Figure 2.6: Post-test questionnaire. Likert-type scale answers regard satisfaction, repeatability, fatigue and difficulty. On the right, all session attempts are listed.

2.2 Exergames Description

The exergames are presented individually, detailing the required actions, the motor or cognitive functions targeted, the game level and scoring logic. Each game supports semi-autonomous activation through a gesture-based selection system, eliminating the need for a mouse or keyboard. This was made possible by the integrated hand-tracking system and an intuitive, user-friendly Human-Machine Interface (HMI).

2.2.1 Airplane - SynArea Consultants

A 3D scene in which the participant controls an airplane using coordinated movements of the torso and arms (Figure 2.7), which simulate the function of wings. The objective is to avoid obstacles and fly through colored rings (from which score is computed) distributed along the path. Difficulty levels are determined by the aircraft's speed settings, as well as the number and spatial arrangement of the rings. In the more advanced levels, the color of each ring determines whether the airplane should fly through it or avoid it. Passing through an incorrect ring results in a point deduction.

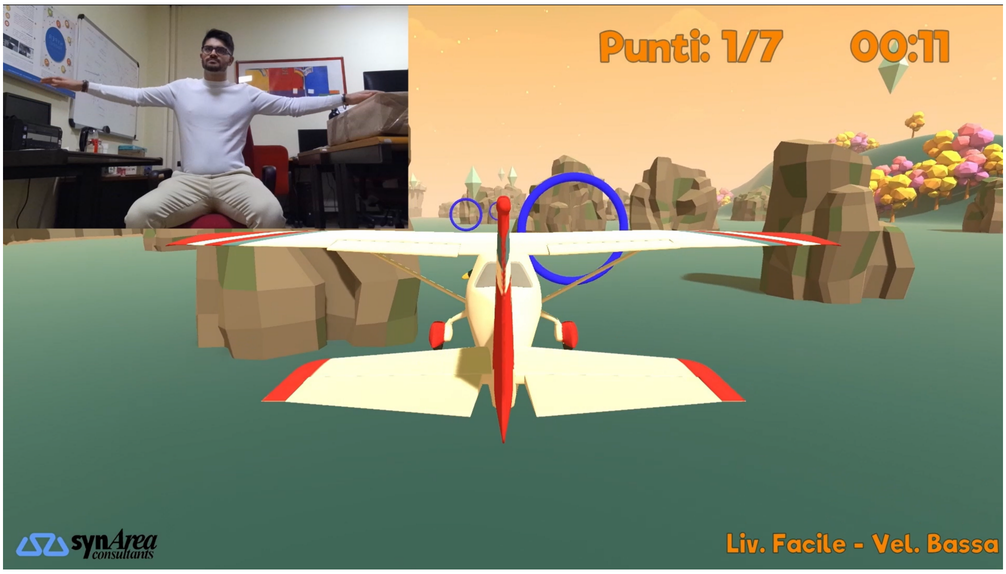

Figure 2.7: Airplane exergame demonstration.

Game level is computed as follows:

$$\texttt{GameLevel} = (\texttt{level difficulty} \times 3) + \texttt{level velocity}$$

where `level difficulty` and `level velocity` spans from 0 to 2, defined as:

- 0 = Easy
- 1 = Medium
- 2 = Hard

From a cognitive perspective, this exergame engages several key functions. The dynamic 3D environment stimulates *selective and sustained attention*, requiring the participant to stay focused on relevant stimuli while maintaining performance over time. In advanced levels, the color-dependent decision-making activates *working memory*, *inhibitory control* and *cognitive flexibility*, as players must recall and adapt to rules in real-time. Additionally, the spatial navigation task enhances *visuospatial reasoning* and *processing speed*, supporting both cognitive training and motor-cognitive integration in rehabilitation contexts.

2.2.2 Tile - SynArea Consultants

In order to press tiles, this exergame is designed to stimulate upper limb activity through frontal extension and positioning movements (Figure 2.8). These actions promote motor control, coordination and movement precision, while also engaging more complex motor patterns, upper limb strength and postural alignment of the head and trunk.
In addition to its physical goals, the game targets cognitive functions, particularly *sustained and selective attention*. The participant must remain focused on executing the correct movements while selecting the appropriate keys, which are randomly prompted by the system to compose a melody. This dual-task structure supports motor-cognitive integration and encourages engagement through musical feedback. Game levels and scores reflect the total amount of tile to press (5, 7 or 10) to complete the exercise.

Figure 2.8: Tile exergame demonstration.

2.2.3 Skewer - SynArea Consultants

Game's target is to compose a skewer following a certain order of ingredients (Figure 2.9). The primary functional objective of this exergame is to stimulate upper limb motor coordination, either unilaterally or bilaterally, in combination with grasping and pinching actions. These movements promote fine motor control, coordination and precision in reaching for and manipulating virtual objects. The game leverages Google MediaPipe's hand-tracking capabilities, which allow for real-time detection of hand landmarks via a standard webcam.
In addition to motor stimulation, the exergame engages several cognitive functions. The participant must maintain *selective attention* to identify and interact with relevant targets, while *visuomotor integration* is required to align hand movements with visual cues. A key cognitive component involves selecting the correct ingredient displayed on a virtual blackboard: this task activates *working memory*, *semantic processing* and *decision-making*, as the participant must recall the target item and distinguish it from distractors. The grasping and pinching actions further reinforce *executive functions* such as planning and response inhibition, particularly under time constraints or when incorrect selections result in point deductions. While scores goes up to the number of ingredients inserted in order (maximum value is 6), game level is computed as follows, resulting in six distinct levels ranging from 0 (easiest) to 5 (most challenging):

$$\texttt{GameLevel} = (\texttt{level difficulty} \times 2) + \texttt{hands number}$$

where `hands number` takes values 0 (single mode) or 1 (both hands mode) and `level difficulty` ranges from 0 to 2, corresponding to the values:

- 0 = Easy
- 1 = Medium
- 2 = Hard

Generally, within each session, both different hand modes were proposed to the participant, keeping the same difficulty level.

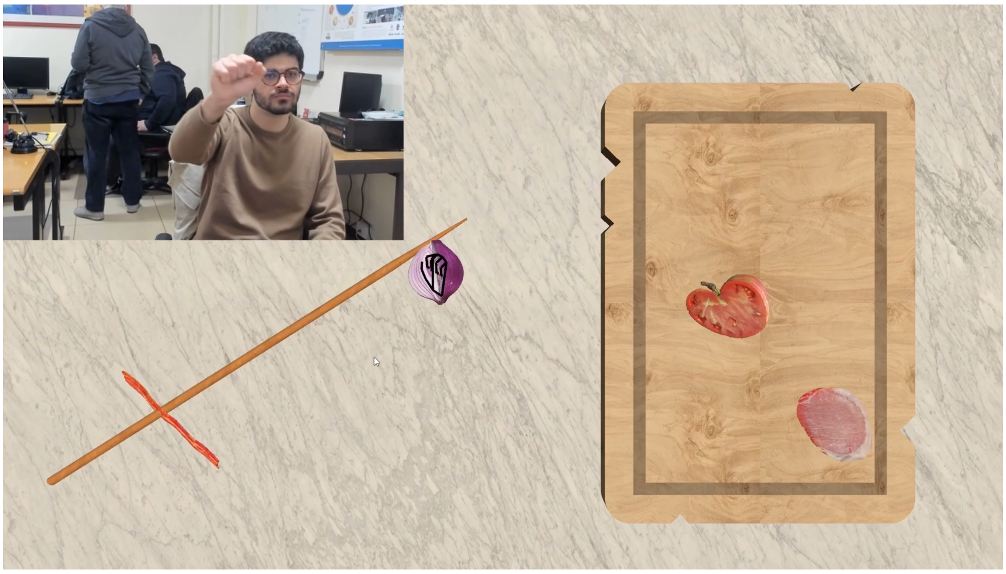

Figure 2.9: Skewer exergame demonstration.

2.2.4 Ski - SynArea Consultants

This game, designed to simulate skiing across varied tracks (Figure 2.10), is the only one in which both upper and lower limbs are actively engaged. The activity involves rhythmic, alternating stepping movements of the legs combined with coordinated arm lifts. Despite its simplicity, the exercise targets multiple rehabilitative domains, including motor coordination, muscular strength, mobility and postural trunk balance.
In addition to its physical goals, the exergame also engages several cognitive functions. The participant must maintain *sustained attention* and *motor-cognitive coordination* to execute the correct movement pattern while navigating the track. The need to avoid obstacles and collect coins as rewards introduces elements of *anticipatory planning*, *visual scanning* and *decision-making under pressure*. These demands promote cognitive flexibility and reinforce the integration of perception, action and reward-based motivation.
Difficulty levels are implemented by introducing tracks with increasingly complex layouts, requiring greater precision, timing and cognitive control to navigate successfully. Game level is computed as follows:

$$\texttt{GameLevel} = \texttt{level velocity} + (\texttt{track difficulty} \times 3)$$

where `track difficulty` ranges from 0 to 3, corresponding to the values:

- 0 = Straightaway

- 1 = Curve
- 2 = S-turn
- 3 = Closed track

`Level velocity` spans from 0 to 2, defined as:

- 0 = Easy
- 1 = Medium
- 2 = Hard

Figure 2.10: Ski exergame demonstration.

2.2.5 Gym - CNR-IEIIT

This exergame is designed to stimulate upper limb motor function through exercises inspired by common gym routines. The activity consists of a series of arm lifts-performed either unilaterally (one arm at a time) or bilaterally (coordinated arm movements), executed in both lateral and frontal directions.
The game includes the following sequence of exercises:

- Repetitions of right-arm lifts
- Repetitions of left-arm lifts
- Repetitions of alternating arm lifts (starting with the right arm)

- Repetitions of simultaneous arm lifts

In order to complete the game, each of the four exercises must be performed in both the lateral (arms lifted to the sides) and frontal (arms lifted forward) modalities. In addition to its physical objectives (targeting motor coordination, strength and range of motion) the exergame also engages cognitive functions. The participant must maintain *sustained attention* and *sequencing ability* to follow the prescribed order of exercises and switch between movement types. The need to recall and execute specific movement patterns also stimulates *working memory* and *motor planning*, supporting the integration of cognitive and physical rehabilitation goals. Scores are assigned as follows:

- `Correct Repetition` = +100

- `Wrong Repetition` = -25

A four-level scale is implemented, where the number of repetitions increases in increments of five-from a baseline of 5 up to a maximum of 20.

Figure 2.11: Gym exergame demonstration.

2.3 Deployment of the Pilot Study within the Association

As summarized in Table 2.1, the collection process involved the coordination among multiple actors: the AAPP-AIP coordinators, the CNR-IEIIT, the SynArea Consultants employees, the neurologist and the patients. This process lasted from the start of November 2024 until Easter festivities of 2025, with plenty of time devoted to identifying participants, planning sessions and accommodating individual needs. The pilot study consisted of two phases:

Table 2.1: Project Timeline Overview

Date	**Task Description**
5 Nov 2024	Web meeting with the Association to define thesis aspects, especially the exergame testing
10 Nov 2024	Project divided into two phases: feasibility study and data collection
10 Jan 2025	Visit to the Association to introduce the activities and prepare an info document for potential participants
13 Jan 2025	Drafting message to circulate among AAPP-AIP professionals to recruit Phase 1 volunteers
17 Jan 2025	Drafting newsletter for AAPP-AIP members
20 Jan - 10 Feb 2025	Meetings with volunteers for Phase 1
11 Feb - 4 Mar 2025	Planning weekly calendar to identify patients for Phase 2 (Tab 2.2)
5 Mar 2025	Project presentation at AAPP-AIP (Figure 2.12) and participant clinical assessments by a neurologist
10 Mar 2025	Start of Phase 2
18 Apr 2025	End of Phase 2

- **Phase 1**: First feasibility study and search for ideal target
- **Phase 2**: Data collection organized in gamified sessions

2.3.1 Phase 1 of Pilot Study

This phase involved the search for heterogeneous participants (see upper graphs from Figure 2.1), willing to try the five exergames, which, at that stage, still relied on Azure Kinect technology. For this part of the study, volunteers were asked to participate only once and to answer the same pre & post questionnaires described in Figures 2.4 and 2.6. 26 subjects showed up to contribute, among which there were cases of DBS, early and late PD stages and a person confined to wheelchair. The primary goal was to perform a usability study to assess the range of patients' disease stages for which the difficulty level of the exercises is suitable, enjoyable and effective. This was done considering game performances as well as by collecting feedback, impressions and suggestions directly from the players. Outcomes were used to better align the development of the in-game activities with the needs and preferences expressed by the patients. Secondary purpose was to select the best responsive volunteers, both in terms of game performances and survey evaluations, to invite them to participate in the second phase. Findings indicated that the gamified exercises were best suited for individuals in Hoehn and Yahr stages two to three (see chapter 1 subsection on the scale).

2.3.2 Phase 2 of Pilot Study

The last phase of the study, which lasted for more than a month, was dedicated to the data collection. Gamified sessions were divided in 1-hour slots with frequency of one or two times per week, depending on participant availability. Among the 26 patients visited in phase 1, 14 of the most responsive participants agreed to take part in this phase. Data collection consisted not only in gathering game performance data and survey responses, but also joint motion capture and video recordings of main motor UPDRS tasks, asked to be performed before and after each session. The collection made in this phase was later used as unique base for the data analysis described in section 3.

Table 2.2: Example Weekly Schedule: One-Week Snapshot (subject to modifications)

Time	Monday	Tuesday	Wednesday	Thursday	Friday
7:00-8:00	6				
8:00-9:00	1		18	18	
9:00-10:00	5		21	14	13
10:00-11:00	2		17	20	5
11:00-12:00	10		19	2	14
12:00-13:00	20		7	22	7
13:00-14:00	22		6		
14:00-15:00	17				
15:00-16:00	13		10		
16:00-17:00	21				

Figure 2.12: Presentation to the patients in AAPP-AIP.

Overall Collection Statistics

In phase 2, a total of 123 sessions were held and personally supervised, corresponding to a total of at least 123 hours of data collection. The total number of tries is reported in Figure 2.13.

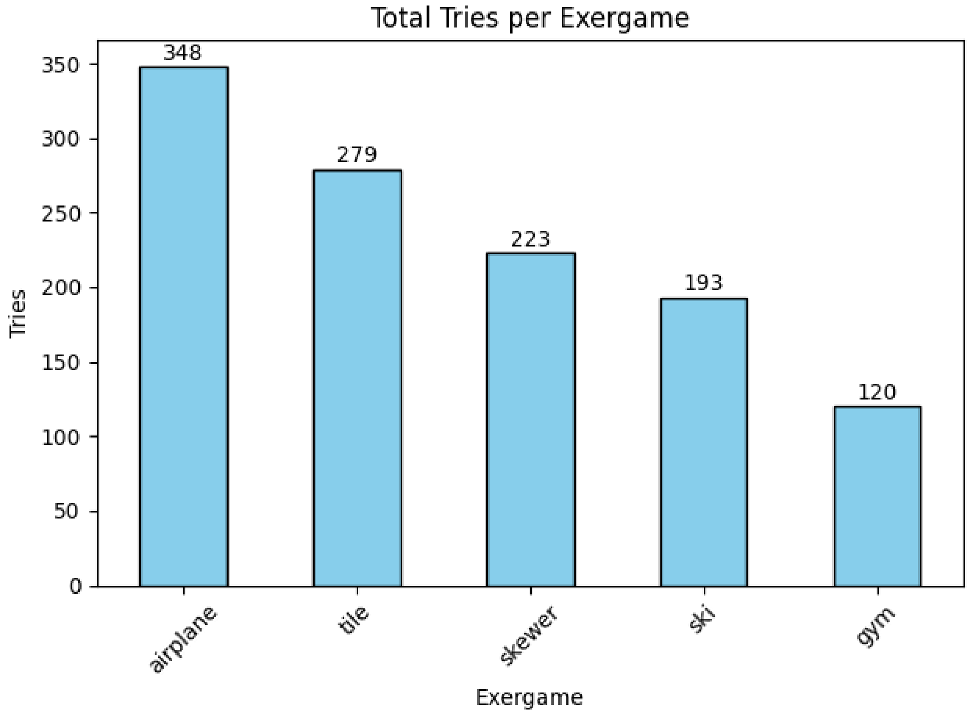

Figure 2.13: Total number of tries for each exergame.

Notably, the game with the fewest trials was Gym, which was proposed only once per session due to its longer duration.

2.3.3 Data Types and Ranges

The data collected during the pilot study can be categorized into four main types: **game data**, **video recordings**, **questionnaire responses** and **clinical and demographic data**. Each type was stored in standardized formats for ease of analysis, reproducibility and cross-reference.

Game Data

Game sessions generated two distinct data streams stored in `.json` format:

- **Motion data**: Frame-by-frame tracking of joint positions using 3D Cartesian coordinates (x, y, z) per joint. These data support kinematic analysis and movement quality assessment.

- **Game results**: A series of integer-based scores reflecting user performance across different exergame sessions.

Video Recordings

Two types of video files were recorded throughout the work:

- **Exergame attempts**: Gameplay recordings that could be used in the future to qualitatively assess user engagement, motor strategies and any discrepancies with tracked motion data. The recorded sessions may be reprocessed by third parties to extract motion data from the video material without requiring patients to undergo the rehabilitation protocol again.

- **UPDRS tasks executions**: Videos captured participants via smarphone performing selected items of the MDS-UPDRS scale, recorded both before and after each session for clinical comparison. Tasks included Sit-To-Stand, Leg Agility, Gait (from two perspectives, frontal and horizontal), Finger Tapping, Toe Tapping and Hands Movement.

Questionnaire Data

Self-reported questionnaires were presented to participants pre and post sessions. Responses used a Likert-type scale[1] ranging from 1 (lowest) to 5 (highest) and were saved in `.csv` format for analysis.

Clinical and Demographic Data

Additional data was compiled in structured files:

- **Patient information:** Collected in a `.csv` file, including demographic data and anonymized participant identifiers.

- **Clinical scores:** Stored in `.xlsx` format, incorporating Hoehn and Yahr staging and itemized UPDRS assessments filled out by a neurologist.

[1] psychometric response format commonly used in questionnaires to measure attitudes, opinions, or perceptions. It present a graded scale, where 1 means "strongly disagree" and 5 means "strongly agree".

2.4 Hardware Setup for Data Collection

This section describes the hardware setup adopted for the data collection throughout the pilot study. Two distinct configurations were used for Phase 1 and Phase 2, reflecting the technological transition during the course of the project.

Phase 1 - Azure Kinect Setup

In the first phase of the pilot study, the exergames were still integrated with the Azure Kinect system. Participants interacted with the system through Kinect's built-in RGB-D camera. No external webcam was used during this phase.

Phase 2 - Webcam and Display Setup

For Phase 2, the Azure Kinect HW was replaced by Mediapipe SW. A Logitech BRIO 4K Stream Edition webcam was used for high-resolution video recordings and to enable real-time body tracking for game inputs, with key features including:

- Ultra HD 4K resolution (recorded at 1280×720px, 30 fps)
- Automatic light correction with HDR technology
- Omnidirectional microphones with noise cancellation
- Adjustable field of view (FOV) at 65°, 78° and 90°

In both phases, laptop used to support exergames, body tracking and video recording had the following specifications:

- **Operating System**: Windows 11
- **Processor**: Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz
- **RAM**: 16 GB
- **GPU**: NVIDIA GeForce GTX 1650

To enhance the user experience and enable real-time feedback, a secondary 50-inch monitor was installed to display games.
Participants were seated facing the television, approximately two meters away, in a well-illuminated position with a white wall behind them. The designated posture remained consistent across all games when considered individually. Specifically, the setup for Airplane, Ski and Tile was identical: the webcam was positioned horizontally, aimed directly at the participant to ensure full-body framing (see Figure 2.14). For the Gym game, the indicated posture required the subject to stand. During Skewer, the participant was moved closer to the screen (approximately one meter away) and the camera was tilted and mounted on top of the television.

Figure 2.14: Video frame of a patient performing Airplane exergame.

Chapter 3

Analysis & Results

This chapter details the data analysis process implemented after the completion of Phase 2 data collection. Before presenting results, a brief overview of the data preparation pipeline is provided, followed by a general profile of the participants. The analysis is organized into key areas of interest: game results (with emphasis on game level to point out progressions), survey responses, correlation heatmaps (linking game metrics, questionnaire data and clinical evaluations), special patients analysis and unsupervised sessions classification. Special patients analysis consists into the inspection done on three representative patient profiles, designated as the “least severe”, “most severe” and “outlier”. These subjects data is explored in comparative depth. Each area of investigation (excluding clustering) is analyzed on a per-game basis by aggregating data first by patient, then by session.

3.1 Data Preparation

This section outlines the data preparation process, including initial organization, cleaning and preprocessing steps required to produce a consistent and analyzable dataset.

3.1.1 Data Organization

As described in chapter 2, multiple data types were used in Phase 2 analysis, which necessitated an initial organization into structured folders, followed by conversion into standard formats such as `.csv` files to enable compatibility with Python DataFrame libraries.

Game Data

The built-in auto-save process generated a hierarchical folder structure for each game, summarized in Figure 3.1.

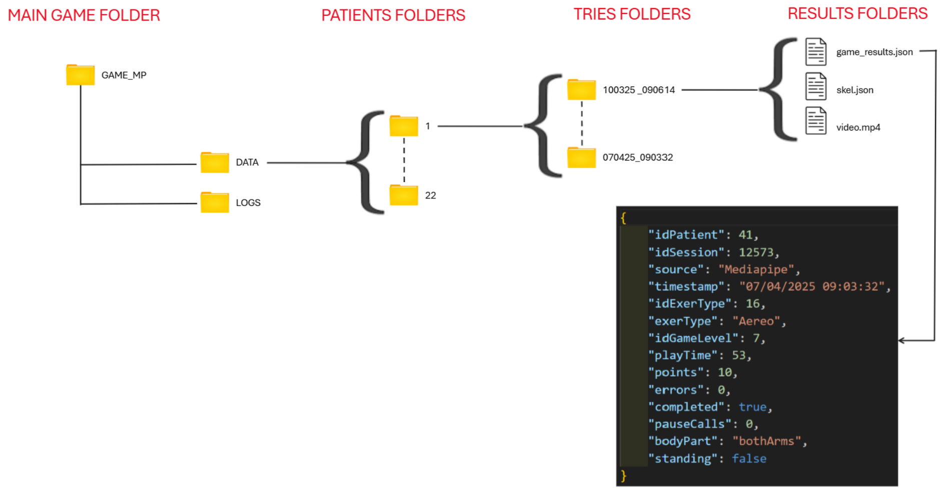

Figure 3.1: Auto-generated hierarchical structure of game data.

Noticeably, no session information was explicitly present in final `game_results.json` files, which were all divided and with the wrong ID mapping. In addition, repeated gameplay attempts were stored as folders sharing the same date but differing by timestamp (hour and minute). The first step in organizing data involved creating a `GAME_session_summary.json` file for each exergame, containing the session summaries for that specific game: primary key was the patient's ID and value was a dictionary having as key the session number and as value the list of folders indicating the different tries. Each patient's folders were iteratively accessed to compute game metrics and generate preliminary `.csv` result files, storing all tries tabular data.

UPDRS Tasks Video Recordings

Since videos were recorded with a smartphone, proper filing was required after acquisitions, ensuring that motor assessment clips were properly sorted by patient ID and session type (pre/post). Files were later saved on company cloud for future usage, as not analyzed as part of the present study. Clinical scores, patient descriptions and survey answers did not require any additional organization step, being already well separated and, apart from medical results, already in `CSV` format.

3.1.2 Data Cleaning & Preprocessing

The cleaning process primarily targeted game results, with a focus on session and attempt integrity checks: only complete, uncorrupted and uninterrupted records were considered valid for inclusion in the final dataframes. For all exergames, columns that were meant to be numeric were converted into appropriate numeric types (integers or floating-point values).
Session number indicators were standardized to ease future merging of data, while timestamps were converted into python-readable datetime objects.
The Gym exergame scores were normalized to align them with the range of other game metrics, minimizing distortion in visual analyses.
Questionnaire answers were subjected to a renaming process, replacing numerically labeled columns (such as `answer_2`) with semantically meaningful attribute names, followed by a feature selection phase where non-relevant metrics were discarded.
Patient and clinical evaluation databases were subjected to the same transformation pipeline as the surveys, ensuring consistent naming conventions and formats across shared columns.
Merging was a crucial step to put together exergames' results into a single complete dataset, using as key fields the patient ID and the session number. An 'exergame' identifier column was added to preserve row origin. Due to heterogeneity in metrics across games, only a subset of universally relevant features was retained for cross-game analysis.
These are: **game time**, **game level**, **score** and **errors**. In particular for the tile game, a **tile time** field obtained by dividing the game time by the number of points (and thus tiles pressed) was included in the analysis but not stored as a column in the merged database.
Patients and clinical results were merged together on the patient ID and the result was also merged with survey data, keeping the session number indicator.
Finally, this DataFrame was merged with the output of the combination of games data, obtaining a fully informative starting point for more advanced researches in later stages.
All cleaning and merging operations were performed using custom Python scripts developed for this study. Analyses ware conducted by aggregating merged databases by patient ID or session number, obtaining different points of view of the results evolution process and knowledge on best performers.

Software Tools

All data processing, cleaning and analysis tasks were performed using **Python 3.10**. The workflow leveraged several open-source libraries and modules, including:

- **pandas**, **csv**, **json**, **pathlib**: For tabular data manipulation, I/O operations and directory traversal.

- **datetime, functools, itertools, re, collections**: For time management, functional operations and data structuring.

- **NumPy, SciPy**: For numerical computations and geometric analysis.

- **matplotlib, seaborn**: For data visualization and custom plot generation.

- **scikit-learn (sklearn)**: For preprocessing, clustering (DBSCAN, KMeans, Agglomerative) and evaluation metrics (silhouette score).

3.2 Patients Presentation

Before diving into the analysis, a general presentation of the patients is given in Table 3.1, while main clinical evaluations from neurologic assessment are shown in Figure 3.2 and 3.3.
Participants ranged in age from 54 to 82 years, with total UPDRS scores spanning 8.5 to 62.
Despite the apparent heterogeneity of participants, thirteen out of fourteen participants were classified within Hoehn and Yahr stages 2 or 3 (stage 2: $n = 8$, stage 3: $n = 5$). The only patient labeled as stage 4 was intentionally included as the most severe case among the admissible subjects selected in phase 1. Purpose was to explore his responsiveness to gamified therapy over a prolonged period of time to prove effectiveness, despite the tougher conditions.

Participant Demographics

Patient ID	Age	Years with PD	Most Compromised Side	Sex
1	54	24	RX	M
2	80	4	RX	M
5	66	3	RX	M
6	81	12	RX	M
7	75	4	RX	F
10	74	5	RX	M
13	65	10	RX	M
14	66	3	LX	M
17	78	9	RX	M
18	75	7	RX	M
19	82	2	NO	M
20	75	4	RX	M
21	67	4	RX	F
22	67	15	RX	M

Table 3.1: Phase 2 patients: demographics and most compromised side (RX = right, LX = left, no = no compromised side).

Clinical Profile

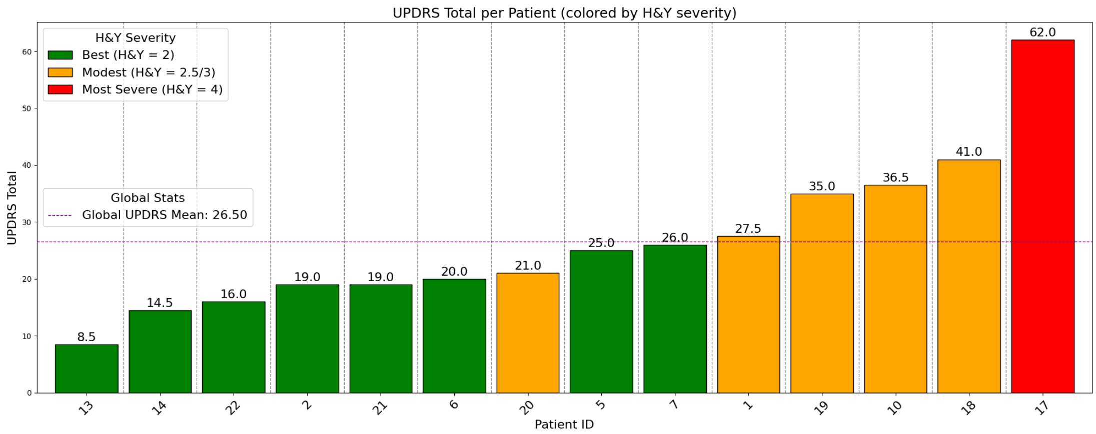

Figure 3.2: Total UPDRS values of patients, colored by H&Y scale.

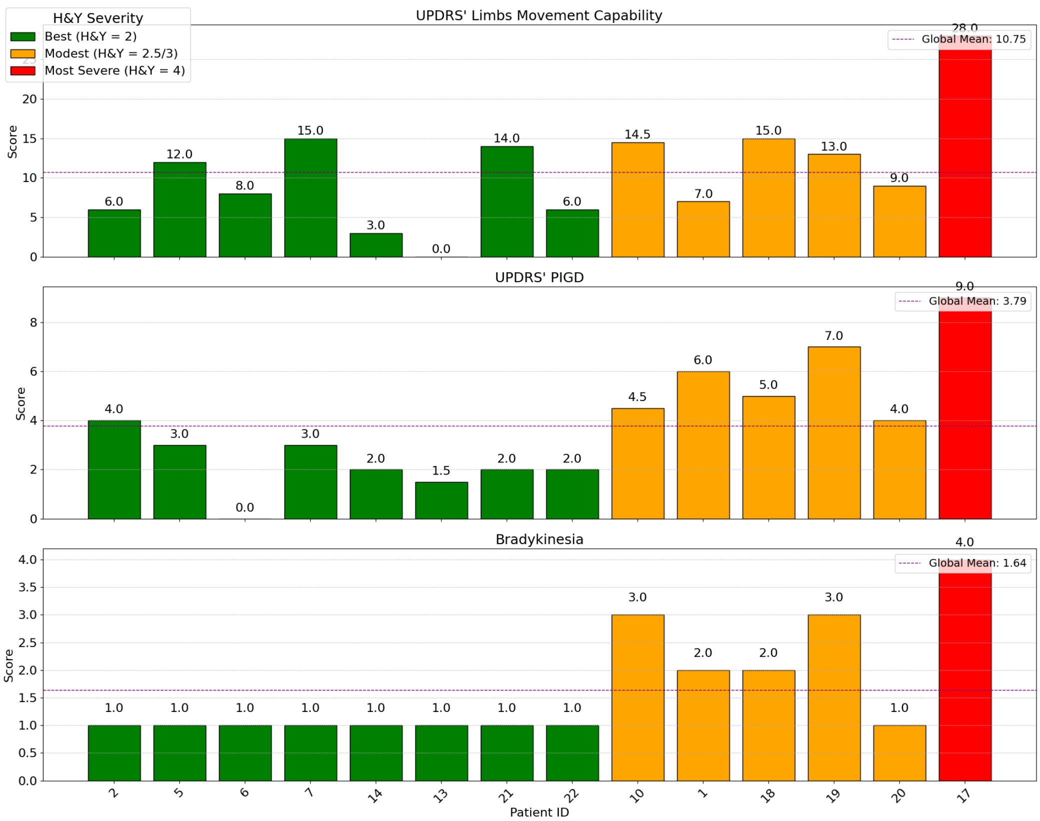

Figure 3.3: UPDRS' main motor features values per patient, colored by H&Y scale.

3.3 Game Results Analysis

A subset of the available exergame metrics was selected to highlight the most relevant indicators for performance assessment. Results were first categorized by aggregation type, then analyzed by game. Chosen indicators were:

- **Airplane**: *score* and *game level*, chosen to assess participants' adaptability and progression across difficulty tiers.
- **Tile**: *time to press a tile* and *errors*, indicators of reaction speed and cognitive-motor coordination under pressure.
- **Skewer**: *game time*, used to quantify endurance, accuracy and velocity.
- **Ski**: *game level*, *game time* and *errors*. They reflect adaptability to a dual-task challenge involving obstacle avoidance and timed execution.
- **Gym**: *game level* and *score*, which summarize task completion and in-game success across multiple motor tasks. Score values were normalized to percentages to align with other metrics.

3.3.1 Per Patient Aggregation

This section aims to identify the top-performing patients based on aggregated performance metrics across sessions. In order to do so, each exergame's DataFrame was grouped by patient ID, computing the overall mean, standard deviation and maximum values for all metrics previously listed. For each metric, a summary DataFrame was generated containing one row per patient, ordered by descending mean value of that metric. These summaries were used to generate Matplotlib-based bar charts, where for each patient:

- The **bar height** represents patient's average performance
- The **gray shaded area** shows the ± standard deviation
- The **black triangle marker** indicates the maximum value achieved

A **dashed red line** shows the global mean across all session records of all patients. The standard deviation was calculated using the `.std()` method from `pandas` library, which quantifies the dispersion of a set of n values $x_1, x_2, \ldots, x_n$ from their arithmetic mean $\bar{x}$:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.1}$$

where:

- s is the sample standard deviation,
- $\bar{x}$ is the mean of the observations,
- x_i represents each individual measurement,
- n is the total number of observations for that patient.

Airplane

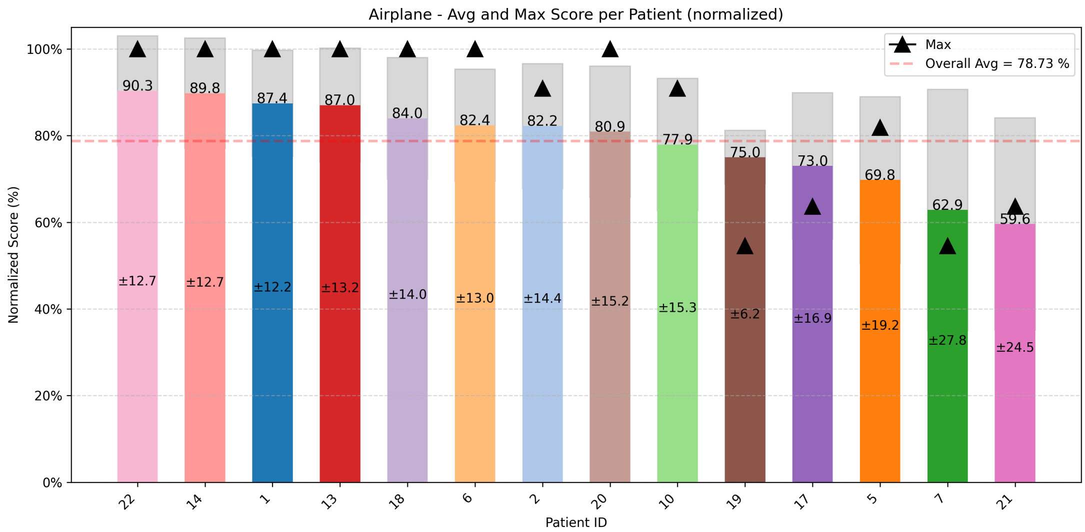

Figure 3.4: Average and Maximum normalized Score obtained by patients across sessions (expressed in percentages).

In figure 3.4, initial insights into patient consistency can be drawn by examining the standard deviations (std), maximum and means normalized score. Normalization enables evaluation of the ratio between actual and maximum scores, penalizing subjects that reached lower levels at their best, since higher levels bring possible higher scores and thus larger percentages. In particular, subjects with higher normalized scores and lower standard deviations (patients from number 22 to 18) demonstrate more stable and efficient motor-cognitive performance across sessions. Participants with moderated standard deviation and near-average mean scores have a variability consistent to session to session changes in motor and cognitive aspects. Patients with high standard deviations and low mean scores—such as 17, 5 and 7—are the most inconsistent: their performances may spike in some sessions and drop in others. In addition, case of participant 19, with low std and mean, is an example of a consistent poorly performer. Maximum level analysis reveals that

50% of patients reached the global maximum score, achievable with three most engaging game levels marked as "difficult". Interesting are the cases of patient 2 and 5. The first, despite being among the top performers, never hit the global maximum score, while the latter, although being the one with the second lowest mean score, has obtained at least once a quite high result.
In Figure 3.5, per patient game level progression for Airplane exergame is shown.

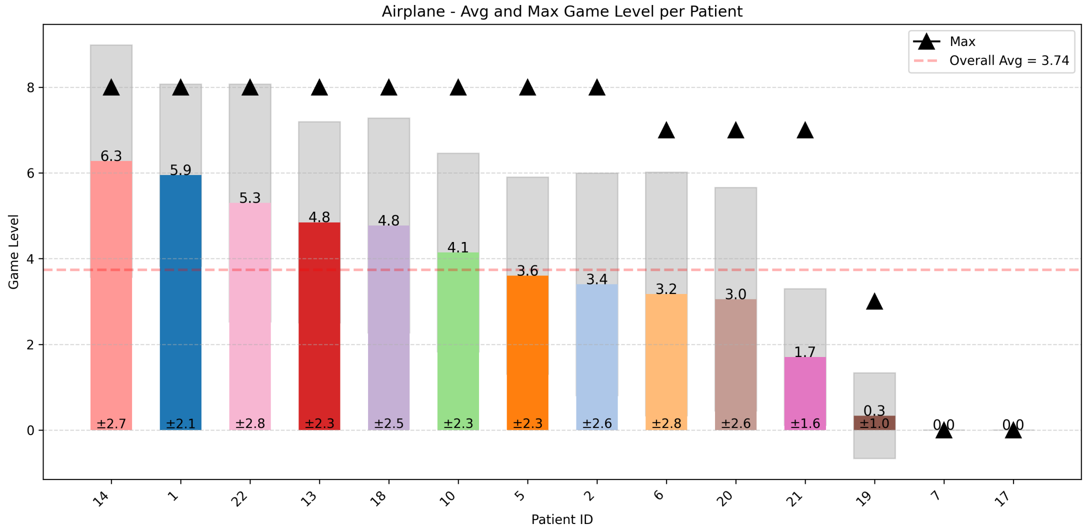

Figure 3.5: Average and Maximum Level obtained by patients across sessions.

Top performers (which are the same outlined for 3.4) are patients with an average value 1 unit higher than the mean level. All of them reached the absolute max level, including patient 5 and 2, which have below overall mean average game level, indicating an important later sessions improvement. Remaining patients never reached maximum level, with subjects 19, 7, 17 that found early a difficulty plateau that could not overtake. Despite the differences in level reaching, the relatively uniform standard deviation indicates that study group members were all similarly consistent in their behavior: game's difficulty progression produced a balanced cognitive-motor challenge across the participants, making this metric a reliable discriminator for differences in capability and adaptation.

Tile

By observing the average and maximum times to press a tile in Figure 3.6, a clear understanding of participants reaction times can be achieved. Both mean and maximum results cover a wide performance spectrum: patients ranged from the fastest patient 14 (3.5 seconds as average) to patient 19 (slowest, almost 14 seconds

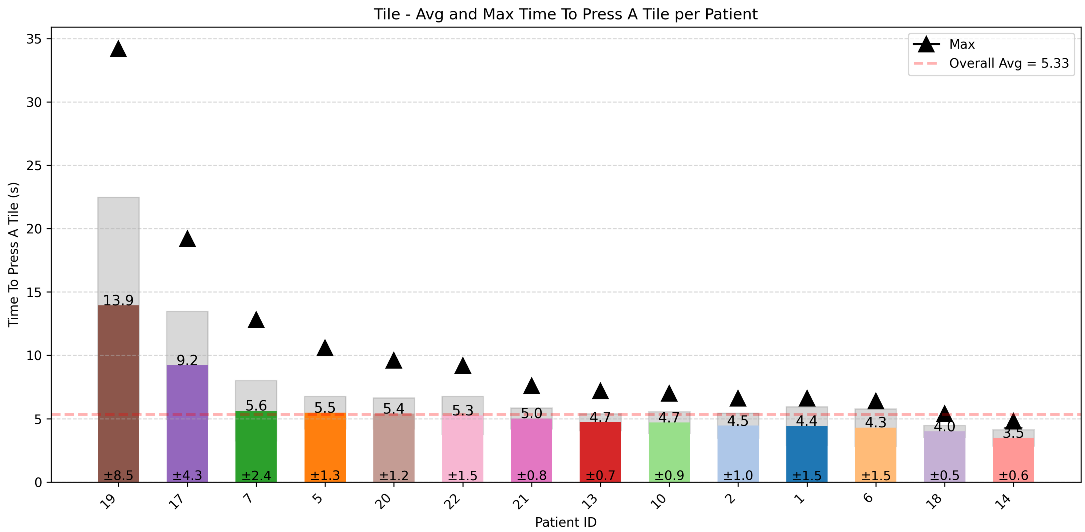

Figure 3.6: Average and Maximum Time to press a tile obtained by patients across sessions.

as average). The slowest patients (19 and 17) also show the largest standard deviations (±8.5 seconds and ±4.3 seconds), consistent with pronounced motor impairment evaluations. All participants, excluding the least reactive 17 and 19 and the quickest 14 and 18, had as average values numbers close to the overall mean, with deltas between the range of half a second. Fastest participants mentioned before were also consistent in their performances, with very low standard deviations. Figure 3.7 reveals several notable patterns in error distribution through patients.

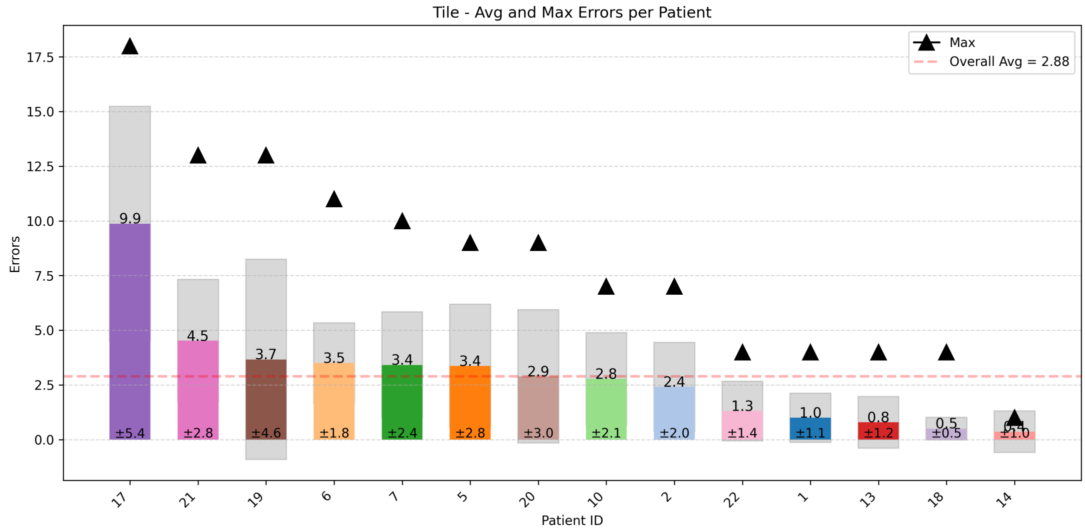

Figure 3.7: Average and Maximum number of Errors obtained by patients across sessions.

Error spectrum is wide: maximum values range from more than 17 to less than 1, while mean values vary from 10 to 1. Patient 17, despite being one of the slowest in pressing tiles (see previous analysis), is also the one who makes the most mistakes, validating the clinical outcomes that ranked this subject as the most severe. Patient 6 instead, which was one of the fastest, is also one of the least precise. Finally, all patients beside patient 14 have had a game try where they made more errors with respect to the overall average number. Subject 14 instead can be considered as the top performer: not only the fastest, but also the most consistent and precise, without any sessions exceeding the average error threshold.

Skewer

Figure 3.8 illustrates the variation in Skewer's gameplay duration across participants, resulting in a large spectrum of values: times shift from the highest time execution attempt (more than 11 minutes and a half) performed by patient 19 to the lowest

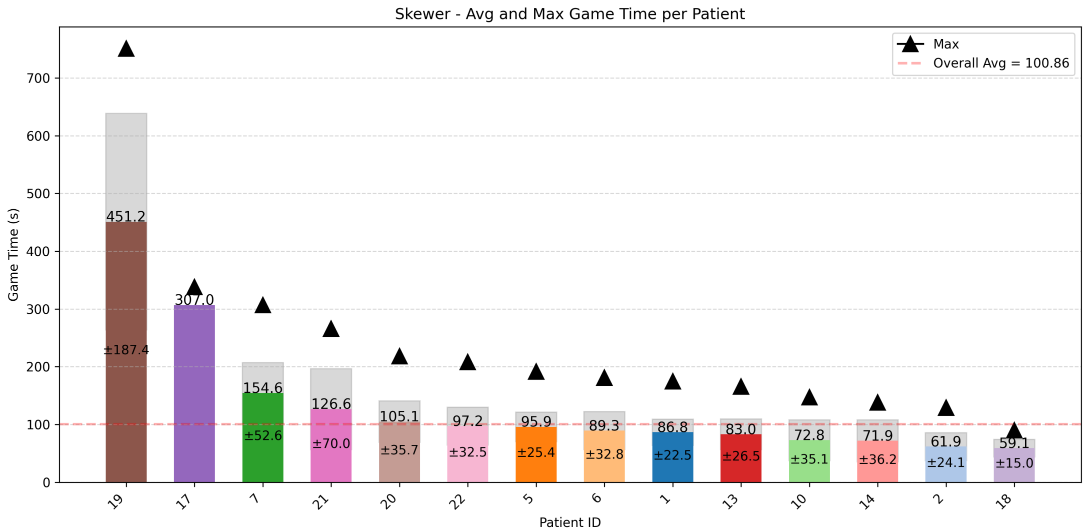

Figure 3.8: Average and Maximum Game Time obtained by patients across sessions.

maximum time execution, performed by patient 18 in just under 100 seconds. Important is to notice that as maximum value is intended the worst performance obtained by a patient. Least quick subjects are 17 and 19, the same outlined in the tile game investigation, while the fastest are 2 and 18. Notably, the standard deviation trend indicates that slower patients are also the least consistent, while most speedy ones are more continuous.

Ski

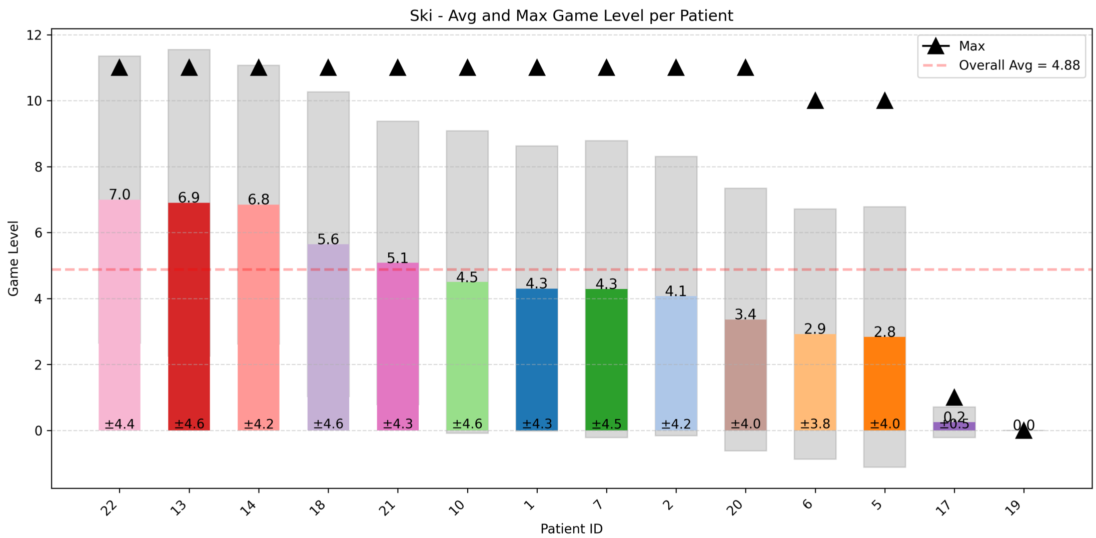

Figure 3.9: Average and Maximum Level obtained by patients across sessions.

Ski's game level related indices visualization in Figure 3.9 allows for several key observations. Firstly, overall mean level is less than one unit below half maximum level: this validates the difficulty of the exergame. If we consider in addition the maximum game level reached by patients, it is noticeable that 10 out of 14 patients reached global higher number, meaning that progress was slow and mainly occurred during the last sessions. Only patients 5, 6, 17 and 19 failed to reach higher difficulty mode, with last two struggling to escape the least challenging ones. These subjects are the same outlined in previous games metrics as more in difficulty. Besides these participants, all the others showed an uniform standard deviation, similarly to the Airplane game. The challenge was engaging also for this exergame, leading to the conclusion that these two interventions were the most difficult and arduous ones. Patients 22, 13 and 14 distinguished themselves the most, in complete accordance with their UPDRS evaluations.
Figure 3.10 illustrates a wide range of game durations across patients: metrics

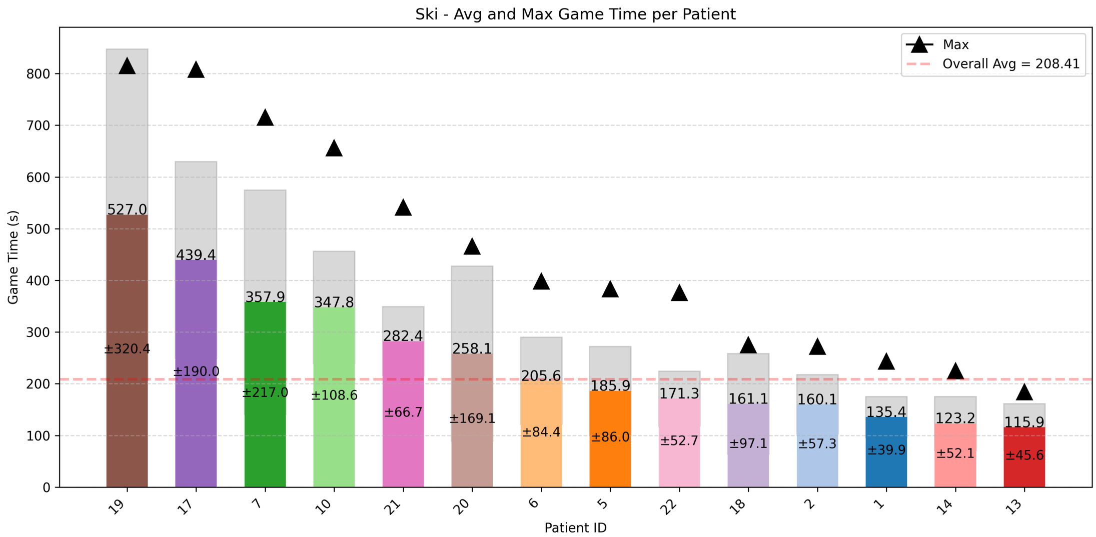

Figure 3.10: Average and Maximum Game Time obtained by patients across sessions.

vary from patient 19 values (maximum of more than 13 minutes, mean close to 9 minutes) to the ones of patient 13 (maximum try lasted nearly 3 minutes, while mean value is 2 minutes). Notably, the std variation across patients is directly proportional to the self mean game time. Patients 7, 10 and 21, differently from subjects 13 and 14, despite being among the best in game level reached, needed more time in average to complete the executions. Finally for the Ski plots by

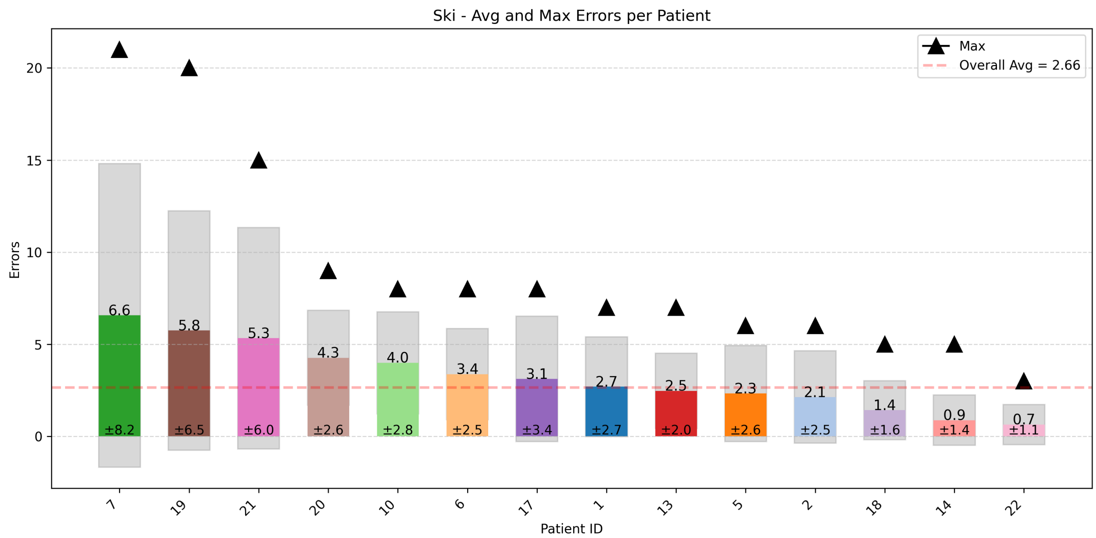

Figure 3.11: Average and Maximum number of Errors obtained by patients across sessions.

patient aggregation, error metric distribution is visualized in Figure 3.11: also in this case, errors ranges widely both in mean value (from 6.6 to 0.7) and max absolute value (from 20 to 3). Patients 7, 19 and 21 had the most fluctuating executions. In particular, subject 21, despite being an above-average performer, had prolonged and error-prone trials. Patients 14, 18 and 22 committed fewer errors on average with respect to other participants, despite playing hardest levels and in shortest times. These, together with patient 13, are the Ski top performers.

Gym

Figure 3.12 presents patient-level distributions for the Gym exergame. The std appears to be uniform across all subjects, meaning that they reached a difficulty plateau in a limited number of tries. This plateau was reached by only 3 out of 14 patients, since by considering the absolute maximum level completed, all the remaining ones executed the exergame at highest degree of difficulty. Best performers include those with average values above the overall mean: these are, as

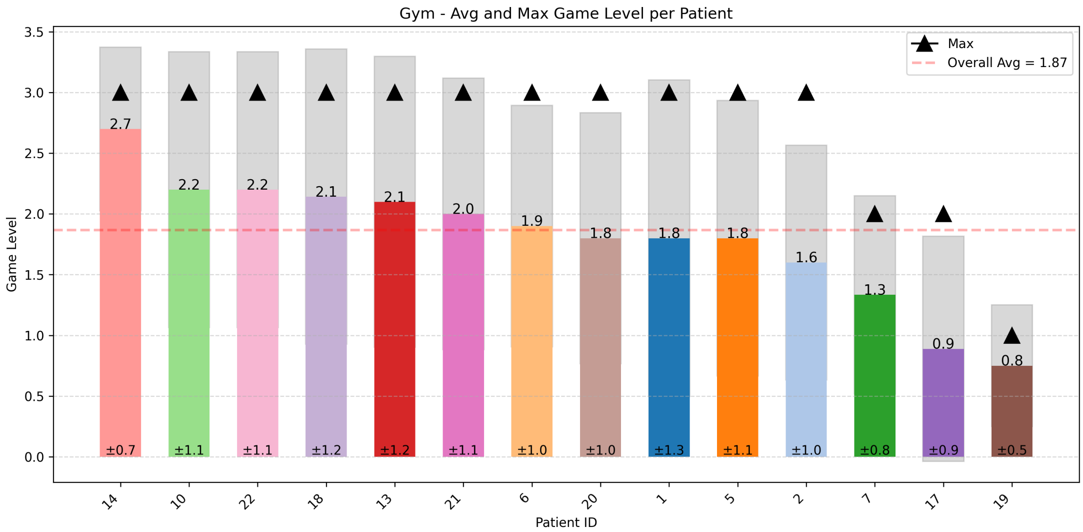

Figure 3.12: Average and Maximum Level obtained by patients across sessions.

for the majority of previous analyses, number 13, 14, 18, 22, with the addition of patient 10 and 21. Additionally, analysis of Figure 3.13 on normalized total score,

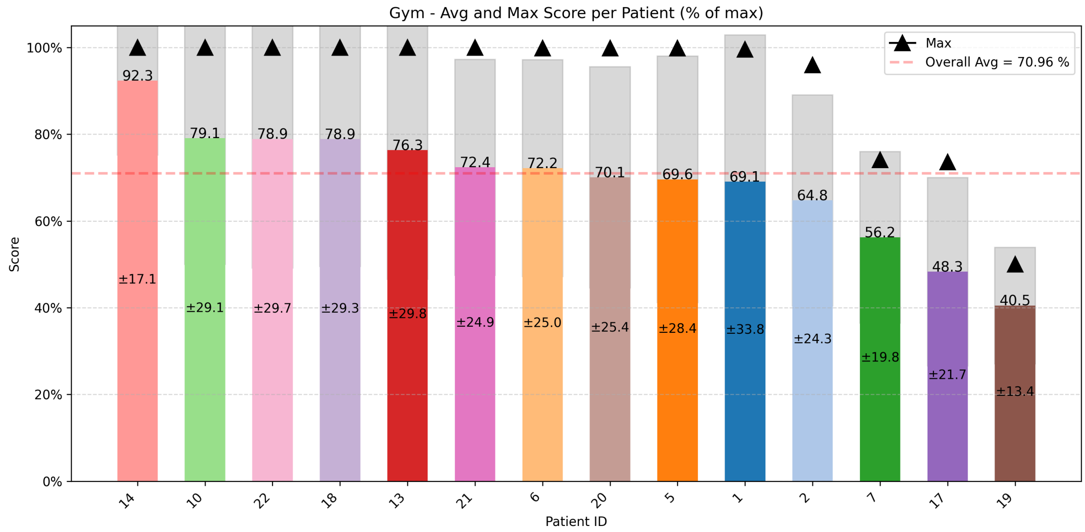

Figure 3.13: Average and Maximum Score obtained by patients across sessions.

gives evidence that score is heavily correlated with game level: as seen in Chapter 2, `+100` points were assigned for a correct repetition, while `-25` was the penalization for a wrong movement. Higher level means more repetitions, thus more potential total points. This is why best performers are the same outlined for game level.
Despite clinical conditions, patient 18 resulted one of the top performers in almost all metrics analyzed in this section. From now, in all the following inspections, it will be referred as the "outlier" patient. In the final investigations, particular attention will be paid to this subject, together with the "best" and "most severe" patients, respectively number 14 and 17.

3.3.2 Per Session Aggregation

This analysis focuses on the progression of performance across rehabilitation sessions, aiming to uncover trends and variability in the selected metrics over time. To achieve this, each exergame's DataFrame was grouped by session number. For each metric, the **mean**, **standard deviation** and **maximum value** were computed across all participating patients.
The resulting session-level summaries were used to generate Matplotlib bar charts, structured as follows:

- The **bar height** represents the mean value of the selected metric across all patients for that session.

- The **gray shaded rectangle** visualizes one standard deviation ($\pm\sigma$) around the mean.
- A **black triangle marker** denotes the maximum value reached for that metric in the corresponding session.

Additionally, a **dashed red line** indicates the global average across all session entries in the dataset for that metric.

The selected metrics for each exergame remained consistent with the per-patient analysis.

Airplane

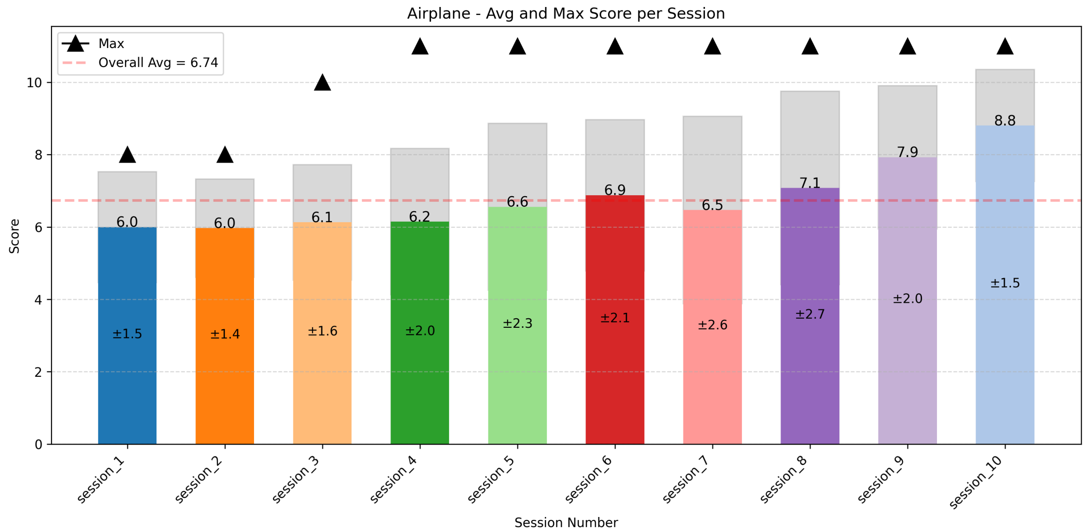

Figure 3.14: Average and Maximum Score obtained in sessions across patients.

Considering Airplane's visual display of score metric illustrated in Figure 3.14, it is possible to see the global patient evolution of performances. The first session where at least a patient was able to complete an execution with maximum score was the fourth. Standard deviation values increased by more than one unit from session 4 onward, suggesting that substantial performance variability was present among patients, with subject stuck on lower levels or on higher degrees of difficulty modes but unable to reach elevated scores. Considering average scores, it is possible to see that globally means tend to rise directly proportional with session number: sessions up to number 6 yielded average scores below the global mean, while successive sessions (excluding session 7) exceeded it.

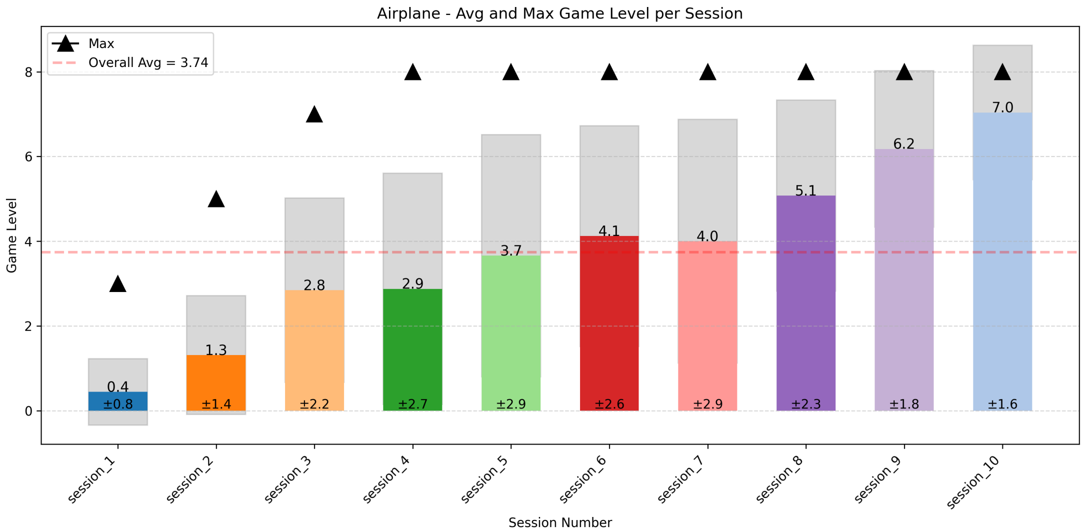

Figure 3.15: Average and Maximum Level obtained in sessions across patients.

Figure 3.15 clearly illustrates a learning curve and performance improvement trend of patients across sessions for level: ceiling score was first reached in session 4, while mean metric values increase from start to end. Standard deviation is lower on first and last two sessions, where majority of individuals was either learning how to deal with the game in easier modes or either on most challenging scenario. The proximity of the overall mean to half the cap level suggests the game posed a balanced challenge.

Tile

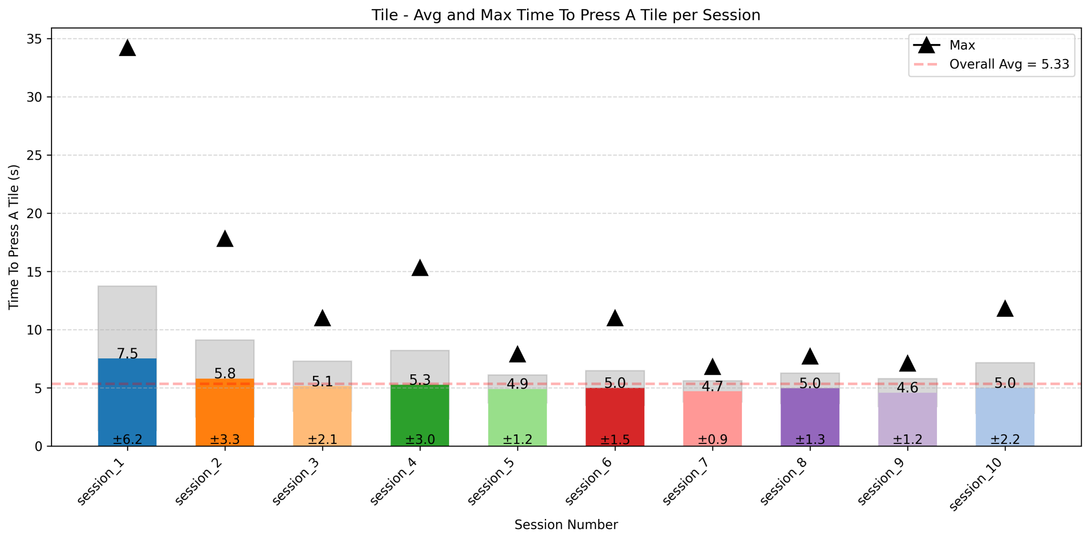

Figure 3.16: Average and Maximum Time to press a tile obtained in sessions across patients.

Studying time to press a tile metric plot in Figure 3.16, insights into the nature of the tile game can be derived. Peak session values indicate that first session was the hardest in general for patients, where they were still trying to figure out how to play the exergame. Until session 6 included, there were patients taking above average times to press a tile, while, from next one, mean and maximum session values where close to the overall average. Exception is last session, possibly reflecting late-stage adaptation among slower learners reaching higher difficulty levels. Mean values correspond with the observed oscillations in standard deviation. Considering errors in Tile game, aggregated per session, showed in Figure 3.17, it is noticeable how mean error numbers were close to the overall mean for all sessions except the last, slightly lower. Maximum error number per session were high compared to other games, ranging from 7.5 up to 17.5: the Tile game appears

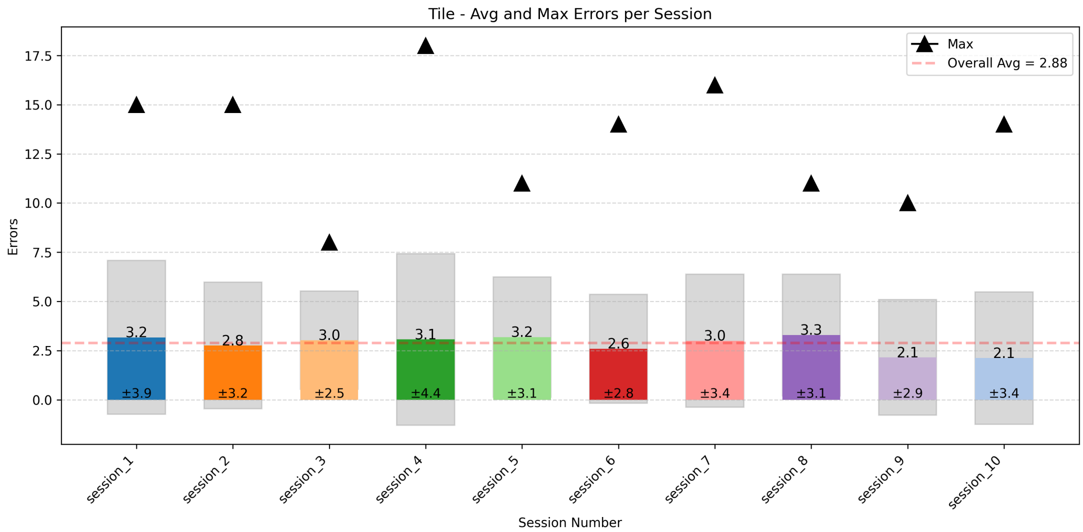

Figure 3.17: Average and Maximum number of Errors obtained in sessions across patients.

to demand the highest degree of precision among all five exergames. Also, the std was quite uniform across all sessions.

Skewer

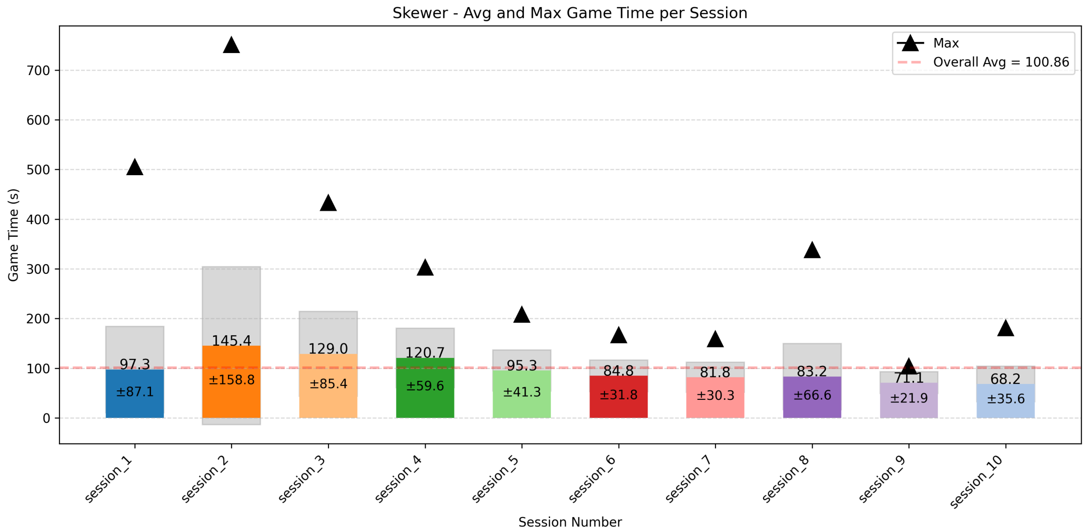

Figure 3.18: Average and Maximum Game Time obtained in sessions across patients.

Figure 3.18 shows trends in game time performance for the Skewer exergame. It reveals a negative trend involving all sessions above the second for the mean values and all session included between 2 and 7 for the peak intra-session times. The final three sessions were excluded due to outlier performances involving unusually slow executions. Trend described begins from when patients started dealing with the two hands mode. Standard deviation followed the same evolution path.

Ski

Digging into Figure 3.19's display of Ski levels across sessions, an initial inter-session progression trend is clearly observable. Standard deviation increases up to session 6, after which it declines in proportion to session number. This implies a global improvement of performances in later tries, with a gradual increase in the metric value that differs significantly intra-session, depending on patients' ability. First two sessions can be considered of initial setup for the game, with mean values close to the baseline scores. Although upper-bound level is reached from fourth session, the overall mean (below half peak value for game level) indicates that this exergame was considered above-average as difficulty, with intra-session discrepancies between

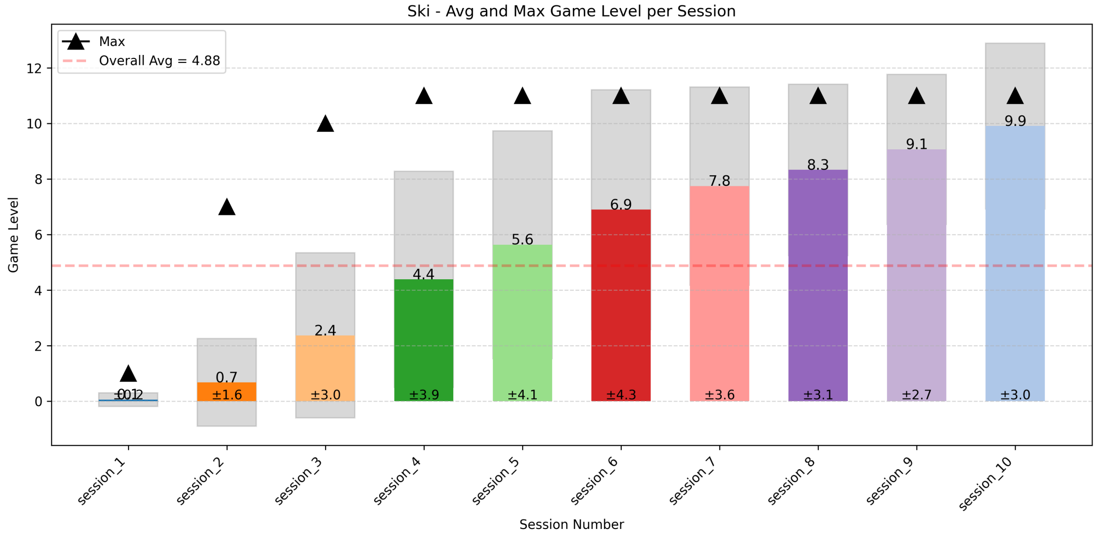

Figure 3.19: Average and Maximum Level obtained in sessions across patients.

patients.

Figure 3.20 exhibits consistent game time performance across sessions, with mean

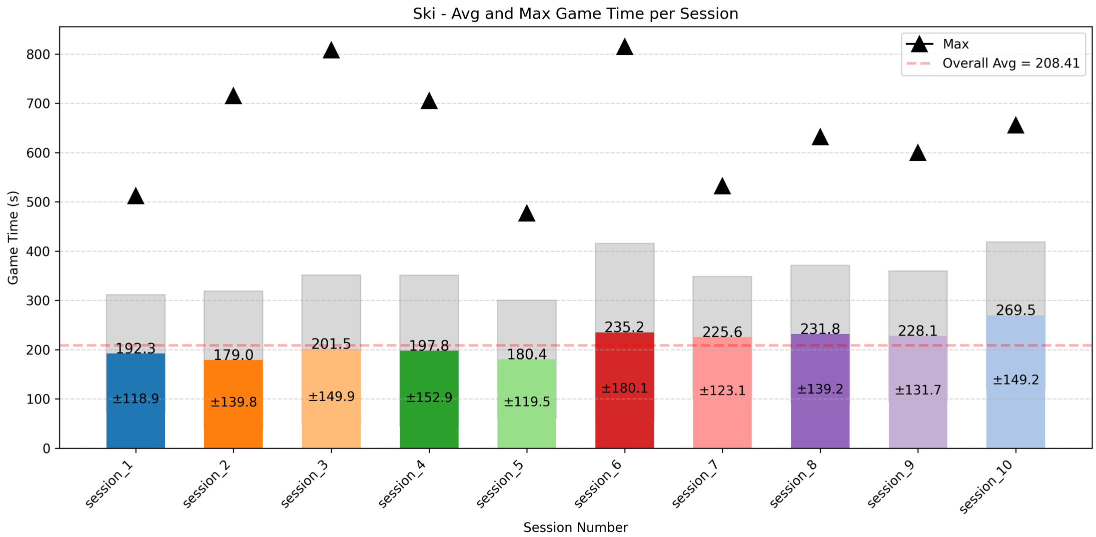

Figure 3.20: Average and Maximum Game Time obtained in sessions across patients.

session values close to overall average and a stable but high standard deviation across them, with values near or higher than 2 minutes. Considering that higher levels required longer execution times due to the track length and shape, these results indicate the patients' overall ability to deal with the exergame, following a capped 10-sessions trial-and-error process.

Analysis of error metrics in Figure 3.21 shows how, despite the increasing difficulty of the game throughout sessions, the quality of the execution remained overall steady. Notably, the values for session 6 were above the overall mean, supporting the hypothesis of a correlation between execution time and error frequency. Inter-session mean error is close to 3, which is a good result considering the presence of worst attempts with up to 20 errors. Standard deviation is stable across all sessions, highlighting the variability in patient performance throughout this exergame.

Gym

Gym's level plot showing aggregated per session metrics in Figure 3.22 suggests that this exergame is more suitable for later-stage progression: maximum level is reached starting from session 3 and average metric values per session are above overall mean since session 5, with a general plateau from session 7, where mean value is

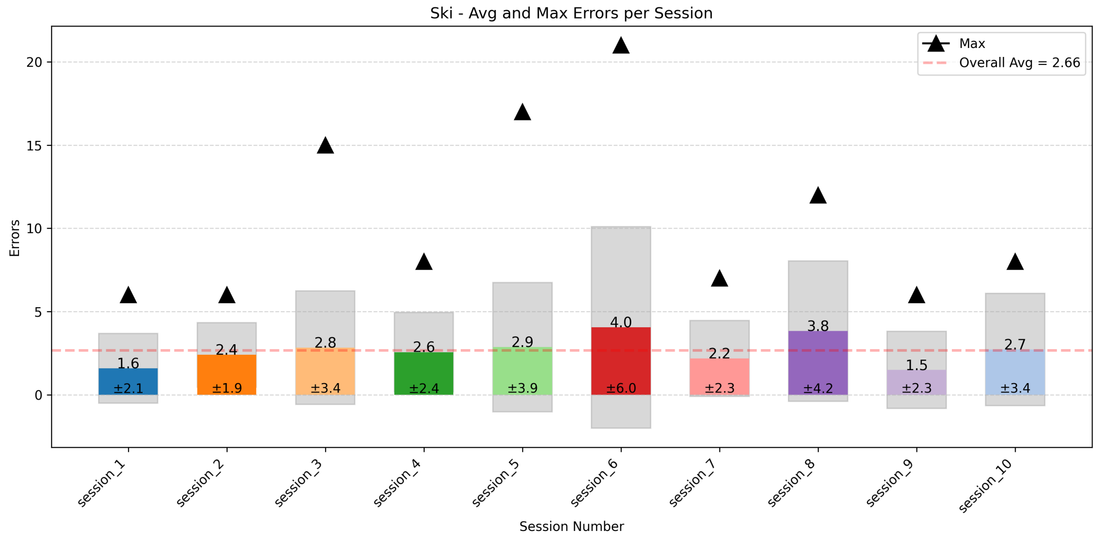

Figure 3.21: Average and Maximum number of Errors obtained in sessions across patients.

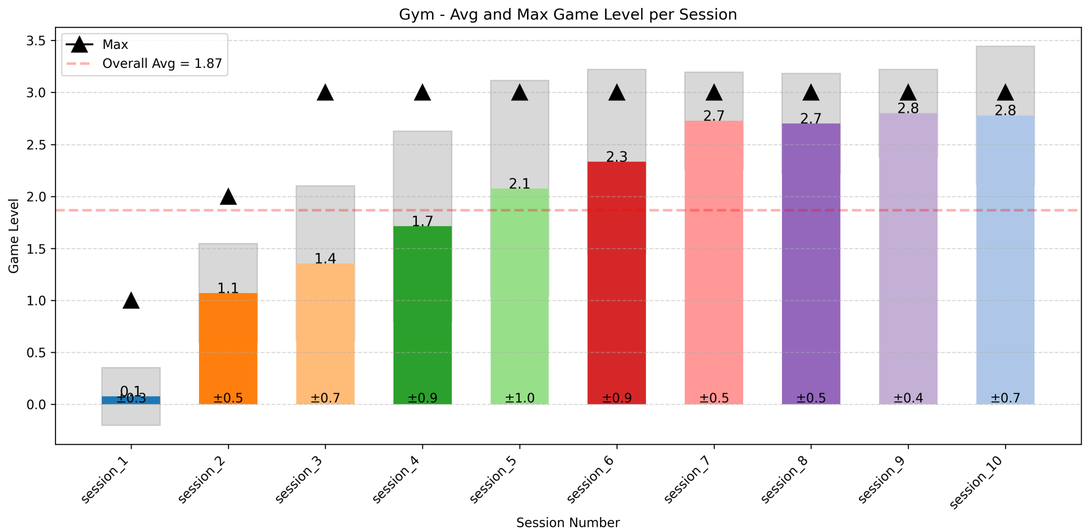

Figure 3.22: Average and Maximum Level obtained in sessions across patients.

close to the upper bound. Standard deviation is steady and low throughout all sessions, reflecting the ease with which most participants reached higher difficulty levels.

Figure 3.23 reinforces the strong correlation between score and game level in Gym,

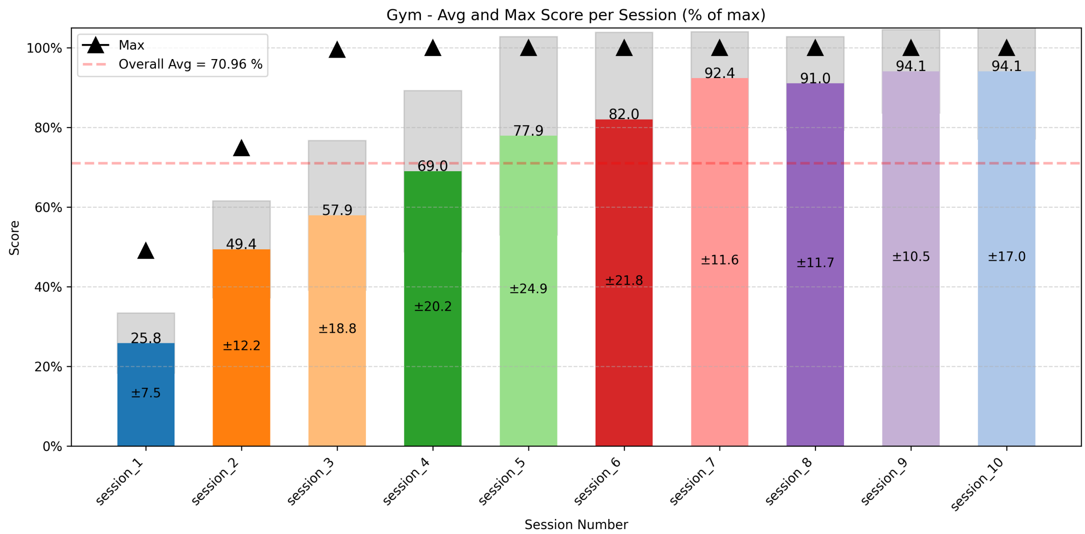

Figure 3.23: Average and Maximum Score obtained in sessions across patients.

as previously described in the per-patient aggregated analysis and supports the same conclusions observed in the level-based evaluation.

3.4 Maximum Level Study

As stated before, special focus was dedicated to the game level metric. This can be used to track patients progression in exergames throughout sessions, in order to:

- Find better and worse performing patients.
- Prove the learning ability of participants, analyzing inter-session differences.
- Rank the difficulty of an exergame.

3.4.1 Patients Ranking based on Maximum Level

This investigation relies on prior game-level analyses done on exergames' game level plots, aiming at finding relationships between game level reached and clinical evaluations. In the binary matrix presented in Figure 3.24, patients were sorted by

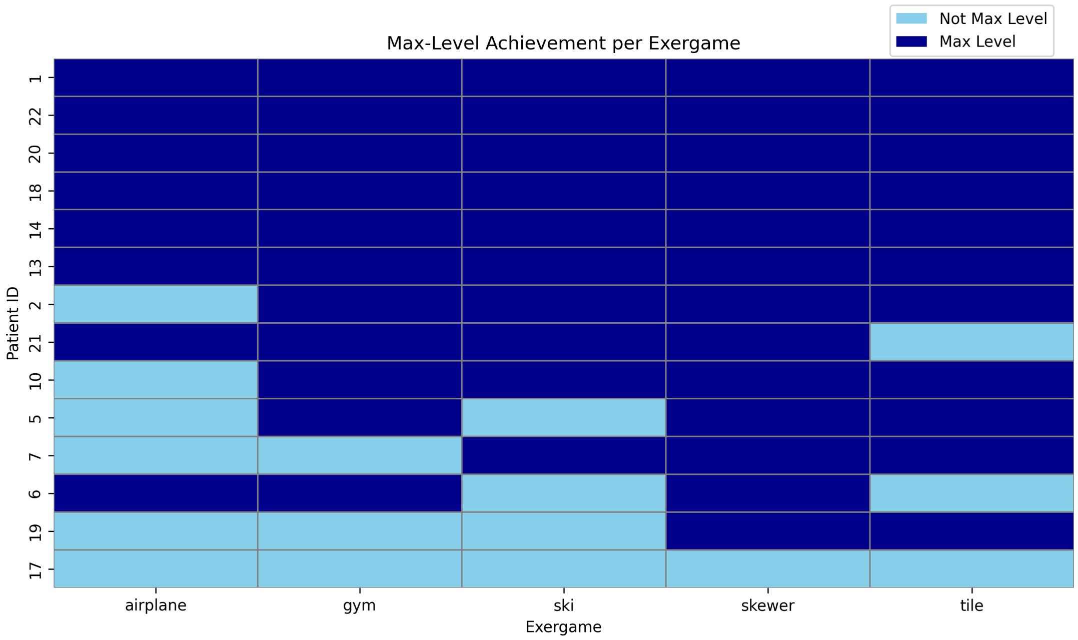

Figure 3.24: Binary Matrix pointing out patients that reached maximum levels in exergames.

the total number of exergames in which they reached the maximum level, allowing visual grouping of highly engaged or skilled individuals. The visualization was generated as a color-coded heatmap, with light and dark shades distinguishing max-level achievement.
After a first look, patients 1, 13, 14, 18, 20 and 22 reached ceiling level on all exergames. Among these, individuals 13, 14 and 22 fall into the low PD stage category, while 20 in the medium. Patient 1 is classified as high UPDRS, but the clinical score of 27.5 is at the lower bound of this category and is derived by non-motor symptoms, like voice-related complications. Subject 18, as noted before, can be considered an outlier, noted as high UPDRS individual but with great performances.
Participants 2, 5, 6, 7, 10 and 21, which reached peak level in at least three exergames, belong all to medium clinical class, apart from subject 10 which is ranked as high. Remaining patients, which reached maximum level in two or less games, are clustered as high and are the same patients with weaker performance metrics in plots examined in previous analysis scope. Overall, 9 out of 14 patients reached peak level in at least half games and just one subject never reached the most challenging stage of any exergame.

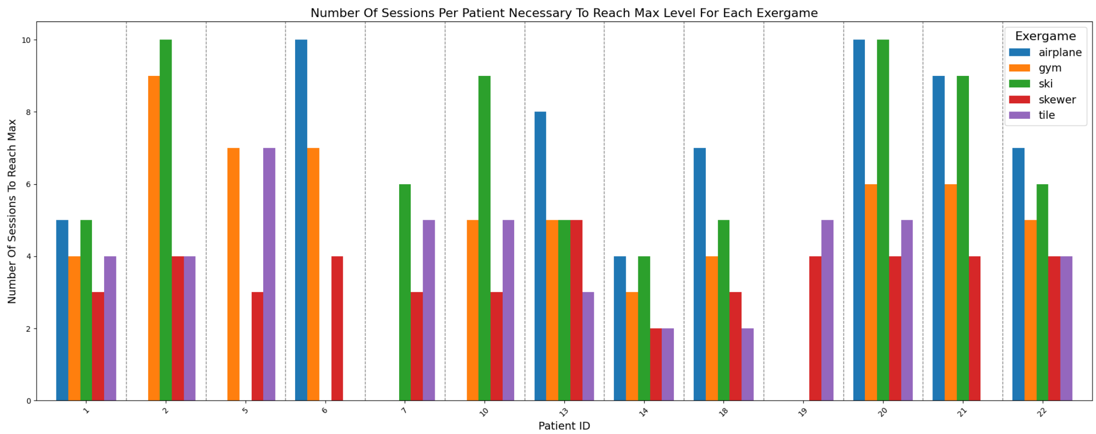

Figure 3.25: Bar plot showing the number of sessions needed for each patient to reach maximum level in any exergame.

The bar plot in Figure 3.25 was generated using a summary dataset containing the session number where a patient hit the global peak level for a game. Separated by a vertical dashed line, patients are on the x-axis, while session numbers are on the y-axis. Order of the games considered is always the same in each subject. It is important to note that if a patient (like number 17) never reached an upper ceiling value in any game it will not be displayed. The same is valid for single exergames, displayed only if subject managed to complete the maximum level.

This plot offers a comparative overview of learning pace and adaptation across game modalities. Patients with lower bar heights across exergames may have reached peak difficulty early, whereas higher bars suggest gradual progression. It can be also indicative of the difference in perceived game-induced challenge. Results clear that Airplane and Ski are the hardest games, while Skewer appears to be the most inclusive exergame.
Apart from subject 20, which can be considered a slow, progressive learner, all other subjects (1, 13, 14, 18, 22) that reached peak levels in all games took at most 7 sessions. Interesting is the case of patient 7 and 19: despite mastering only 2 or 3 out of the 5 exergames, they managed to do so in at most 6 sessions, indicating how different game requirements influenced the outcomes. Patient 19 succeeded in Gym and Skewer which require slow movements and have a static, non-dynamic background. Patient 7 also performed well in the Ski, despite being the only one to play in the preferred standing position.
To rank patients by multi-modal performance, a composite scoring approach was applied on a DataFrame composed of 1163 records (each row is an exergame attempt in a session for a subject, for which also survey answers are reported). It contained:

- Positively-weighted metrics: score, game level, self evaluation, satisfaction, repeatability.
- Negatively-weighted metrics: game time, errors, session overall duration, difficulty and fatigue.

All metrics were normalized, then for negative-weight metrics sign was inverted. A composite score was evaluated per row as mean of features. Finally, aggregating by patient and computing the mean of the new score, ranking was obtained, displayed in Table 3.2. The score difference between top-ranked Patient 14 and bottom-ranked Patient 17 is outstanding. Only the top 6 subjects have positive scores. These value are proportionally influenced (since features were standardized) by questionnaires answers. This is why patient 21, which did not reach top level in all exergames) is in third position, while patient 20 (with highest perceived difficulty through sessions) is third to last. Questionnaire answers will be analyzed in next section.

3.4.2 Global Progression Analysis

This study was conducted by computing the maximum level reached in each gamified session by each patient, considered collectively rather than individually and in some cases grouped by UPDRS tertiles. Analysis was performed using stacked bar plots and point plots.
Point plots are used to illustrate the mean maximum game level reached in each

Table 3.2: Final Patient Ranking.

Ranking	Patient ID	Composite Score
1	14	0.689257
2	13	0.460066
3	21	0.235544
4	1	0.183399
5	18	0.126259
6	22	0.081800
7	6	-0.004346
8	2	-0.066756
9	5	-0.105916
10	7	-0.180273
11	10	-0.319574
12	20	-0.463665
13	19	-0.666743
14	17	-0.799093

session for patients grouped by UPDRS tertiles. These plot show both the average trend and the variability within each group, making it easier to compare progress across severity levels. For each exergame, two versions are presented: one includes all patients contribution, while the other excludes the outlier (patient 18) to show how impactful it was on the per tertile computation.
Cumulative stacked bar plots are used to show how many patients reached each maximum level in each session. This visual representation makes it clear how the distribution of levels shifts over time, highlighting the overall progression pattern and the persistence or disappearance of lower performance levels. It is important to highlight that stacked bar height changes with the session progression, since from session number 5 some patients ceased to participate: total number of participants was 14 at start, 12 in session 6 and 7, 10 in ninth one and 9 in the last.

Airplane

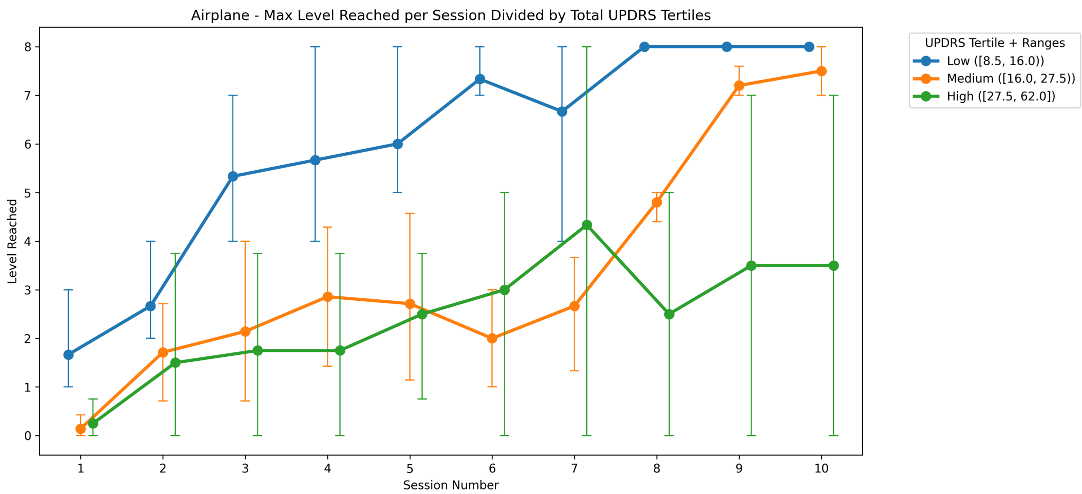

Figure 3.26: Point Plot showing patients' maximum game level evolution in sessions, divided into UPDRS tertiles.

In Figures 3.26, 3.27 tertiles-divided patients' evolution in game level is analyzed: by session aggregation, the maximum session game level for each patient is added and the total is divided by the number of total players for that session, obtaining the mean value of the peak session level. Together with the mean, also the std by session is displayed. Notably, the standard deviation for the high tertile increases with session number, while medium and low UPDRS classification trend is inversely proportional. This indicates that less severe patients tend to adapt to their respective tertile means, especially in last 3 sessions. Most severe patients, instead, appear to be more heterogeneous in their evolution. In Figure 3.27, tertiles evolution

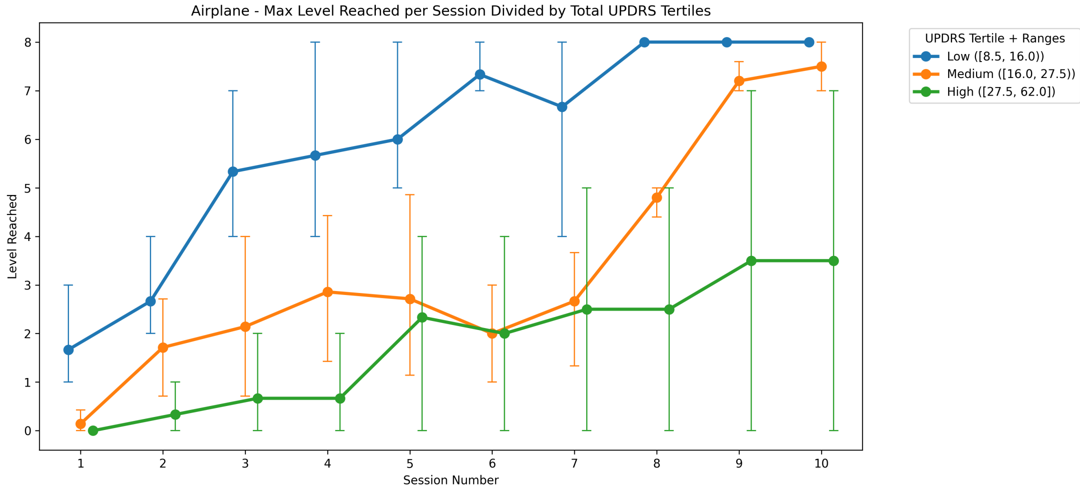

Figure 3.27: Point Plot showing patients' maximum game level evolution in sessions, divided into UPDRS tertiles. No patient 18 in higher tertile.

looks almost totally well separated according to clinical evaluation. Considering instead also the outlier, for all the 7 sessions to which it took part, higher UPDRS score patients appear to be more performative. Stacked bar plot in Figure 3.28

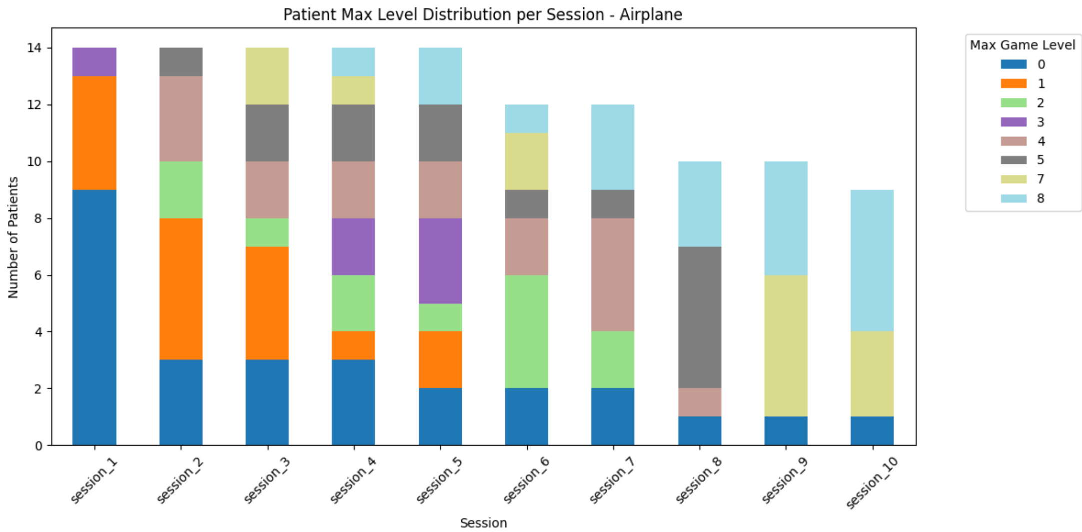

Figure 3.28: Display of number of patients that reached a certain maximum level in sessions.

indicates how level distribution mutated through sessions. Aside from persistent low-level performance attributed to the most severe patient (17), a clear upward trend is observed, with metrics lasting from 5 to 7 sessions.

Tile

By investigating Figures 3.29 and 3.30, it is evident that the Tile game has a differential impact on each severity level patient. For least severe patients it appears to be an easy to master game, with upper ceiling level reached from session 4 and without variance. Medium UPDRS group perceived the game as stimulating, with no global difficulty plateau that slowed down the progression curve. Their std however increased until last two sessions, where lowered, indicating an increment on total number of medium condition patients that successfully reached the top level. Most severe patients' attempts generated on average a learning curve which is not strictly monotonic, having a total of 4 sessions where performances remained identical to the previous ones.
These conclusion are further supported in Figure 3.31, where it is clear how learning curve was overall gradual, with all levels present in all sessions, from the second to the last one. First level persistence is again attributable to most severe patient, number 17, confirming consistent difficulty for the most impaired subject.

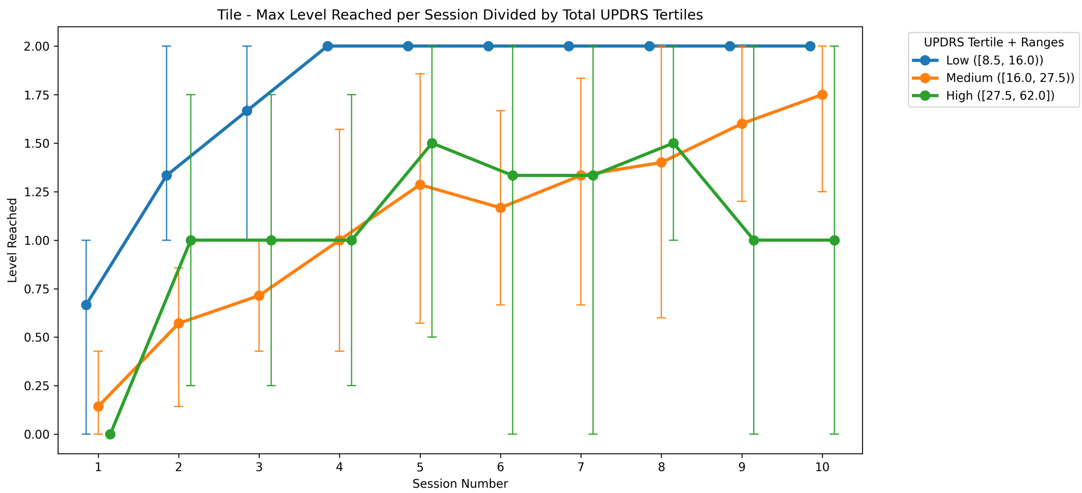

Figure 3.29: Point Plot showing patients' maximum game level evolution in sessions, divided into UPDRS tertiles.

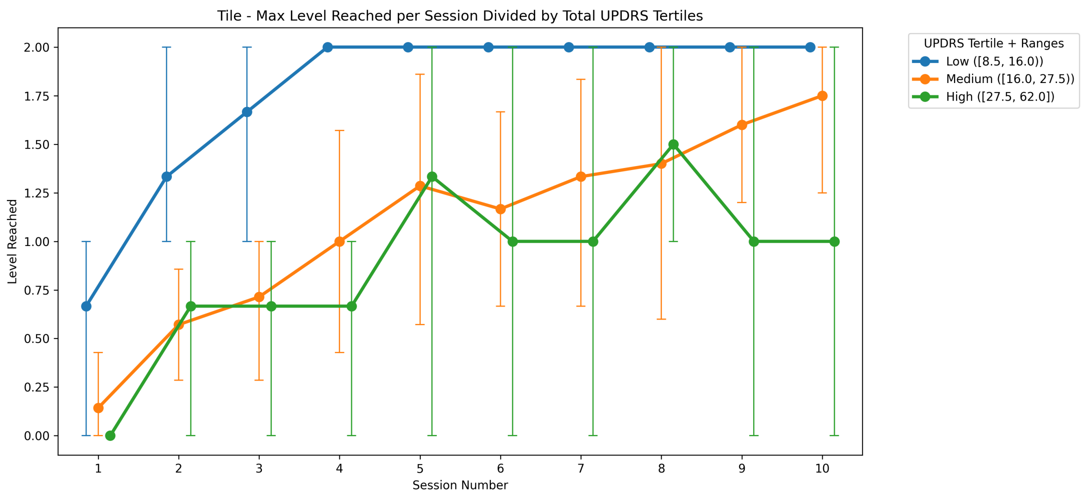

Figure 3.30: Point Plot showing patients' maximum game level evolution in sessions, divided into UPDRS tertiles. No patient 18 in higher tertile.

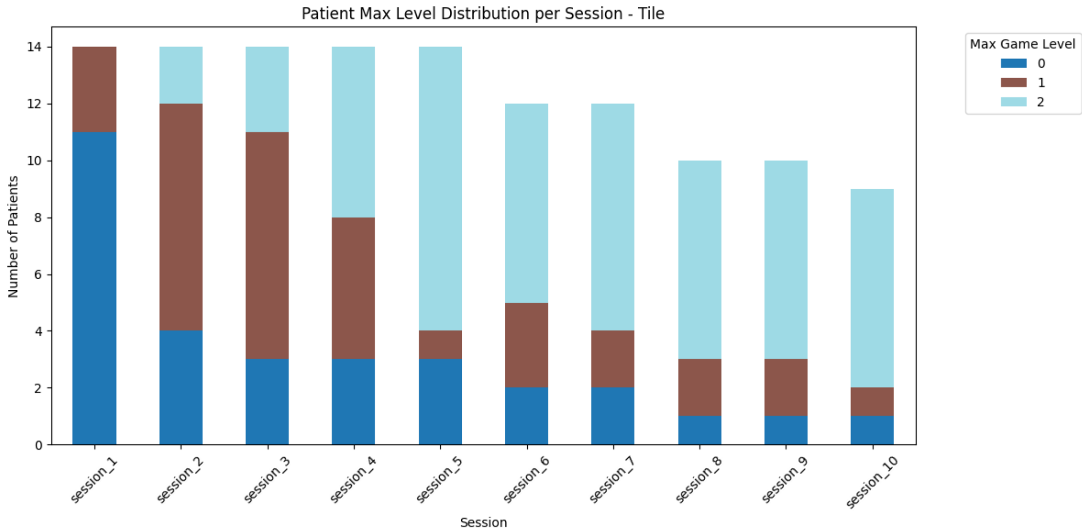

Figure 3.31: Display of number of patients that reached a certain maximum level in sessions.

Skewer

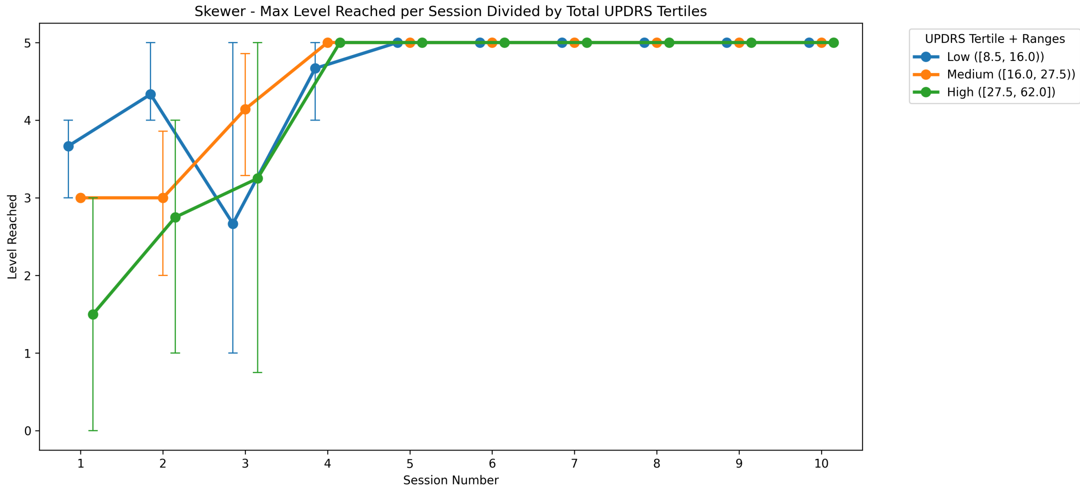

Figure 3.32: Point Plot showing patients' maximum game level evolution in sessions, divided into UPDRS tertiles.

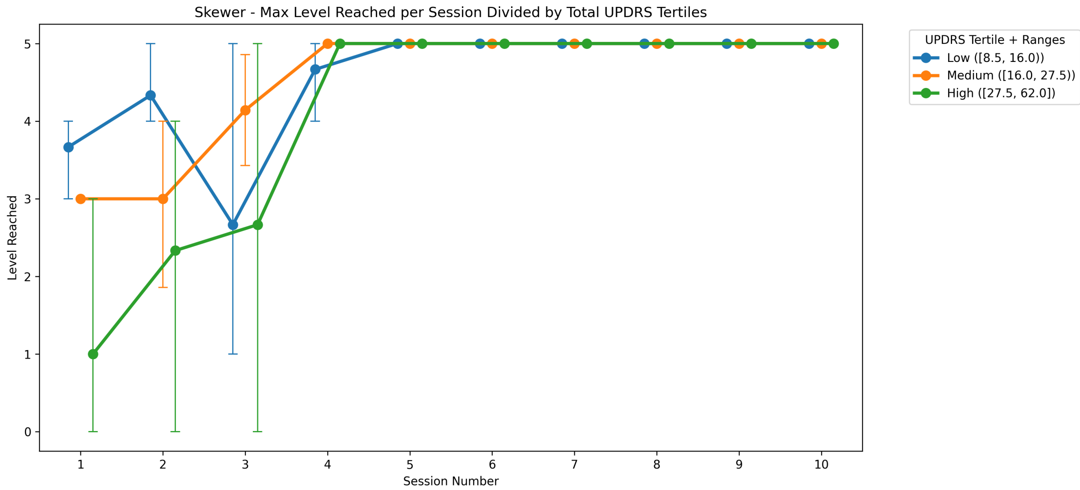

Figure 3.33: Point Plot showing patients' maximum game level evolution in sessions, divided into UPDRS tertiles. No patient 18 in higher tertile.

Considering point plots in Figures 3.32, 3.33, it is evident that patients across all severity categories quickly reached the maximum ceiling, showing a steep progression curve. Variances decreased progressively until session 5, where reached near zero values. This session is also the one where upper ceiling was reached by all players (except number 17, which was unable to conclude also basic level due to motor impediments). In session 3, the mean for low-severity patients was unexpectedly low: they played on average easier modes than medium UPDRS classification participants. In example, patient 13 in that session made just one attempt at level 1. Figure 3.34 illustrates that Skewer was perceived as the least difficult exergame.

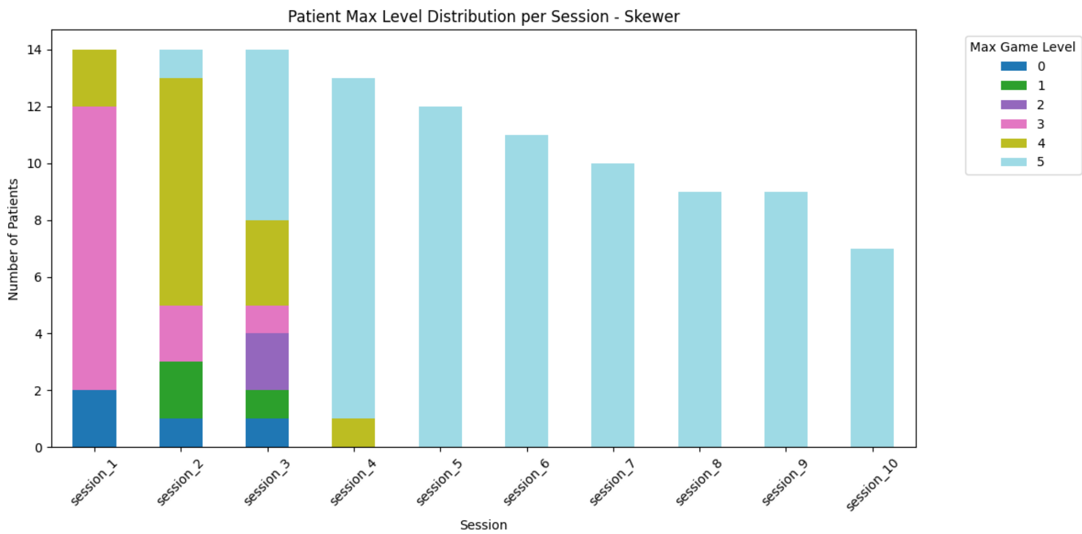

Figure 3.34: Display of number of patients that reached a certain maximum level in sessions.

Base level gets mastered at the end of session 3, with peak levels reached from the early second session. Starting from the fifth session, the distribution of levels within the stacks is concentrated on peak level unique contribution, as seen by the predominance of one color in the session bars.

Ski

Figures 3.35 and 3.36 demonstrate the Ski game's sensitivity in discriminating Parkinson's severity. The low UPDRS group exhibited the highest and most consistent performance, reaching elevated difficulty levels quickly and uniformly throughout all its members. The high UPDRS group showed a less steep and not strictly monotonic progression curve, with greater intra-group variability across sessions compared to the other two groups. The moderate group displayed a steady, monotonic progression, with a standard deviation that, after growing until session 6, decreased towards uniformity in later attempts. From the stacked bar plot in Figure 3.37, it is evident how Ski learning curves, previously described, resulted in a well distributed and consistent pattern across sessions. The only level which all patients

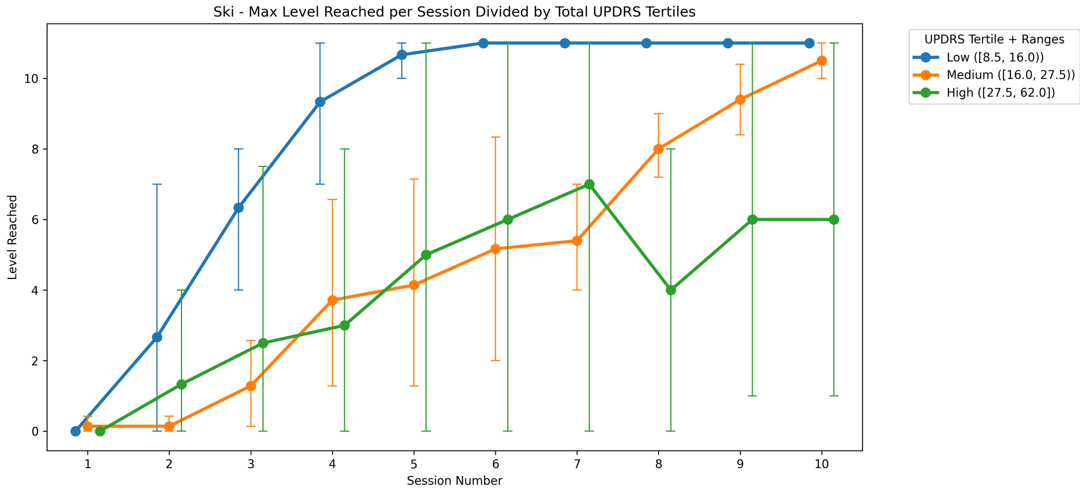

Figure 3.35: Point Plot showing patients' maximum game level evolution in sessions, divided into UPDRS tertiles.

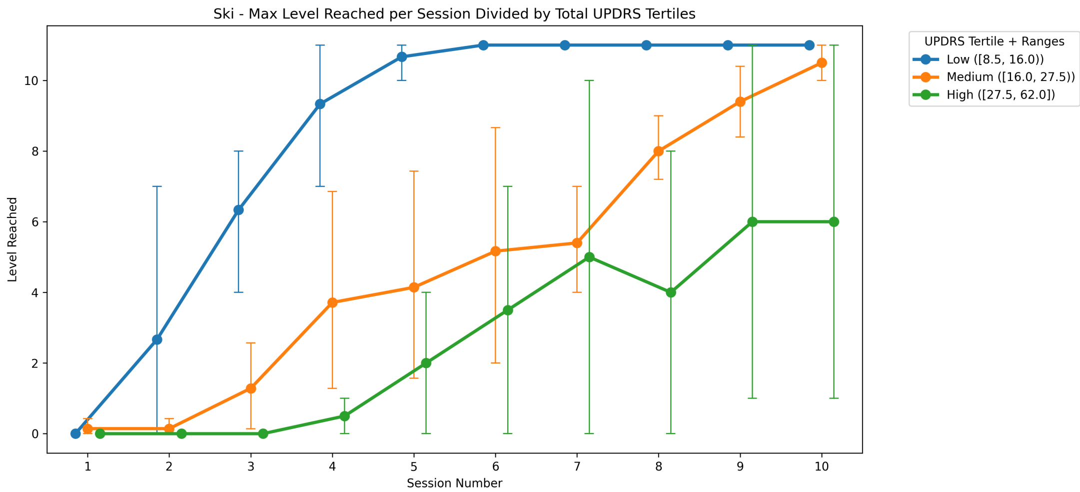

Figure 3.36: Point Plot showing patients' maximum game level evolution in sessions, divided into UPDRS tertiles. No patient 18 in higher tertile.

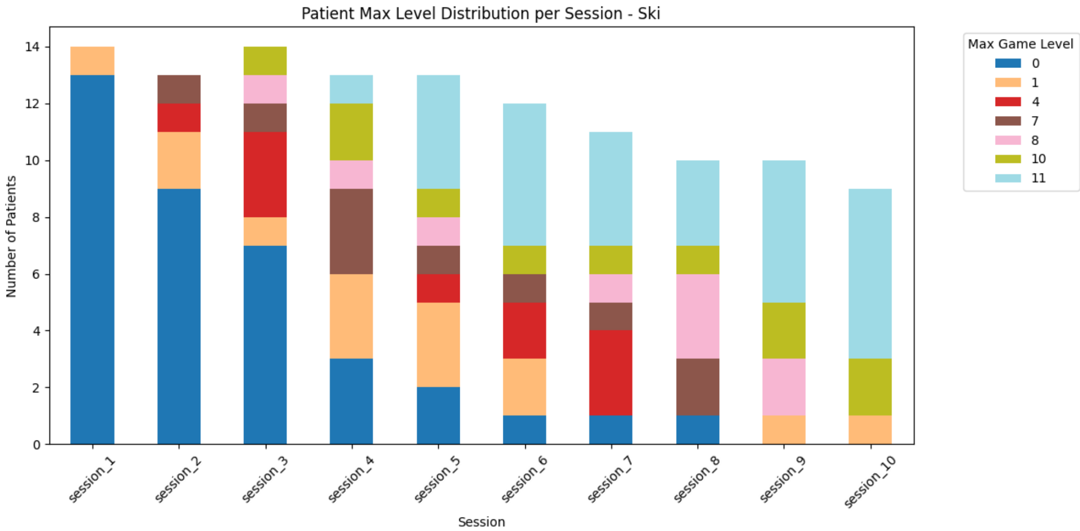

Figure 3.37: Display of number of patients that reached a certain maximum level in sessions.

passed before the end of the collection process was the base level, while participants were dealing with multiple levels until last tries. Given the flexibility granted by an 11-level scale, patients were able to face appropriately scaled challenges over time, without stagnating due to difficulty plateaus. By the end of the data collection, participants either confronted gradually with more challenging scenarios, until finding the right difficulty for them or reached peak level.

Gym

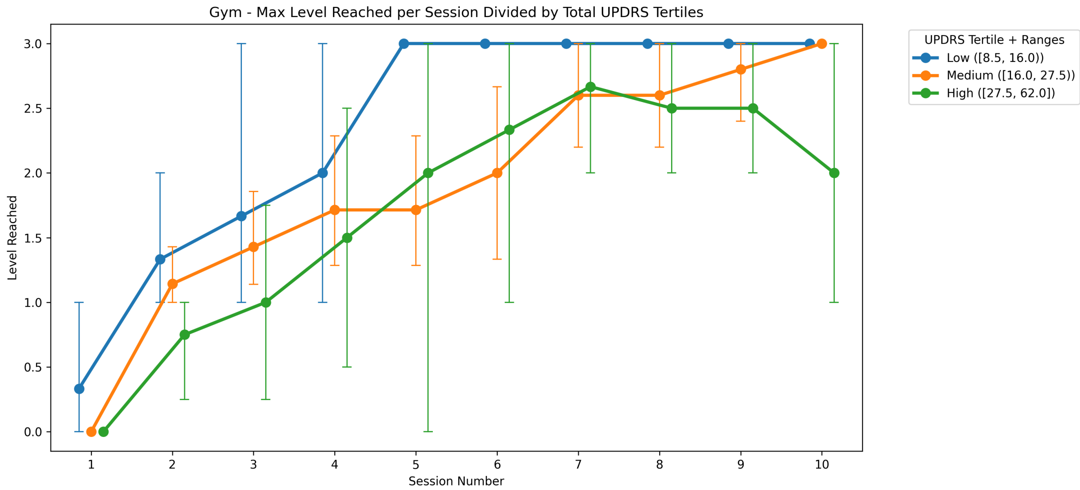

Figure 3.38: Point Plot showing patients' maximum game level evolution in sessions, divided into UPDRS tertiles.

In Figures 3.38 and 3.39, the results clearly show that in Gym exergame, as for the Skewer one, all classes patient learning curves were steep. For low severity patient, upper ceiling was reached uniformly by all members by session 5. For medium UPDRS cases, curve is strictly monotonic with contained standard deviations, which reached near zero value in the last session. Most compromised class patients exhibited greater variability throughout the collection, but still managed to attain average game levels comparable to the medium severity group. By looking at Figure 3.40, it is noticeable how base level was easily surpassed after first session by all patients except the most severe one, which, however, managed to move to the next level after session 5. In the last session, due to personal preference, one participant opted to play at level 1, which was surpassed by all patients by session 6. This is why the final session's stacked bar is not uniform; otherwise, the level distribution would show no inter-patient variability, with all participants playing at maximum level.

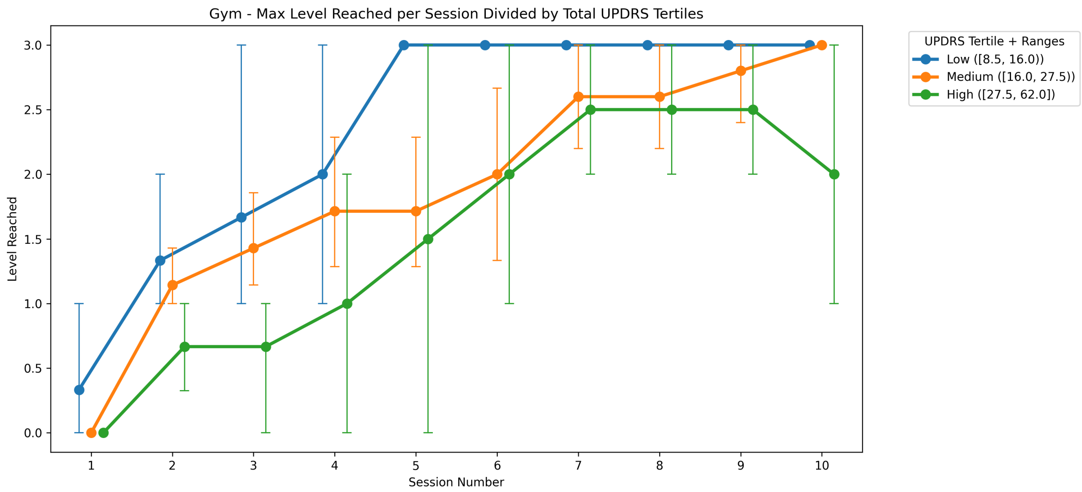

Figure 3.39: Point Plot showing patients' maximum game level evolution in sessions, divided into UPDRS tertiles. No patient 18 in higher tertile.

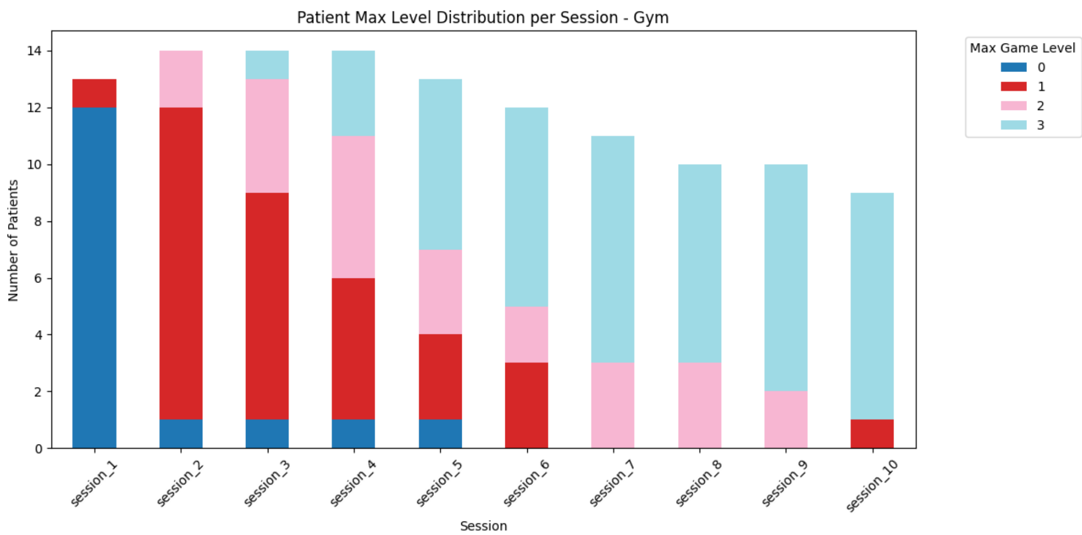

Figure 3.40: Display of number of patients that reached a certain maximum level in sessions.

3.4.3 Exergames Difficulty Ranking

Since game ranking is directly linked to patient progression, regardless of whether patients are analyzed individually or as a group, in order to evaluate which are the most challenging games two plots are examined. First, session-wise patients heatmaps on game level reached in session 1, 5 and 10 are inspected. Then a custom UpSet plot is considered, showing the number of participants that reached peak level in different game combinations. In Figure 3.41, three adjacent heatmaps

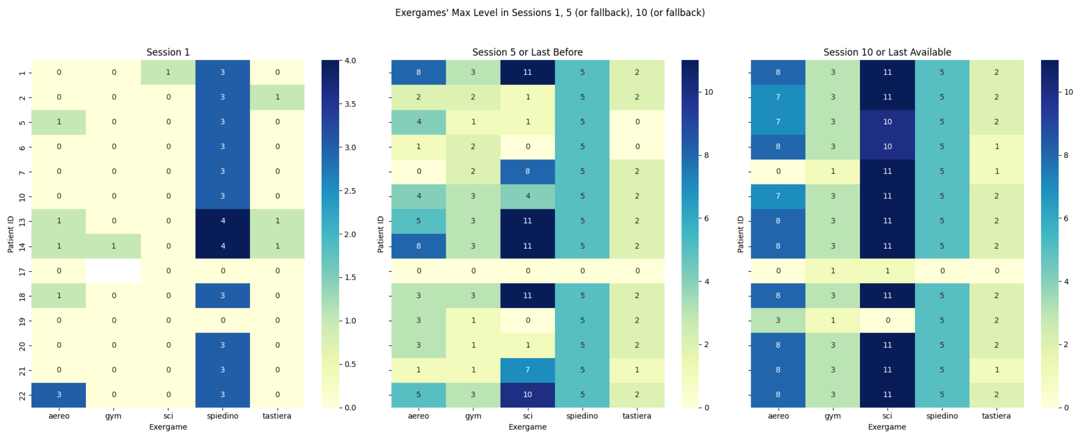

Figure 3.41: Maximum game level matrix for each patient and exergame. Analysis on sessions 1, 5 and 10

are presented, showing the maximum game level achieved by each patient across three distinct time points: Session 1, Session 5 (or the last available session before it) and Session 10 (or last available).
At the end of the first session, the majority of patients managed to reach only relatively low difficulty tiers across all exergames, with higher initial performances observed in Skewer game. By Session 5, a clear upward shift is visible in most games: the majority of patients exhibit improved level attainment, with few individuals able to reach global maximum level in Ski (4 subjects) and Airplane (2).For the remaining games, between 6 and 13 patients were able to arrive at upper ceiling level.
This progression is even more pronounced in last session, which can be used as a meter to rank games:

1. **Airplane**: most difficult game, with the slowest learning curve and least top level players (8 out of 14).

2. **Ski**: second hardest game, characterized by a slow improvement trend and with final 10 players able to deal with maximum difficulty.

3. **Tile**: exergame with a steeper learning curve that flattened in central session for the final best performers, except for subject number 5. Final number of top level players is 10, but is technically 11 in total, since patient 6 regressed in the final sessions.

4. **Gym**: second to last ranked game, with a steep learning curve and final 11 patients that reached top level.

5. **Skewer**: simplest-perceived game, with high difficulty modes reached by most players already at the end of the first session. Final number of top performers is 13 out of 14, the highest.

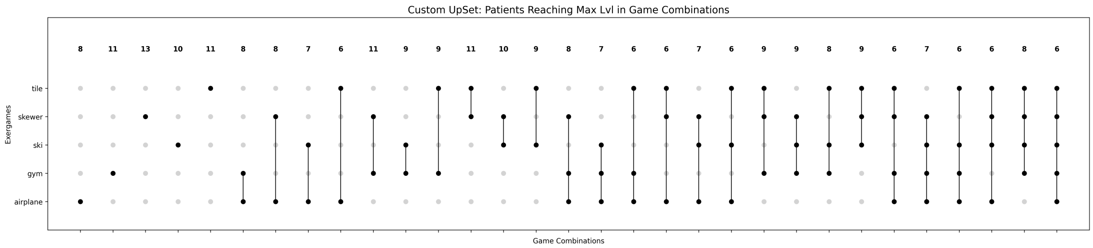

Figure 3.42: Display of number of patients that reached peak level according to different game combinations.

Custom UpSet plot in Figure 3.42 serves as a ranking tool based on the number of patients that succeeded in reaching top level in game combinations. Basing on the values of combination where either Airplane or Ski were included, these two are the most challenging games. In particular, Airplane seems to be the most difficult.

3.5 Questionnaires Analysis

This section is dedicated to the investigation on Likert-scale questionnaire answers. Due to their subjective nature, survey responses may be influenced by external factors such as mood or personal circumstances and thus may result to be uncorrelated with game performances. Results will be firstly divided by type of aggregation, then separated between items. Chosen indicators were:

- **Self Evaluation**: Pre-session self-assessment on overall conditions. Values equal to 5 are intended for best possible conditions.
- **Satisfaction**: Index from 1 to 5 directly proportional to selected value.
- **Replayability**: Item capturing a participant's willingness to continue with the rehabilitation protocol. Maximum value means total patient disposability.

- **Fatigue**: Used to assess the physical impact on participants. Higher values indicate increased tiredness,
- **Difficulty**: Differently from fatigue, allows consideration not only from the motor point of view, but also on cognitive complications.

3.5.1 Per Patient Aggregation

This section aims to identify trends in patient-reported feedback aggregating survey responses across sessions. To achieve this, the survey DataFrame was grouped by patient ID, computing the overall mean, standard deviation and maximum values for each survey question of interest. For every survey item, a summary DataFrame was created containing one row per patient, ordered by descending mean value. Summaries were visualized using Matplotlib bar charts, where for each patient:

- The **bar height** represents the patient's average response for that question.
- The **gray shaded area** indicates the ± standard deviation.
- The **black triangle marker** denotes the maximum response given by the patient.

A **dashed red line** shows the overall mean response across all patients and sessions. The standard deviation was computed using the `.std()` method from the pandas library, which measures how much the individual responses deviate from their mean value:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.2}$$

where:

- s is the sample standard deviation,
- $\bar{x}$ is the mean survey score,
- x_i represents each individual survey response,
- n is the total number of answers for that patient.

Self Evaluation

In Figure 3.43, per patient ranking of self evaluation is displayed. Top performers like 14, 13 and 1 respectively answered 5, 4 and 4 for all sessions, with no variance, showing how their stability in perceived condition also reflected on game metrics.

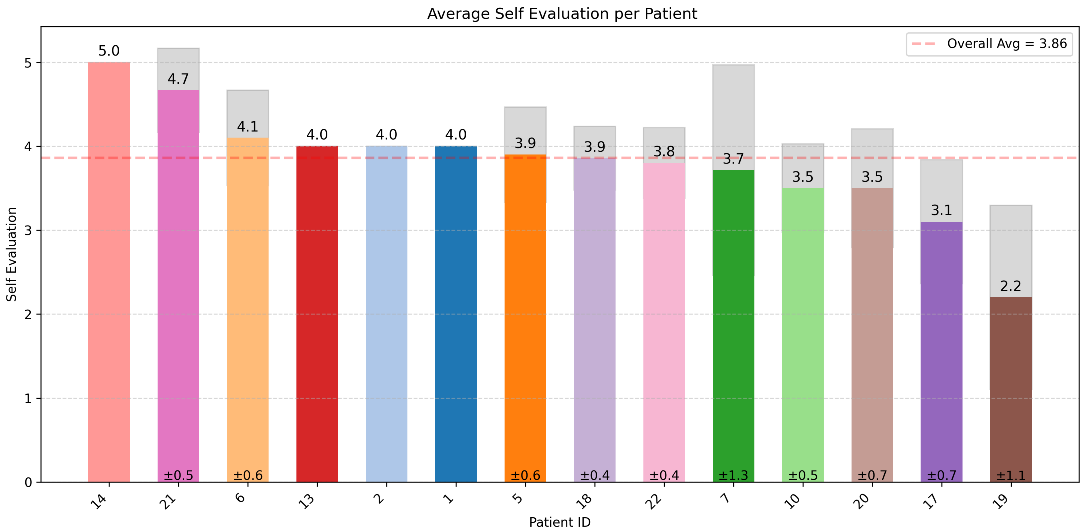

Figure 3.43: Display of patients ranking based on Self Evaluation metric.

Subjects 18, 22 and 20 instead, despite being able to complete all the exergame at peak difficulty as seen previously in Figure 3.24, responded with lower values and a bigger variability (std is near 0.5 units). Participants 17 and 19, which played games mostly at basic levels, have the lowest average self evaluations. Number 7 has the biggest variability, which was reflected in inconsistent session performances across the collection process. Finally, patient number 21 appear to be among participants with higher metric, even if in previous analyses on game metrics was not one of the bests. For this reason, attention will be paid in successive metric inspections to see if this behavior is recurrent.

Satisfaction

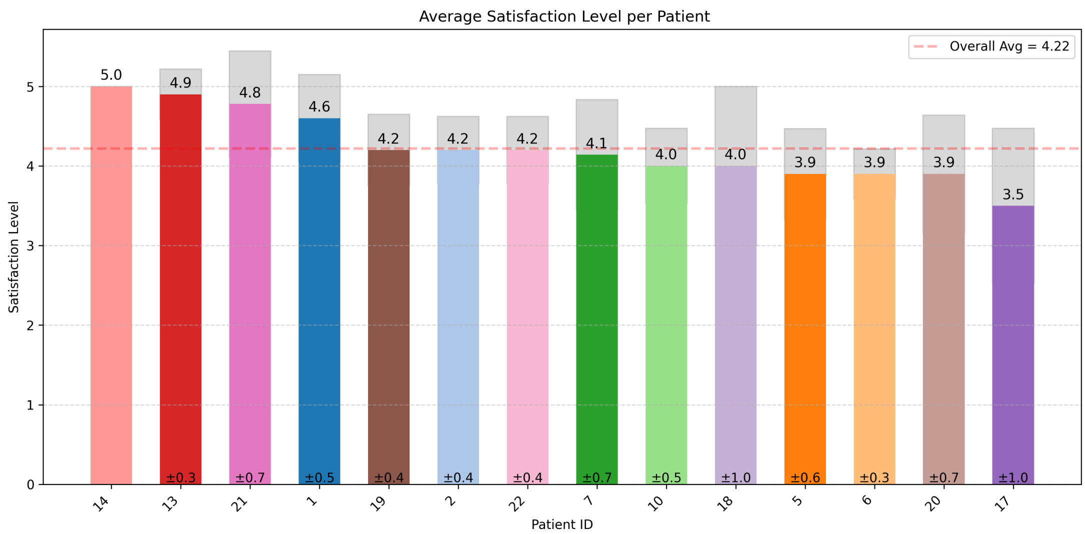

Figure 3.44: Display of patients ranking based on Satisfaction level metric.

By considering Figure 3.44's satisfaction levels, it is immediately noticeable that best performers like 1, 13 and 14 have higher metric mean values. Inversely, the most critical patient 17 was the least satisfied. Subject 19, which is among the ones with slowest learning curve, kept across sessions a mean value identical to subject 22's one, which is a top performer. Despite being able to succeed in all exergames at upper bond levels, numbers 18 and 20 again gave low-point answers. Also for this metric, patient 21 showed to be more satisfied than expected based on performance metrics. Overall average for participants was 4.2 out of 5, indicating high usability and patients' approval of the exergames.

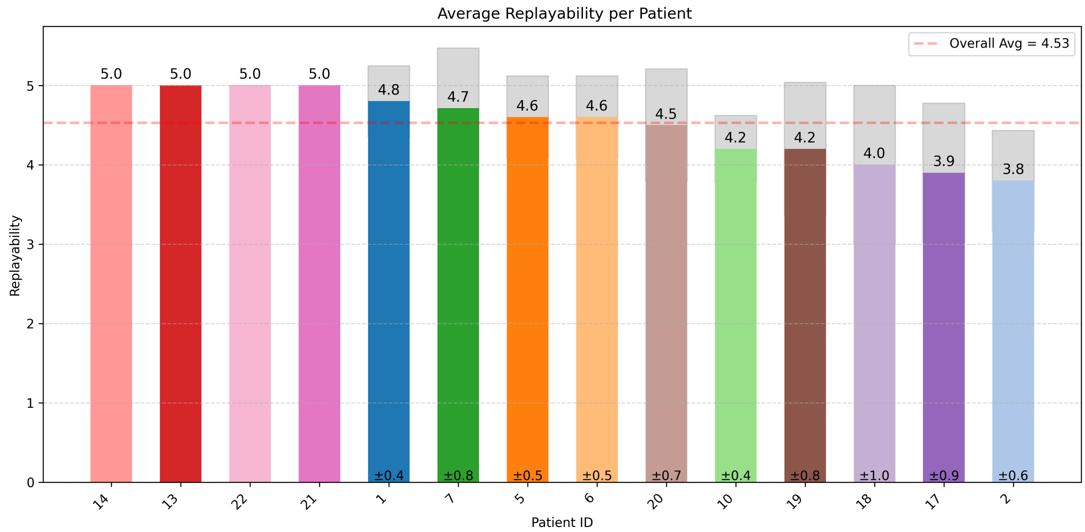

Figure 3.45: Display of patients ranking based on Replayability metric.

Replayability

Repeatability answers reported in Figure 3.45 are also a good indicator on quality of gaming experience proposed to the group. Overall, this metric was rated 4.5 out of 5, showing how positively games were welcomed by patients. In particular, most of top performers plus number 21 answered with maximum score in all sessions with no variance. Most severe cases (subjects 17 and 19) replied with overall means value below global average. Interesting are the cases of subject 18 and 2: these participants performed well but may have experienced diminishing engagement due to limited challenge.

Fatigue

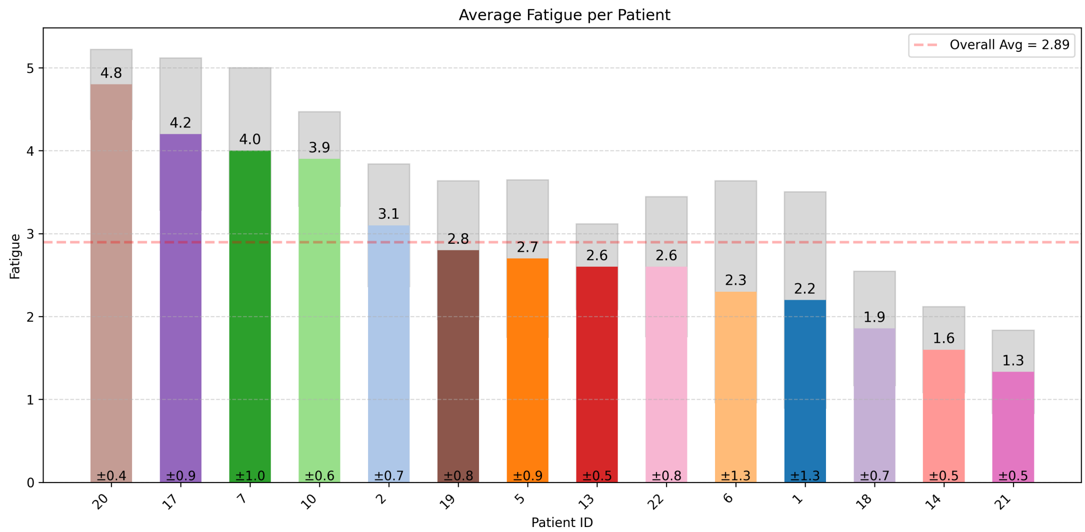

Figure 3.46: Display of patients ranking based on Fatigue metric.

Analyzing Fatigue metric in Figure 3.46, there is a clear inverse relationship between top performers and fatigue: best players perceived less fatigue overall, while most severe patients were more tired at the end of each session. Interesting is the case of patient 20: despite being able to complete all games with peak level, reported a fatigue mean value of 4.8 out of 5, which is the highest. Inversely, subject 21 was the least exhausted participant. Overall mean group fatigue was near 3 units out of 5, which is a good value for rehabilitation-marked games. Subject 2's elevated fatigue score may explain his lower replayability rating highlighted before.

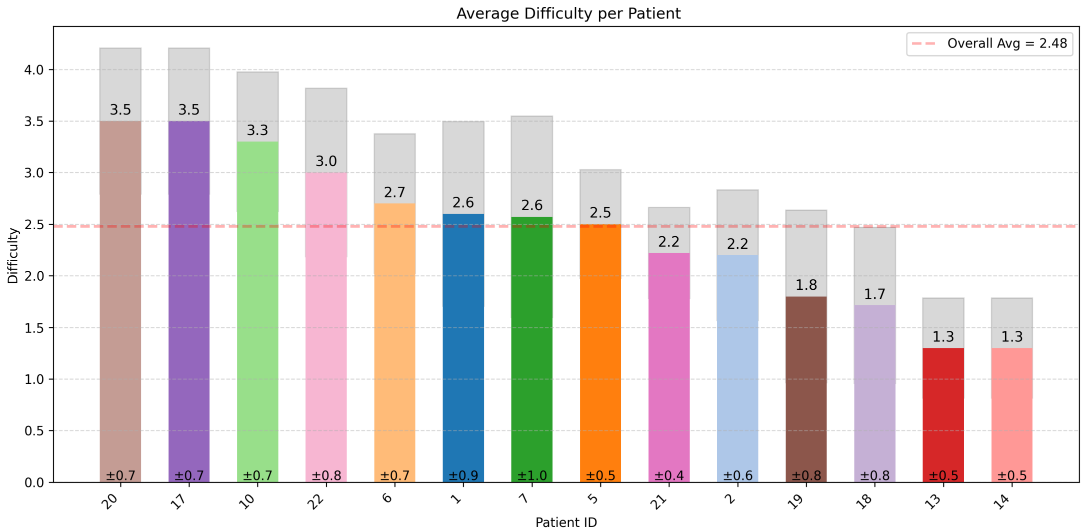

Figure 3.47: Display of patients ranking based on Difficulty metric.

Difficulty

Patients ranking based on difficulty metric is displayed in Figure 3.47. Overall mean difficulty is just below Fatigue's one, indicating correctly that exergames are focused more on energy exertion than on cognitive difficulty. As for the previous metric, subject 20 is the one with highest mean metric value. Lowest-metric patients are the top performers (13, 14 and 18) with the exception of number 19. Most severe patient 17 expectedly has the second highest mean value. Finally subject 22, ranked as one of the best in game metrics, self-reported his gaming experience as the fourth more in difficulty.

3.5.2 Per Session Aggregation

This analysis examines patient-reported survey responses across rehabilitation sessions, aiming to uncover trends and fluctuations in the selected metrics over time. To achieve this, questionnaire DataFrame was grouped by session number. For each metric, the **mean**, **standard deviation** and **maximum value** were computed across all participating patients.
The resulting session-level summaries were used to generate Matplotlib bar charts, structured as follows:

- The **bar height** represents the mean value of the selected metric across all patients for that session.
- The **gray shaded rectangle** visualizes one standard deviation ($\pm\sigma$) around the mean.
- A **black triangle marker** denotes the maximum value reached for that metric in the corresponding session.

Additionally, a **dashed red line** indicates the global average across all session entries in the dataset for that metric.

Self Evaluation

The first metric examined in session-level aggregation is self evaluation, displayed in Figure 3.48. Intra-session averages span between the lowest 3.5/5 (obtained in the first one) and the highest 4.2/5, obtained in the eighth. Being a non game dependent metric, its general trend is erratic, with standard deviations that can reach values up to 22%. This parameter results to be influenced not only by the heterogeneity of the group, but also by inter-sessions intra-patient variations.

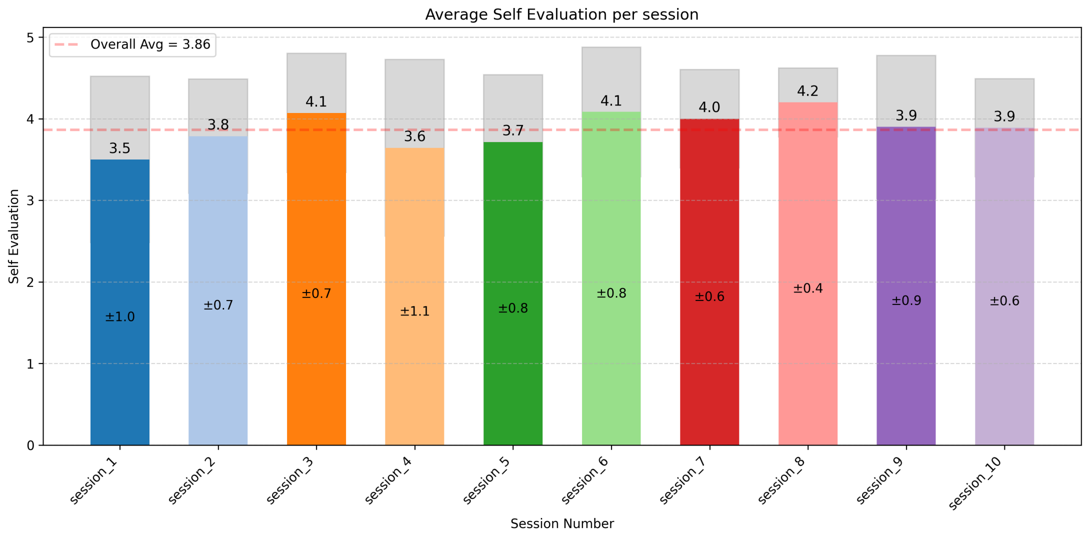

Figure 3.48: Display of session-based progression considering only Self Evaluation metric.

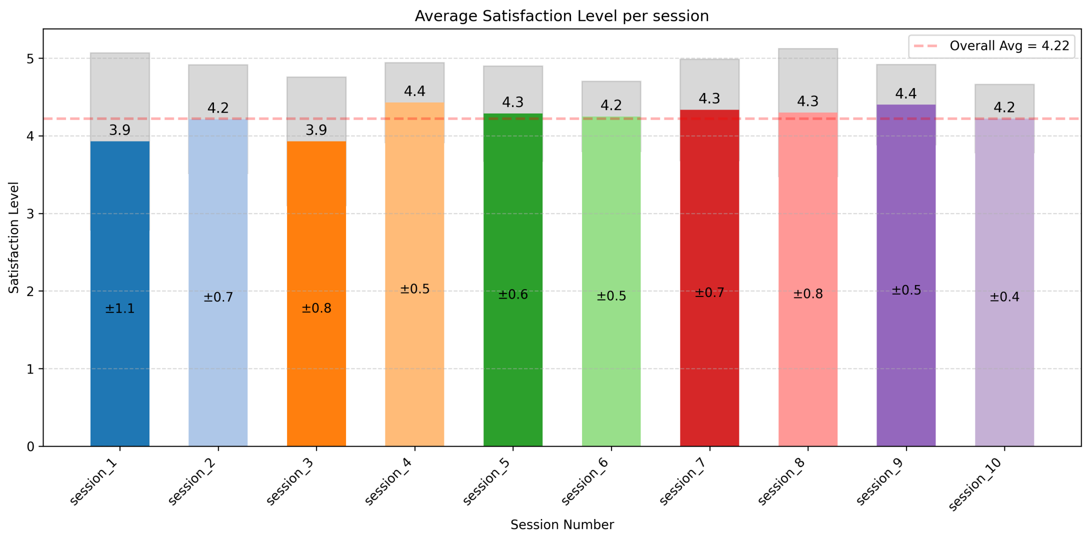

Figure 3.49: Display of session-based progression considering only Satisfaction level metric.

Satisfaction

Figure 3.49, which shows the satisfaction levels averaged through patients per session, suggests that the intervention was positively received throughout the study. Starting with an initial mean value of 3.9 out of 5, value increased until stabilization, reached around its overall mean of 4.2/5. Interesting is also to notice how first and last sessions were respectively the one with highest standard deviation and the one with the lowest.

Replayability

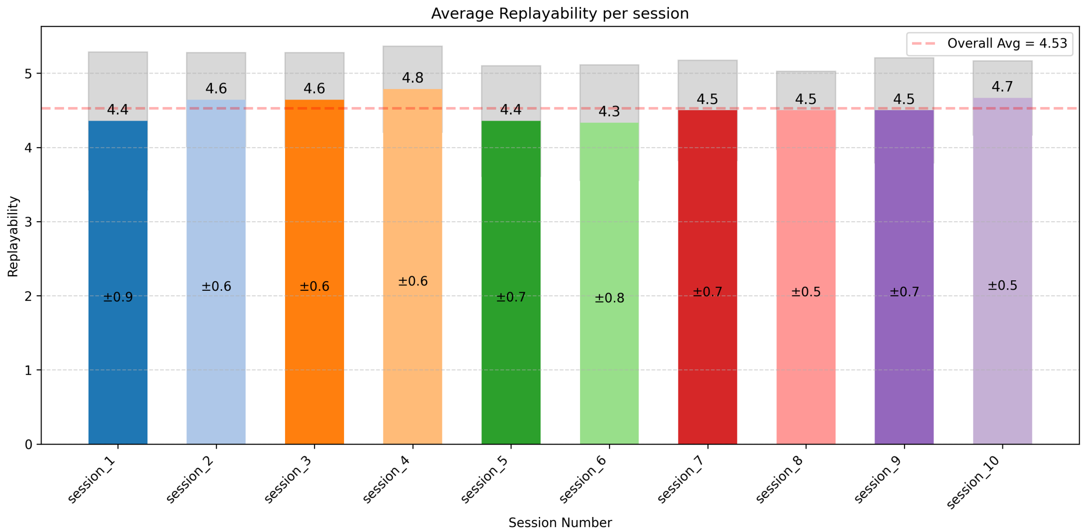

Figure 3.50: Display of session-based progression considering only Replayability metric.

In contrast, replayability trends displayed in Figure 3.50, reveal different dynamics: a positive trend characterized the first 4 sessions, becoming negative in fifth and sixth. From session number 7, another positive trend can be found, lasting until the end of the collection process. Standard deviation values range between 0.5 and 0.9 units, with highest found in first session and lowest found in session 8 and 10. Overall replayability is really high, with a value of 4.53/5, indicating study's success in keeping patients entertained and willing to continue playing. The alternating trend across sessions (positive, then negative, then positive) may reflect perceived difficulty in exercises like Ski or Airplane.

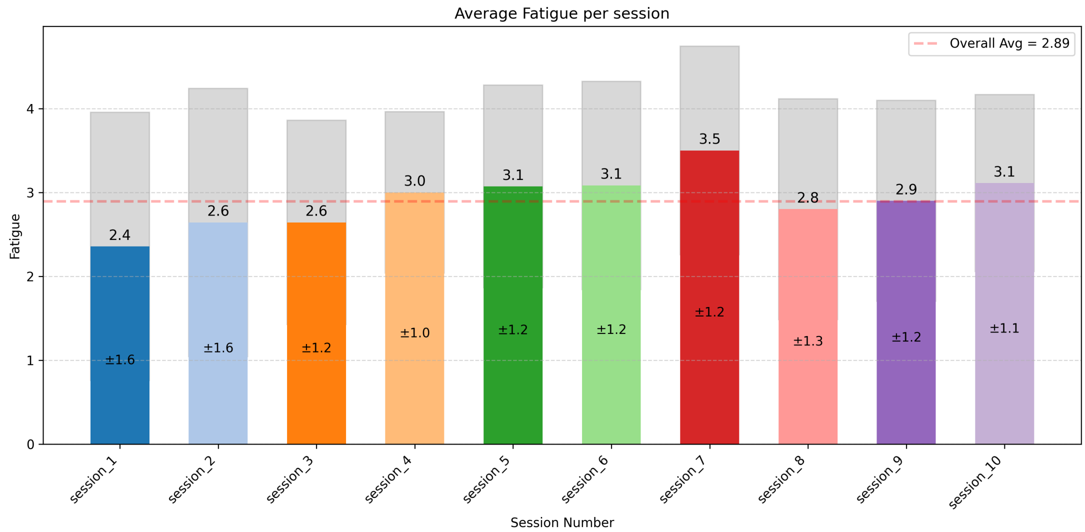

Figure 3.51: Display of session-based progression considering only Fatigue metric.

Fatigue

Considering the fatigue index displayed in Figure 3.51, a general positive trend is visible, starting from the lowest value of 2.4/5 obtained in session 1 up to 3.1/5 in the last one. Specifically first positive trend ends with the highest value of 3.5/5 in session 7 to later decrease in next one and incrementing in last two. Standard deviation captures participants' varied responses to evolving gameplay demands throughout sessions: values ranged between ±1.0 and ±1.6 on a scale with maximum mean score of 5, suggesting how differently this metric was evaluated by participants. Overall fatigue is near 2.9/5, which, as concluded in the analysis of same metric but for patient aggregation, is very good since one of main goal of exergame is physical rehabilitation.

Difficulty

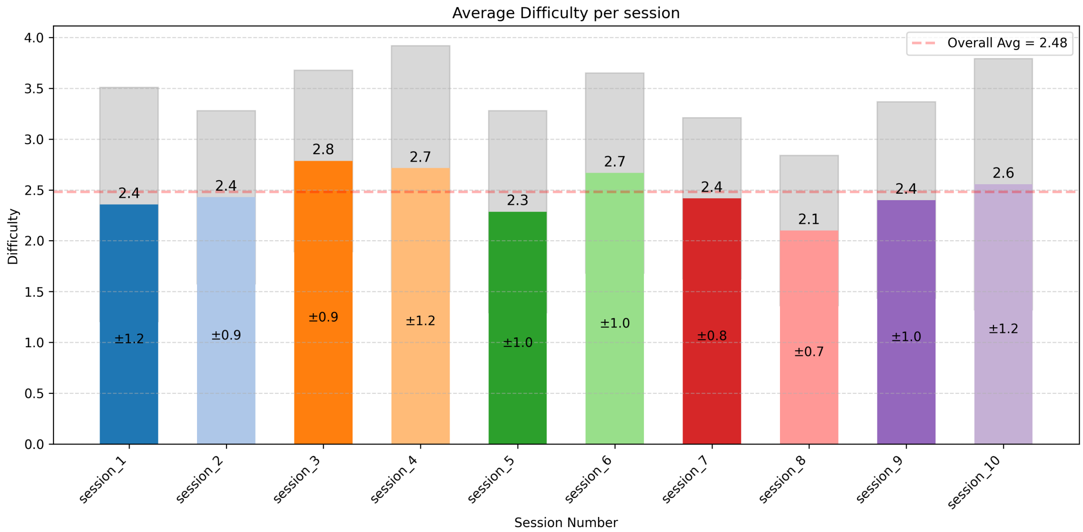

Figure 3.52: Display of session-based progression considering only Difficulty metric.

Analysis of the difficulty metric displayed in Figure 3.52 indicates an erratic trend. Considering that overall inter-sessions metric mean is close to 50% of the maximum, game levels progression kept experience cognitively stimulating, promoting more an increment of fatigue rather than of difficulty. Standard deviation, whose inter-session average is around 1 unit out of 5, captures heterogeneity of participants' response to the games evolution.

3.6 Correlation Heatmaps

Heatmaps were computed to visualize the correlation between multiple exergame performance metrics, session-specific survey responses and clinical evaluations for each patient. Specifically, for each exergame and for the aggregated dataset, relevant numeric variables (game times, levels, scores, errors) and questionnaire responses were merged with patient-level UPDRS assessments and characteristics. Specifically, these included years since diagnosis, Hoehn & Yahr stage and domain-specific UPDRS sub-scores like item involved in the estimation of bradykinesia, limb movements and postural instability & gait difficulty (PIGD).
To estimate these relationships, three standard correlation coefficients (Pearson, Spearman and Kendall) were compared systematically. Based on summary statistics of mean, minimum and maximum absolute correlation values, Spearman's rank correlation was ultimately selected for the final heatmaps. This choice is particularly appropriate given the repeated-measures structure of the dataset (up to ten sessions per patient with multiple game attempts per session) and the ordinal nature of some variables (Likert-scale questionnaire answers and UPDRS severity stages). Spearman's method does not assume linear relationships or normality and it captures monotonic relationships more robustly than Pearson's correlation, which is limited to linear dependence.
The resulting correlation matrices were masked for the upper triangle to enhance readability and were formatted to point out even weak but potentially meaningful relationships. This approach provides an interpretable, comprehensive overview of how motor performance and patient self-reported sensations relate to clinical disease severity across different exergames and over multiple sessions. In Figure 3.53 is displayed a global correlation heatmap considering game data, survey data and clinical scores. Starting by game data, it is important to notice that, considering all tries ever done by any patient for whichever exergame, variance across gameplay metrics is substantial: correlation exceeding 25% can be thus considered high. Keeping this in mind, game level and score metrics showed the strongest connections with items from questionnaire answers and clinical evaluation. In particular, high correlations are obtained with medication time delay (time shift from last levodopa-based medicament), self evaluation, satisfaction index, total UPDRS score and its feature regarding limbs movement. Among game metrics, game time is connected with game level, which is correlated with score and errors. Survey answers appear to be well connected with all clinical score, with values bigger or equal to 20%. Due to the volatile and variable nature of PD disease, the variable indicating the years since diagnosis is mostly uncorrelated, while the age item is connected with game score, errors and all survey and clinical variables, describing how influent it is on playing experience and perception, usually aggravating symptoms. Clinical scores are intuitively well bonded among themselves

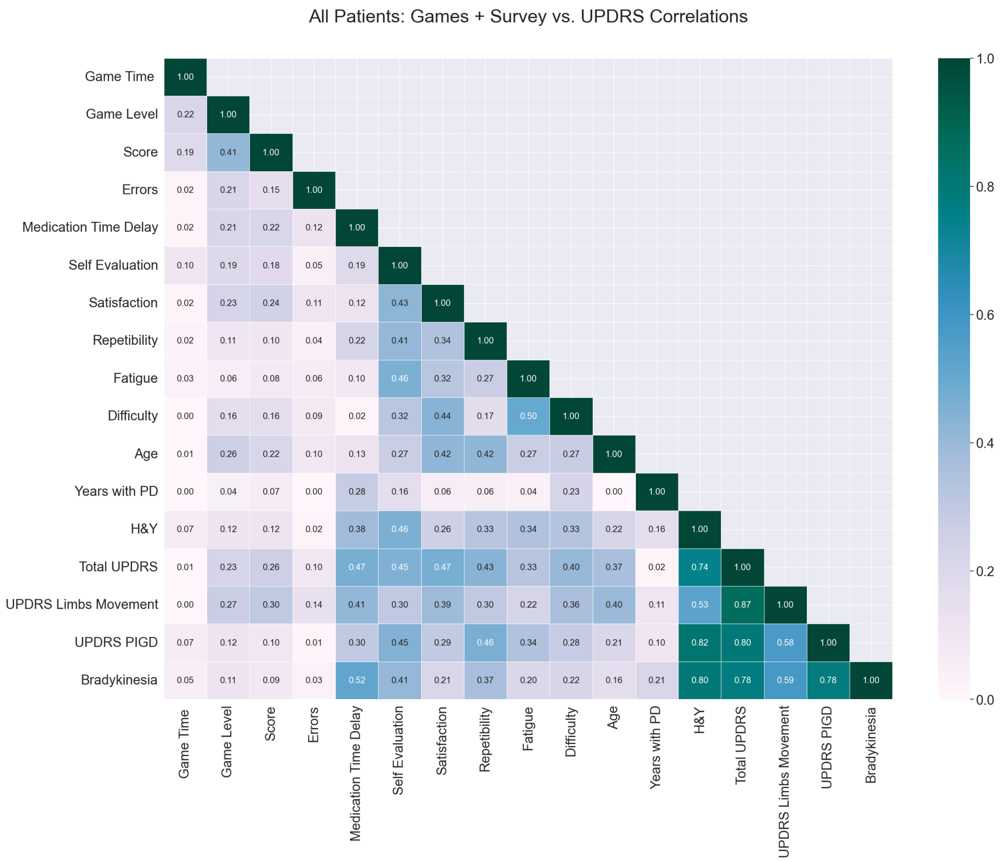

Figure 3.53: Global correlation heatmap involving game data, survey data and UPDRS clinical evaluations.

and correlated the most with level and score, two key metrics to evaluate patient learning curve.
Below, single games correlation matrices will be displayed and analyzed. Since survey responses were collected once per session and clinical condition were evaluated only one time before the data collection, focus will be put mainly in game-specific metric correlations with these two other classes of data.

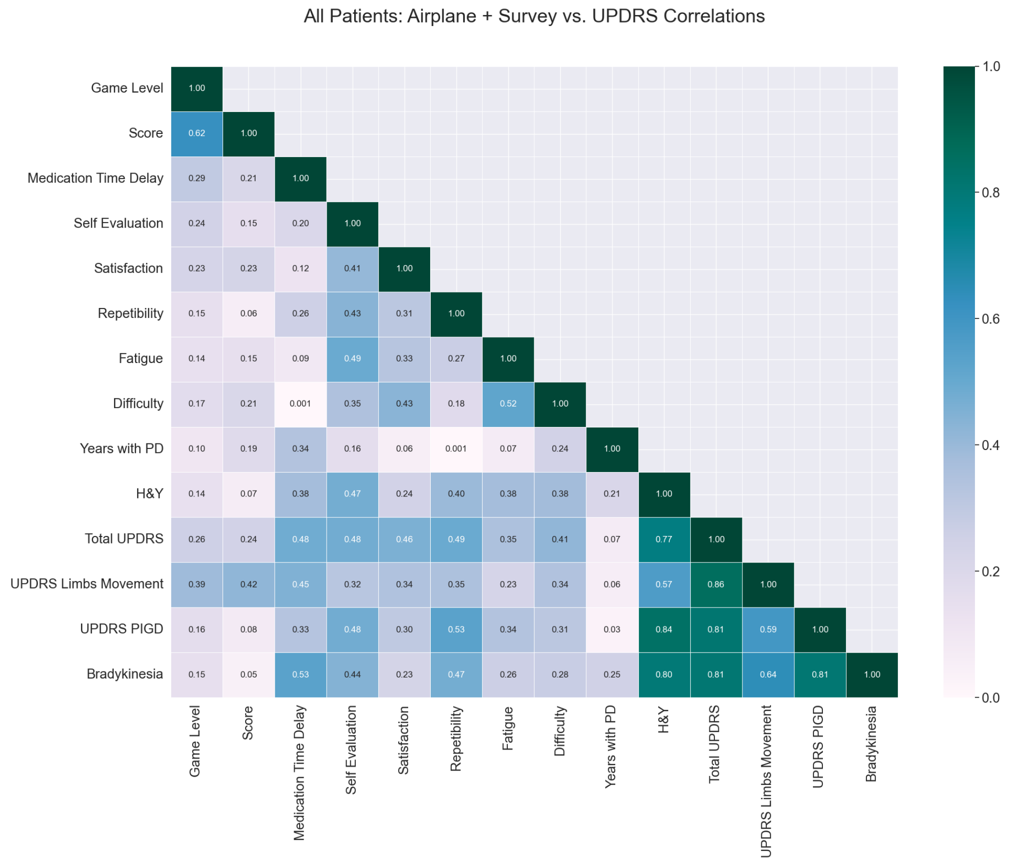

Figure 3.54: Airplane-specific correlation heatmap involving game data, survey data and UPDRS clinical evaluations.

Considering Figure 3.54's Airplane correlation heatmap, it is easily noticeable how score and game level metrics are similarly linked with survey and clinical items as in precedent global heatmap. Game time and error are not shown since for this exergame they were not significant: the first was heavily linked with the velocity but, for the level formulation, could have been misleading, while the latter was mostly non-informative due to the possibility to commit errors only in more challenging levels. However, also in these, values were close to zero on average and thus not selected for the heatmap. The key insight is that UPDRS is well linked with game metrics, the limb movement sub-score in particular. H& Y scale instead seems to be less correlated. Game level is in general more connected with other items, while score is less linked, with the exception of difficulty, year since diagnosis and limbs movement. The two examined game indices are highly correlated. While

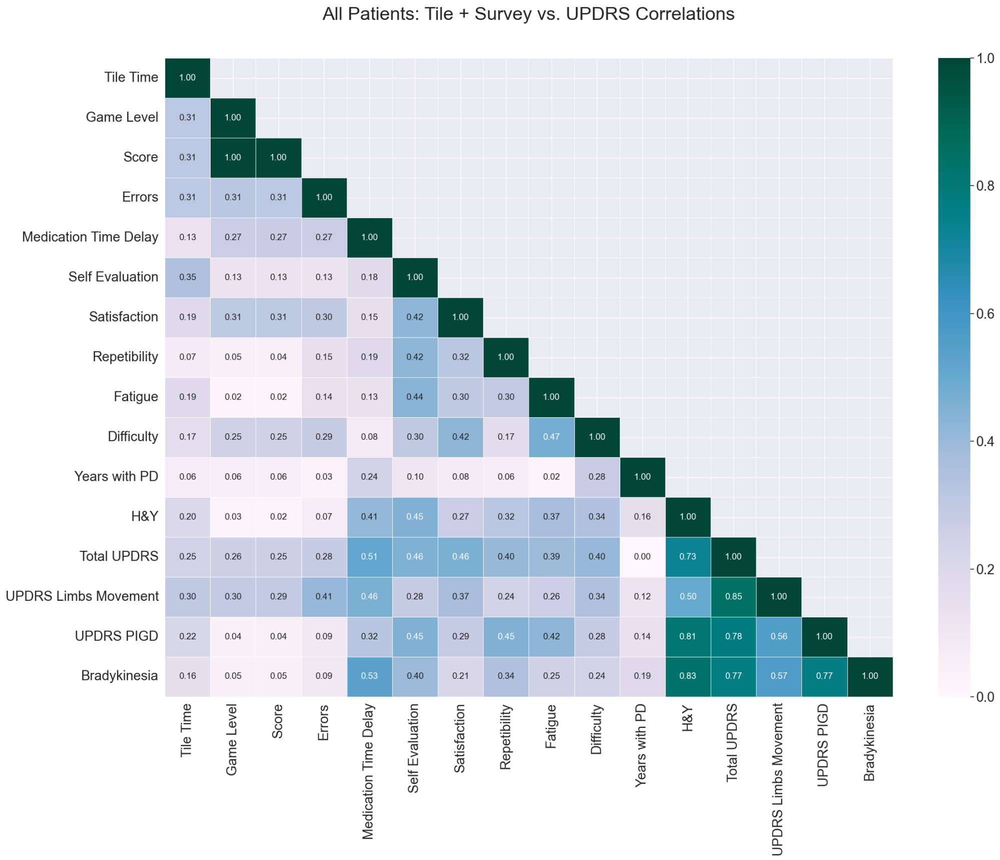

Figure 3.55: Tile-specific correlation heatmap involving game data, survey data and UPDRS clinical evaluations.

examining Tile's heatmap in Figure 3.55, it is immediately clear that score and game level are deeply tied. These metrics, together with errors made, seems to have affected significantly questionnaires' difficulty and satisfaction and to be linked with medication's timing. On the other hand, the perceived condition expressed in self evaluation parameter appear to influence the average time needed to press a tile. UPDRS' metric regarding limbs movement and its total score are fairly correlated to game metrics.
By focusing on tile time, it looks like this parameter is correlated also with bradykinesia, PIGD and H&Y classification. For the Skewer heatmap analysis presented in Figure 3.56, only game metrics reported are game time and game

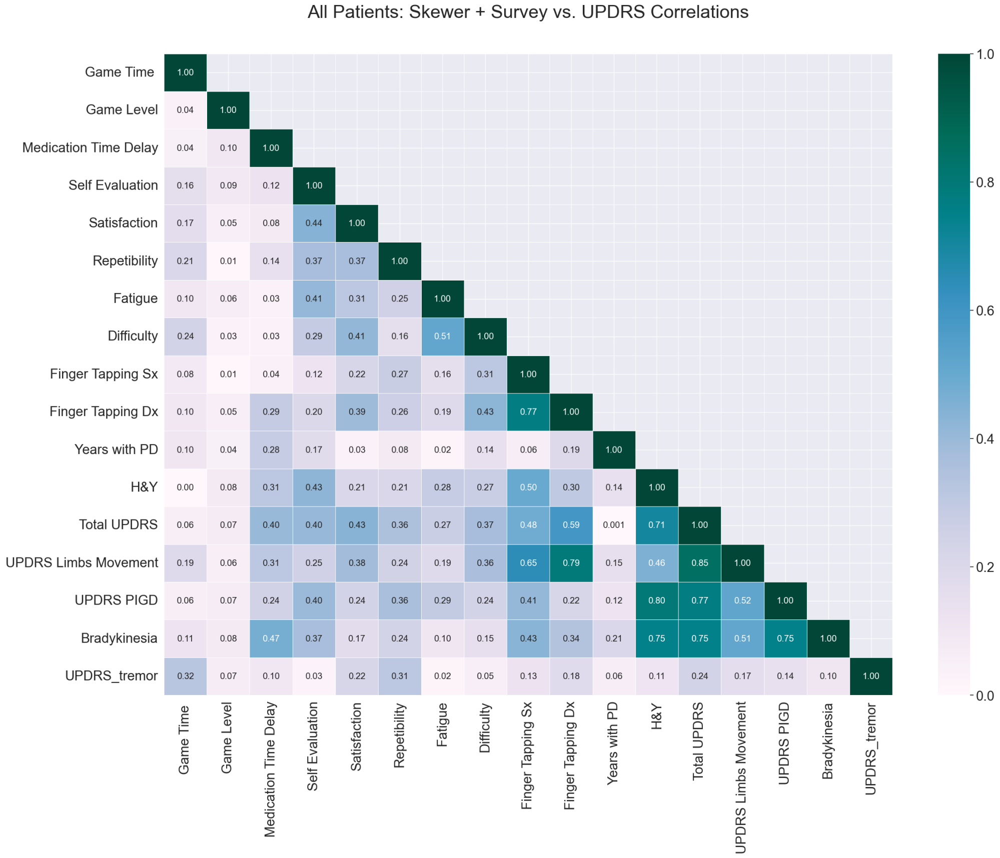

Figure 3.56: Skewer-specific correlation heatmap involving game data, survey data and UPDRS clinical evaluations.

level, since score and errors were respectively identical and close to zero on average. Two new clinical parameters are displayed, which are the tremor and the finger tapping evaluation, divided into left and right sides. Game level, due to the steep learning curve described in previous sections is uninformative, while game time looks correlated with most survey answers and to the clinical metrics of limbs movement and tremor. Game-specific newly included clinical tasks of finger tapping appear to influence survey parameter differently: the right evaluation affinity with those items is on average 10% higher with respect to the left. Finger tapping parameters result to be correlated also with other clinical meters. Tremor instead is only linked clearly with satisfaction and repeatability indices, closely related to finger tapping and total UPDRS scores. As for the Skewer, Ski's heatmap in

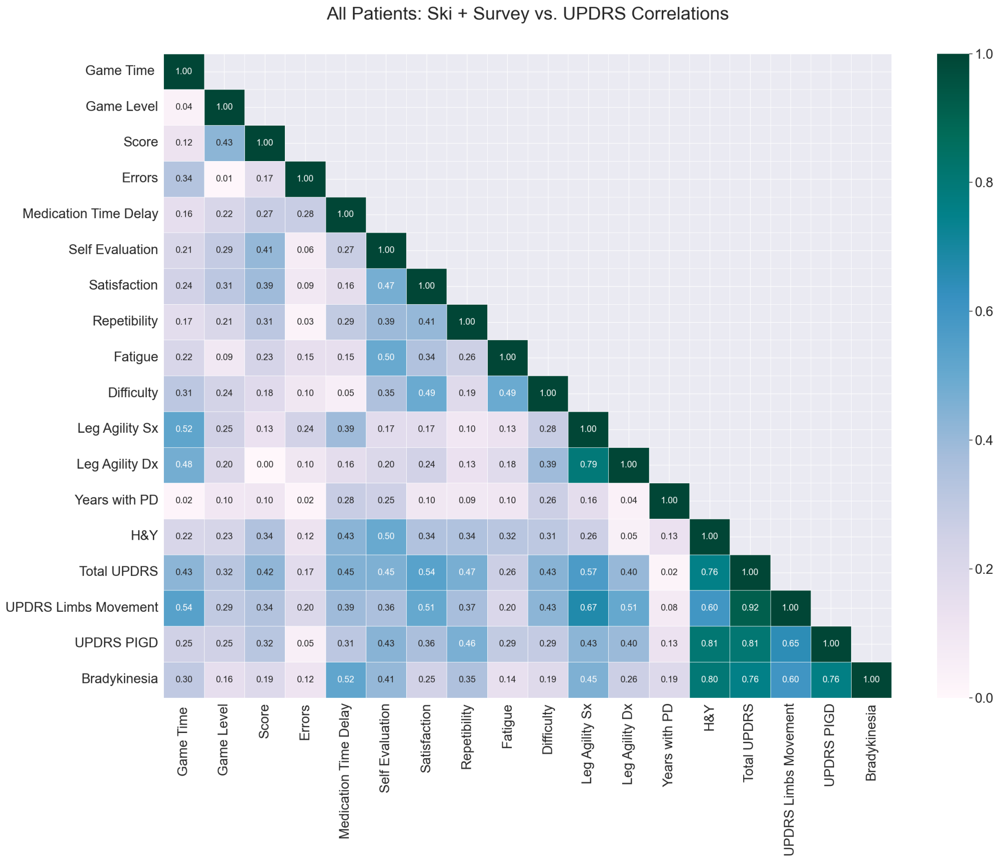

Figure 3.57: Ski-specific correlation heatmap involving game data, survey data and UPDRS clinical evaluations.

Figure 3.57 displays a game-specific clinical item: leg agility, divided in left and right evaluations. These parameters are well correlated with other clinical metrics, always differently for the same parameters between left and right. The left-leg

item exhibited, on average, 10% stronger correlations. The new clinical item is significantly correlated with game time, which makes sense overall.
The game metrics analyzed are game time, level, score and error: the first three are relatively connected with all survey and medical items, while the fourth can be considered in general uninformative. Game time in particular is influenced by limbs movement task, bradykinesia and total UPDRS value. Notably, the correlation between game time and errors reflects a plausible mechanism: every time a patient collided with the sides of a track, he lost velocity, resulting in longer completion executions.

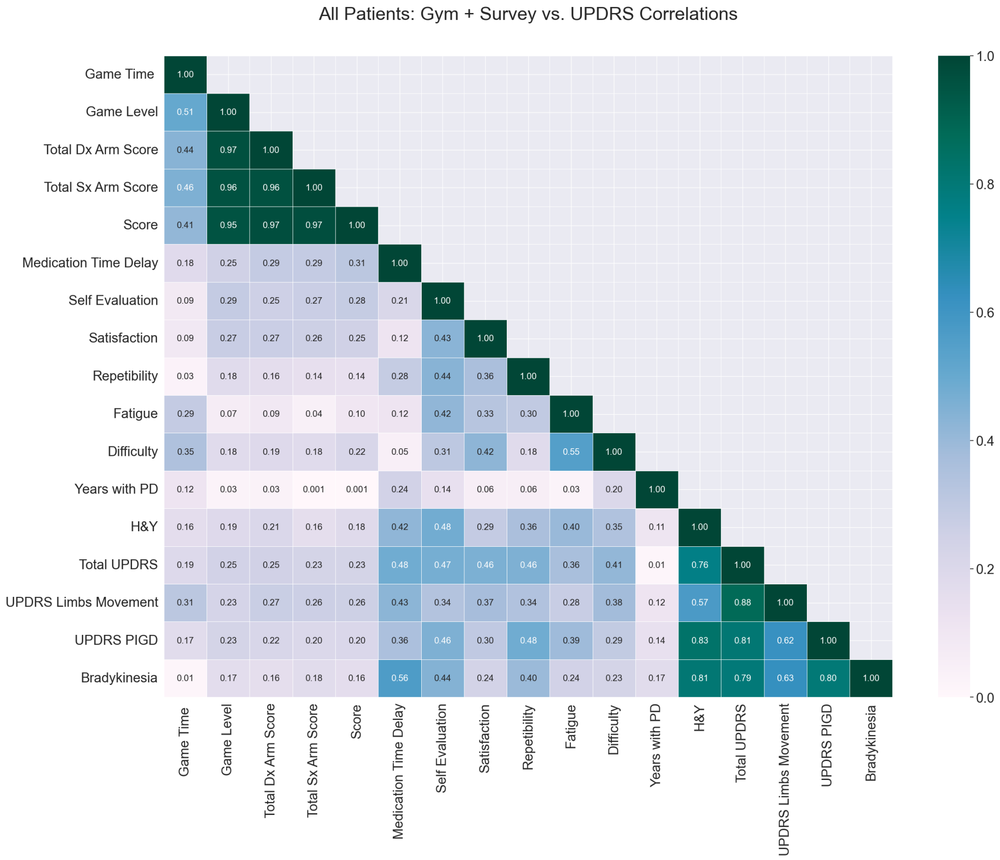

Figure 3.58: Gym-specific correlation heatmap involving game data, survey data and UPDRS clinical evaluations.

Finally, Gym's heatmap is considered in Figure 3.58 for the analysis. Considered metrics include game time, level and score, divided into total left and right arm scores. By a first inspection, looks clear that there are no significant differences between arms. Performances seem to be influenced by self evaluation and appear to influence the responses on satisfaction and repeatability (if excluding game time) and on fatigue and difficulty (if considering only game time). Limbs movement metric, together with PIGD and total UPDRS are the most correlated clinical evaluations. Years since diagnosis showed negligible correlation across variables.

3.7 Special Patients Analysis

In previous sections of this chapter, three patients were designed as "special". In particular they were referred to as follows:

- **Patient 14 - Best**: subject that exhibited the steepest learning curve and most positive game metrics. Clinical evaluations, displayed in Figures 3.59 and 3.60, reflect his performances.
- **Patient 17 - Most Severe**: participant which, accordingly with his highest total UPDRS assigned, concluded the study period with almost no improvement in major game parameters, resulting in far from the mean survey answers.
- **Patient 18 - Outlier**: One participant showed substantial mismatches between clinical severity ratings and actual in-game performance, consistently ranking as a top performer despite a higher UPDRS score. This inconsistency may be explained by the timing of the clinical assessment, which was conducted roughly 210 minutes after the last Levodopa dose, likely coinciding with an “off” state characterized by more severe motor symptoms and possible dyskinesia. In contrast, gameplay sessions occurred predominantly during the “on” state, within 66 to 100 minutes after medication intake.

Table 3.3: Average Game Metrics per Special Patient

ID	AS (%)	AL #	TT (s)	TE #	SKT (s)	SL #	ST (s)	SE #	GL #	GS (%)
14	90.0	6.3	3.5	0.4	71.9	6.8	123.0	0.9	2.7	92.3
17	73.0	0.0	9.2	9.9	307.0	0.2	439.0	3.1	0.9	48.3
18	84.0	4.8	4.0	0.5	59.1	5.6	161.0	1.4	2.1	78.9
OA	**79.0**	**3.7**	**5.3**	**2.9**	**101.0**	**4.9**	**208.4**	**2.6**	**1.9**	**71.0**

Note: Metric prefixes indicate the game—A = Airplane, T = Tile, SK = Skewer, S = Ski, G = Gym. Metric suffixes denote: S = score (percentile), L = level, T = time (seconds), E = errors. OA = overall average.

In order to compare these patients, a focused summarization on average game metrics, main clinical evaluation and survey responses in showed below. Firstly games metrics results are displayed in Table 3.3:

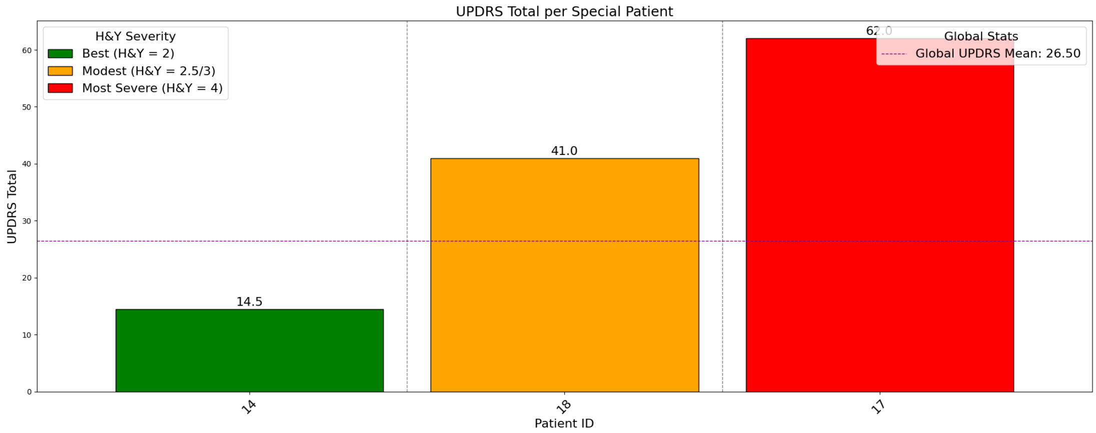

Figure 3.59: Total UPDRS values of special patients, colored by H&Y scale.

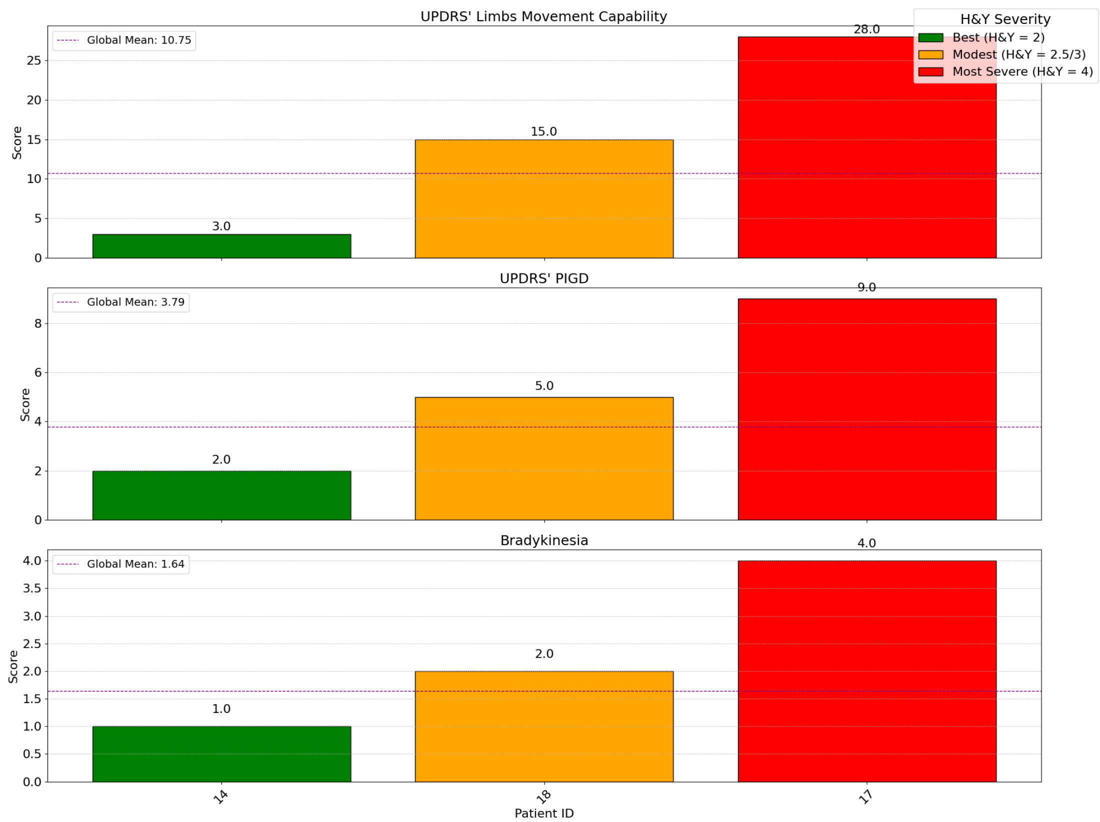

Figure 3.60: UPDRS' main motor features values per special patient, colored by H&Y scale.

- **Patient 14**: exhibited the strongest overall performance, reaching the highest levels in all games (e.g., AL = 6.3, SL = 6.8) and obtaining elevated scores in both Airplane and Gym (AS = 90%, GS = 92.3%). Reaction times were rapid across games (TT = 3.5s, SKT = 71.9s, ST = 123.0s) and error rates remained negligible (TE = 0.4, SE = 0.9), suggesting effective motor planning and sustained engagement throughout the sessions.

- **Patient 18**: demonstrated a moderate performance profile, with high scores (AS = 84%, GS = 78.9%) and relatively high game levels (AL = 4.8, SL = 5.6). While gameplay times were slightly elevated (SKT = 59.1s, ST = 161.0s), they remained within functional range. Notably, tile error rates were minimal (TE = 0.5), but errors increased within the Ski context (SE = 1.4).

- **Patient 17**: struggled significantly across most metrics. Score percentages and game levels were the lowest (AS = 73%, GS = 48.3%; AL = 0.0, SL = 0.2), indicative of limited progress or engagement and gameplay durations were considerably prolonged (SKT = 307.0s, ST = 439.0s), accompanied by high error counts (TE = 9.9, SE = 3.1). These values correspond to pronounced motor impairment, difficulties in task initiation and reduced adaptability.

Then in Figures 3.59 and 3.60 main clinical evaluations are displayed. By analyzing them, it is clear how the best, despite not being the subject with lowest total UPDRS (see Figure 3.2), has main features well below group average. Inversely the most severe is also the one with highest main motor features, UPDRS score and H&Y classification.
A less intuitive insight regards the outlier: total UPDRS and all main motor parameters are far above the mean values. Despite that, patient was labeled as a top performer, being a fast learner able to reach all maximum levels with notably positive other games' indicators. Finally, a brief summary of survey data of these patients is reported in Table 3.4.

Table 3.4: Average Questionnaire Answers for Special Patients: self-evaluation, satisfaction, repeatability, fatigue and difficulty.

ID	**Self-Ev.**	**Sat.**	**Rep.**	**Fat.**	**Diff.**
14	5.0	5.0	5.0	1.6	1.3
17	3.1	3.5	3.9	4.2	3.5
18	3.9	4.0	4.0	1.9	1.7
Overall Avg.	**3.86**	**4.22**	**4.53**	**2.89**	**2.48**

The focus of this section is to compare special patients in more informative exergames' performances. In order to do so, patient-scoped correlation matrices are shown below. In each of them, as for the global matrices presented in previous section, also survey relationship are considered.

3.7.1 Special Patients Correlation Heatmaps

Games for which patient-specific correlation matrices are shown are Airplane and Ski: these proved to be the two most difficult (and informative in general) games. Alongside intra-game correlations between exergame metrics, also the survey connection with games is displayed. It is important to remember that questionnaires were filled only at the end of each session, so answers are not directly connected with specific games.
Clinical evaluations are not present in successive plots since no variability is present intra-patient, thus no correlation: neurologist derived the medical profiles just once before the collection process started.
Due to the nature of questionnaires, asked to be filled just once per session after playing the games, survey parameter correlations across games differ by no more than 10% due to dataset construction. Since all rows correspond to a certain attempt of a game, the only way to insert survey data in it was by putting the same session answers on all exergames multiple tries given a session and a patient. Result of this procedure is the following: if a certain patient in a specific session did more total tries than the average of tries done in other sessions, survey data in current one will weight more. For this reason, intra-survey correlation will not be considered.

Airplane exergame

As seen in previous sections, Airplane most informative game parameters are score and game level. For completion, also game time and errors are reported, but will not be considered. From the inspection of correlation heatmap for patient 14's Airplane performances in Figure 3.61, it is clear how the two game metrics are positively linked. In particular, for this subject, it appears that only fatigue is directly proportional with score and game level, while difficulty is negatively correlated. Despite the negative and quite strong connections with difficulty are too extreme due to the variability of data, this reflects the intended design goal to which these exergames aspire: purpose is to increase fatigue while keeping difficulty contained, ensuring patients have the possibility to rank up without encountering insurmountable difficulty barriers. For the case of individual number 17 displayed in Figure 3.62, it is immediate to see that since he never surpassed base level, no game level and errors (which were possible only on successive difficulties) are shown. Since game time is not informative in Airplane exergame, only score is discussed. This metric correlates positively with satisfaction but negatively with replayability.

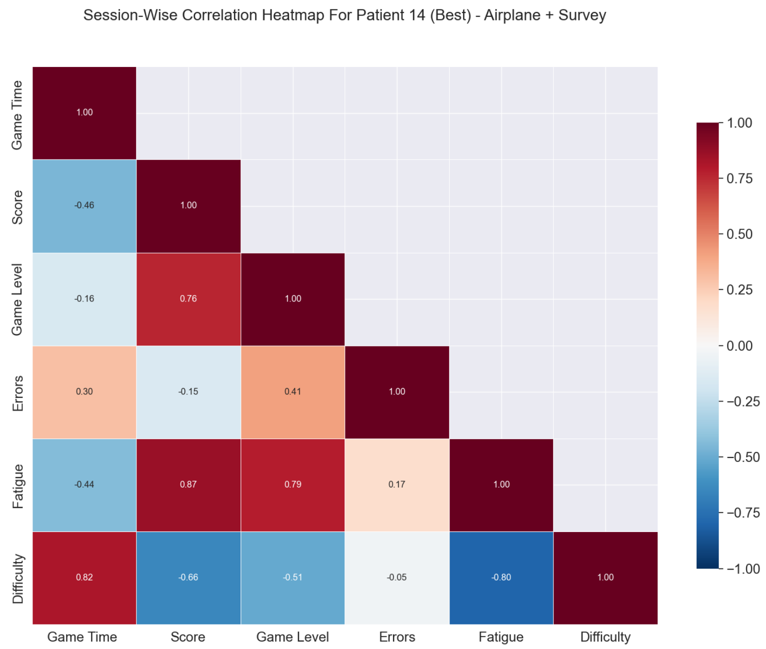

Figure 3.61: Special patient number 14 correlation heatmap - Airplane game.

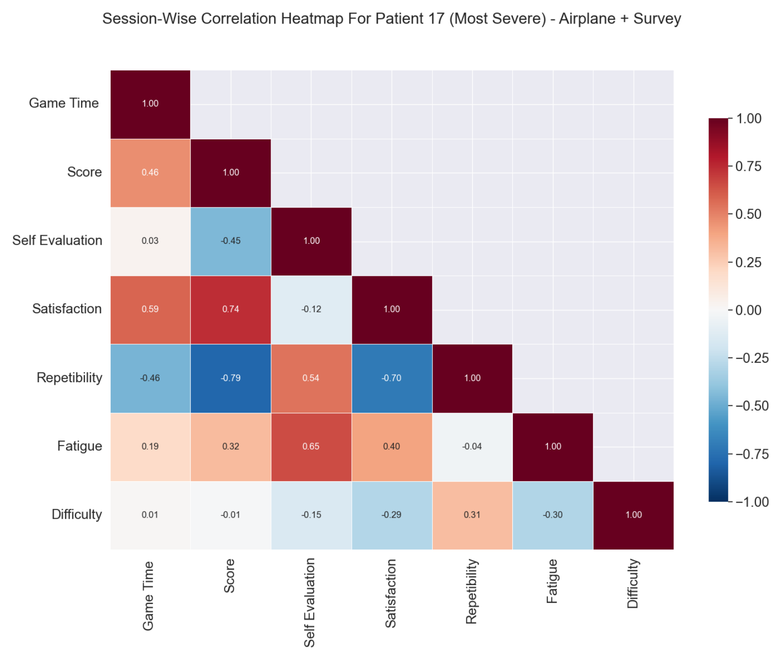

Figure 3.62: Special patient number 17 correlation heatmap - Airplane game.

This might be caused, as seen in Table 3.4, by low overall mean patient scores, which are then linked to low satisfaction values and high repeatability evaluations. Fatigue is quite positively linked with score, while self evaluation given by patient seems not to have been a factor. Game metric is uncorrelated with difficulty.

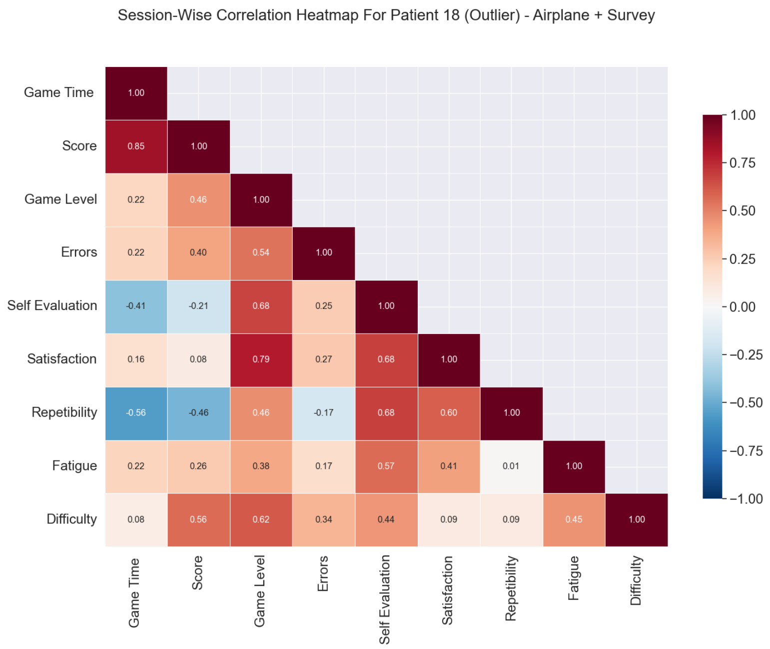

Figure 3.63: Special patient number 18 correlation heatmap — Airplane game.

For subject 18 inspected in Figure 3.63, it is noticeable how score and game level differ in relationships with other survey parameters. This subject reached maximum level within 7 sessions, always keeping a score greater than 80% of the total. This is why score parameters appear uncorrelated, weakly linked or negatively bonded with all survey items. Game level instead is highly correlated with self evaluation, satisfaction and difficulty: as levels increased, more effort was required, enhancing excitement.

Ski exergame

Despite not having been included in game analysis section, score correlations were computed as a supplementary insight on the impact on survey answers generated by collecting coins in the tracks. Inspecting Ski's correlation related to patient 14 in Figure 3.64, evident is how score is connected with game time and level. Since subject average less than an error per session, corresponding metric is uninformative, while game time and level are well positively linked with fatigue. Difficulty is

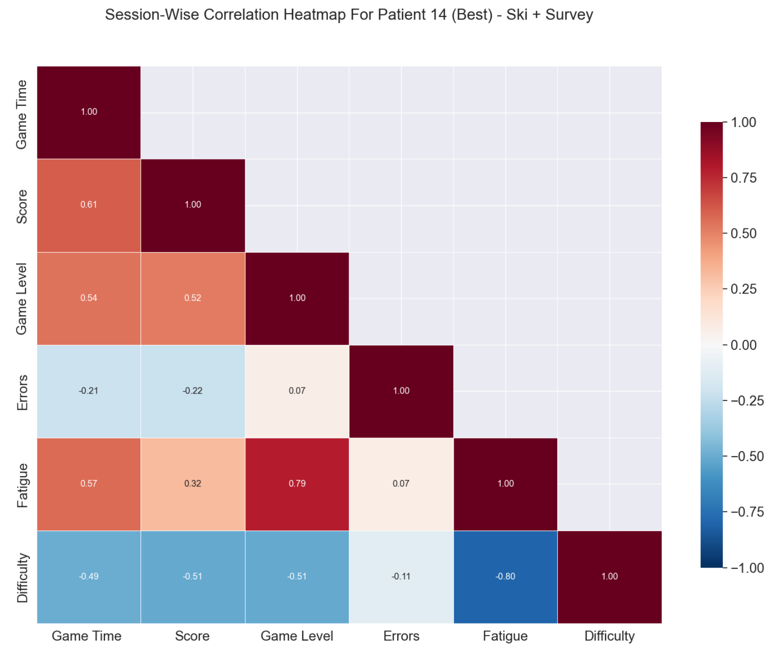

Figure 3.64: Special patient number 14 correlation heatmap - Ski game.

inversely connected with all metrics since subject average value throughout all session was 1.3/5. In Figure 3.65 patient 17's Ski correlation heatmap is displayed.

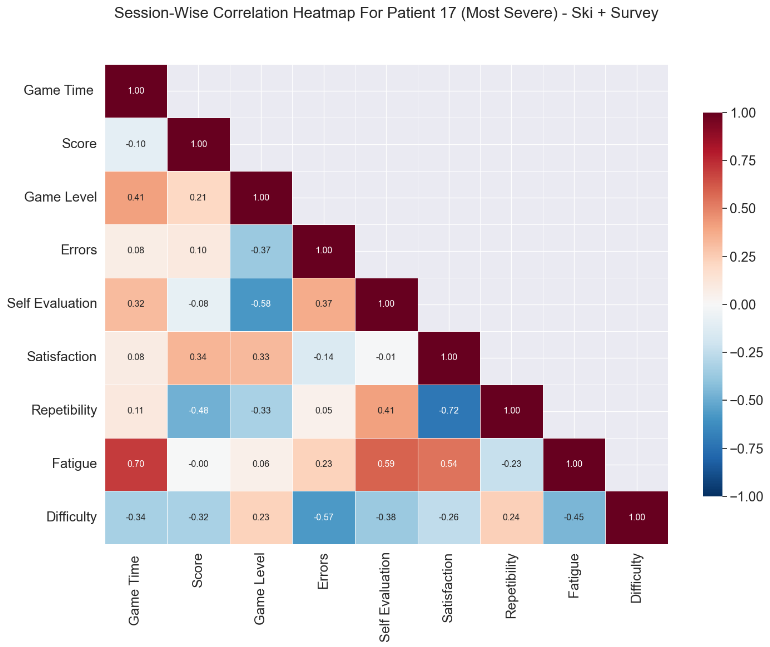

Figure 3.65: Special patient number 17 correlation heatmap - Ski game.

Relevant is the connection between fatigue and game time. Satisfaction seems to be partially linked to score and game level, while errors to self evaluation. Score and game level are negatively correlated to replayability since patient hardly moved beyond base level. Finally, in Figure 3.66, ski game and survey parameters are inspected. Notably, the strong positive correlations between game metrics (game time, score, game level) and perceived fatigue and difficulty indicate how exergame was impactful on those parameters: an increase in level corresponds with higher motor and cognitive challenges. Game time and errors are correctly inversely connected with, respectively, game time and satisfaction. Replayability in general does not appear to be influenced by game performances, which instead appear dependent on self-evaluation scores.

3.8 Session Unsupervised Classification

To categorize rehabilitation sessions into interpretable performance tiers ("basic" vs. "advanced" executions), each patient-session record was treated as an unlabeled observation to which unsupervised clustering was applied. The final dataset comprised 123 rows (one per patient and session) described by 20 features: 15

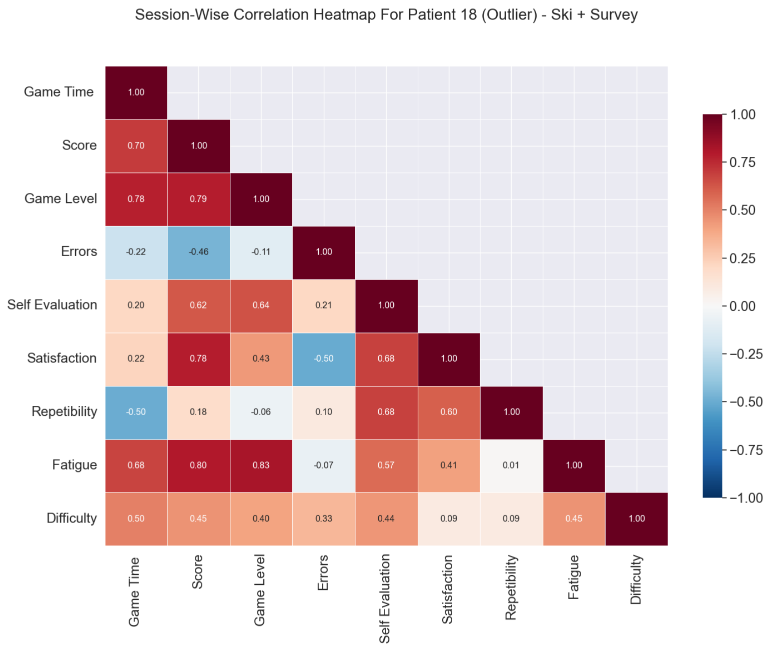

Figure 3.66: Special patient number 18 correlation heatmap - Ski game.

primary performance metrics from exergames (e.g. game level, percentile-normalized score, errors, completion time) and the 5 self-report survey metrics (self-evaluation, satisfaction, repeatability, fatigue, difficulty). Identifier columns ('id_patient' and 'med_delta', regarding time passed since last medicament assumption) were removed prior to analysis and all remaining features were standardized (zero mean, unit variance).

Three families of clustering algorithms were employed:

- Partitional clustering via K-Means: with number of clusters $k = 2 \dots 9$

- Hierarchical Agglomerative Clustering: ranging from $k = 2$ to $k = 9$, consistent with the partitional approach

- Density-Based Spatial Clustering (DBSCAN): considered parameters:

  - **Epsilon (ε)**: $\varepsilon \in \{0.5,\ 1.0,\ 1.5,\ 2.0,\ 2.5,\ 3.0,\ 3.5,\ 4.0,\ 4.5,\ 5.0\}$

  - **Minimum samples threshold**: $\mathit{min\_samples} \in \{2,\ 3,\ 4,\ 5,\ 10,\ 20\}$

In order to select the most appropriate clustering approach, silhouette coefficient[1] for both K-Means and Agglomerative Clustering across described k values were calculated. In addition, a grid search was conducted for DBSCAN over listed values, selecting the optimal configuration by maximal silhouette score (excluding noise points). The resulting performance metrics for all algorithms and parameter settings were summarized in a JSON file (reported in Table 3.5), enabling quantitative comparison of each method's ability to produce cohesive, well-separated session groupings.

[1] calculated utilizing the mean intra-cluster distance between points and the mean nearest-cluster distance

Table 3.5: Clustering Performance Metrics (Silhouette Scores)

Method	k	Silhouette	ϵ	min. samples	noise count
K-Means	2	0.2533	–	–	–
Agg. Clustering	2	0.2115	–	–	–
K-Means	3	0.1794	–	–	–
Agg. Clustering	3	0.2050	–	–	–
K-Means	4	0.1856	–	–	–
Agg. Clustering	4	0.1824	–	–	–
K-Means	5	0.2112	–	–	–
Agg. Clustering	5	0.1923	–	–	–
K-Means	6	0.2099	–	–	–
Agg. Clustering	6	0.1735	–	–	–
K-Means	7	0.1485	–	–	–
Agg. Clustering	7	0.1682	–	–	–
K-Means	8	0.1145	–	–	–
Agg. Clustering	8	0.1775	–	–	–
K-Means	9	0.1185	–	–	–
Agg. Clustering	9	0.1816	–	–	–
DBSCAN	–	0.6785	1.5	3	112

From Table 3.5, best model settings result to be:

- **Kmeans**, $k = 2$
- **AgglomerativeClustering**, $k = 3$

Low values for silhouette were expected due to high inter-individual variability and the gradual nature of patient progression, making strong inter-cluster separation unlikely. K-Means provides a computationally efficient partitional approach, particularly appropriate when spherical clusters are expected and the dataset is continuous and uniformly scaled. Its ability to minimize intra-cluster variance aligns well with the goal of grouping sessions based on overall performance profiles. AgglomerativeClustering complements this by offering a hierarchical perspective, making no assumptions about cluster shape and allowing more flexibility in identifying nested structures among session patterns. It is particularly advantageous for revealing latent transitions or substructure in patient progressions across sessions. DBSCAN's high score for the silhouette is related to its high noise ratio: 112 points out of 123 were not labeled, resulting in an artificially high silhouette score due to limited clustered samples. In general, this model is not suitable for this kind of dataset and thus was discarded: having 20 features, regardless of model parameter

choice, the high-dimensional space consistently led to excessive noise labeling.
Output of unsupervised clustering, obtained using selected best model parameters, were integer-based cluster labels, automatically generated without semantic meaning. To enable interpretation and consistent visual labeling, a custom reordering strategy was applied to map clusters onto a performance spectrum ranging from low to high adaptation.

Feature Weighting Strategy

Each input feature was first classified as either performance-enhancing or impairment-indicative. The attributes `game_level`, `score`, `self_evaluation`, `satisfaction` and `repeatability` were considered positive indicators of engagement and motor-cognitive adaptation and thus assigned a weight of $+1$. Conversely, features associated with difficulty or fatigue (`errors`, `time`, `fatigue` and `difficulty`) were assigned a weight of -1.

Cluster Scoring and Relabeling

For each cluster label produced by the model, the mean value of each feature was computed over all samples belonging to that cluster. These standardized feature means were then combined using the custom weight vector via a dot product:

$$\text{Cluster Score}_i = \sum_{j=1}^{n} w_j \cdot \mu_{ij} \tag{3.3}$$

where w_j is the weight for feature j and μ_{ij} is the standardized mean of feature j within cluster i.

Clusters were then sorted by their total scores in ascending order. The final relabeling remapped the original cluster indices to enforce a consistent semantic order (e.g., Cluster 0 = “Initial”, Cluster 1 = “Setup”, Cluster 2 = “Intermediate”, Cluster 3 = “Advanced”).
This methodology allowed for unsupervised cluster labels to be post-processed into interpretable categories aligned with rehabilitation progression stages without requiring external supervision or manual annotation.
After reordering of clusters and relabeling, for each k-algorithm configuration, each single session point clustering was used to compute the distribution of cluster labels within each session. Two complementary visualizations were produced: separate bar charts (one subplot per cluster) showing the percentage of session records assigned to that cluster and stacked-bar plots illustrating the full composition of clusters per session. Only stacked bar plots are reported in Figures 3.67 and 3.68. This approach was crucial to inspect session-by-session consistency (e.g. does Session 3 predominantly fall into the “Intermediate” cluster?).

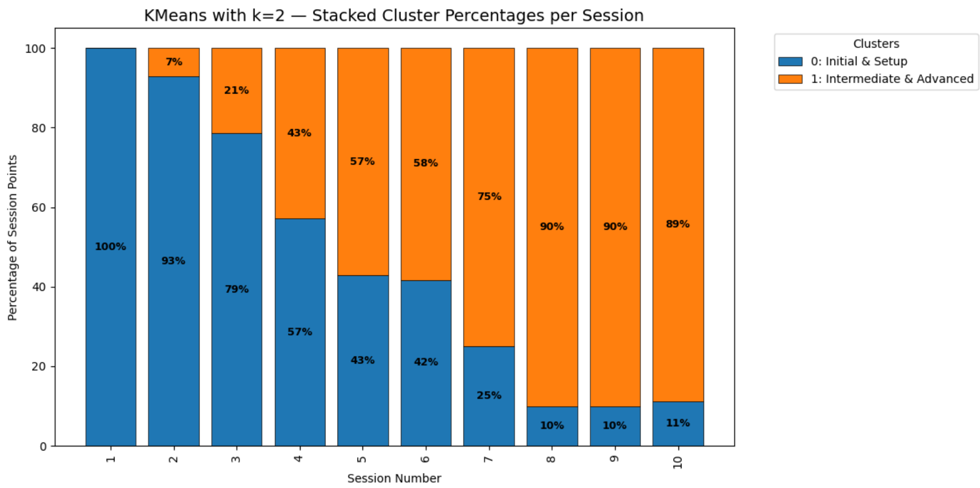

Figure 3.67: Stacked Bar Plots - Percentage of clusters per session, based on KMeans, $k = 2$.

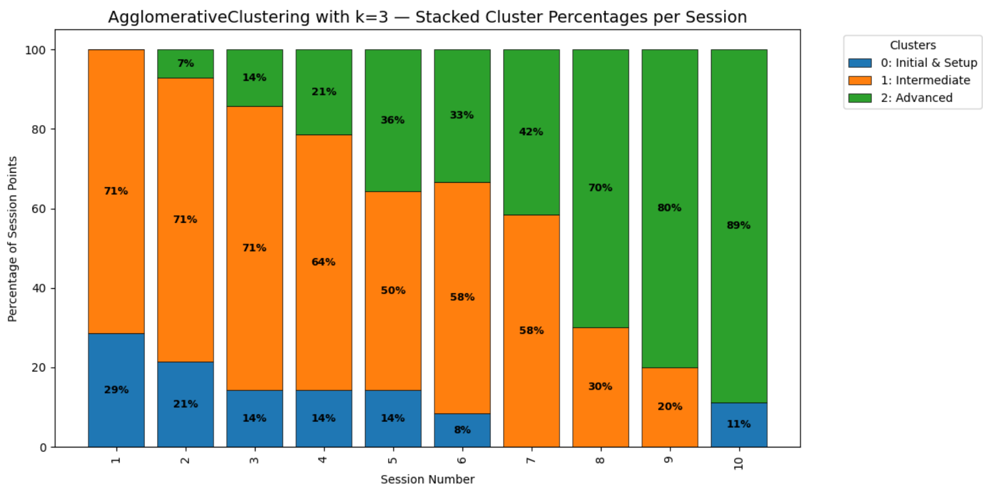

Figure 3.68: Stacked Bar Plots - Percentage of clusters per session, based on AgglomerativeClustering, $k = 3$.

From the comparative inspection of Figures 3.67 and 3.68, it is clear how increasing cluster numbers (thus the corresponding difficulties and learning abilities), distributions shift towards later sessions. In particular, for the 2 clusters selected scenario, sessions 4, 5 and 6 are the most uncertain ones with similar percentages of cluster points. This is expected, since these sessions are the ones in the middle, where best performers reached peak level in most games while more severe subjects were still dealing with lower difficulties. For the 3 clusters obtained with the Agglomerative method, sessions appear to be more clearly separated, apart from session 7: this is the first advanced session, even if still transitional, since only in two cases a personal best or global peak level was reached by any participant in any exergames. This unsupervised classification pipeline thus provided a fully data-driven means of labeling sessions from "initial" to "advanced", without relying on external annotations: this offers an insight into the natural stratification of rehabilitative performance over time. Additional clustering using aggregated single-session features (mean or max) was evaluated. Since by aggregation the granularity of possible different clustering within a session gets lost, results are perfectly clustered. This causes a loss of information, removing all variability coming from different learning curves from patients. For this reason, these solution were discarded. Finally, after computing the per session clusters distribution, the median for each cluster for the 2 solutions with highest silhouette scores was computed. Results are reported in Tables 3.6 and 3.7 and show clearly how Ski's game performances are the most heterogeneous among all exergames, for all clusters and both solutions. Purpose was to find, for each exergame, the most suitable session in relation to each cluster, so that mean selected session features values could be calculated to find the prototype for an initial/setup or more advanced session.

Table 3.6: KMeans, $k = 2$ – Median Session per Cluster–Exergame Pair Based on Minimum Distance to Centroid

Cluster	**Exergame**	**Session Number**	**Distance to Centroid**
0	A	3	0.000
0	S	3	4.243
0	G	2	0.000
0	SK	2	0.000
0	T	3	0.400
1	A	9	0.100
1	S	5	3.211
1	G	4	0.000
1	SK	8	0.250
1	T	9	0.000

Note: For each cluster–exergame pair, the listed session corresponds to the observation with minimum Euclidean distance to the cluster centroid and therefore serves as its empirical prototype. For example, Cluster 0's behavioral pattern is best approximated by Session 3 of Airplane, Session 3 of Ski, etc. Column abbreviations: A = Airplane, S = Ski, G = Gym, SK = Skewer, T = Tile.

Table 3.7: AgglomerativeClustering, $k = 3$ - Median Session per Cluster–Exergame Pair Based on Minimum Distance to Centroid

Cluster	Exergame	Session Number	Distance to Centroid
0	A	3	0.000
0	S	2	3.162
0	G	2	1.000
0	SK	2	4.000
0	T	4	2.256
1	A	5	0.000
1	S	4	4.153
1	G	2	0.000
1	SK	5	0.000
1	T	6	0.100
2	A	6	0.000
2	S	9	3.041
2	G	4	0.000
2	SK	9	0.000
2	T	10	0.100

Note: For each exergame and cluster, the session listed corresponds to the observation with the minimal Euclidean distance to its cluster centroid. These sessions serve as empirical prototypes—meaning that a typical Cluster 0 pattern is best approximated by the feature values in Session 3 of Airplane, Session 2 of Ski and so on. Column abbreviations follow: A = Airplane, S = Ski, G = Gym, SK = Skewer, T = Tile.

In both cases, centroid distances are reported. Non-zero distances are all related to exergames where relevant metrics include time, a factor of great variability, especially if compared with integer number values as levels or scores. Interesting is to notice how the reference sessions change intra and inter clusters, depending on the game. For the two methods, Gym's fast learning curve proved to be a factor for the prototypes, since, to distinguish between type of sessions, close and within half process sessions are sufficient.

Table 3.8: Cluster-Averaged Game Metrics Based on KMeans, $k = 2$

Cluster	**AS** #	**AL** #	**ST** sec	**SS** #	**SE** #	**SL** #	**GS** #
0.0	5.16	2.71	237.00	7.50	4.86	2.71	3893
1.0	8.18	6.70	246.84	7.96	4.71	5.43	5561

Cluster	**GL** #	**SKT** sec	**SKL** #	**TE** #	**TL** #	**TT** sec
0.0	1.07	153.68	3.21	7.36	1.00	5.19
1.0	1.71	83.05	4.50	4.30	1.60	4.58

Note: Values represent cluster-specific game metrics averaged within each prototype exergame-specific session. A = Airplane, S = Ski, G = Gym, SK = Skewer, T = Tile; S = score, L = level, E = errors, T = time. TT is tile time, not total tile game time.
Cluster 0: Initial & Setup; Cluster 1: Intermediate & Advanced.

Table 3.9: Cluster-Averaged Game Metrics Based on AgglomerativeClustering, $k = 3$. These feature values can be used as a prototype for an average, cluster-relative session.

Cluster	**AS** #	**AL** #	**ST** sec	**SS** #	**SE** #	**SL** #	**GS** #
0.0	5.16	2.71	188.88	6.34	1.93	0.07	3893
1.0	5.41	3.36	236.79	7.63	4.21	4.50	3893
2.0	6.25	3.58	259.25	9.15	2.10	9.20	5561

Cluster	**GL** #	**SKT** sec	**SKL** #	**TE** #	**TL** #	**TT** sec
0.0	1.07	153.68	3.21	6.79	1.20	5.75
1.0	1.07	102.75	4.64	6.92	1.42	5.17
2.0	1.71	72.15	4.50	4.00	1.60	5.30

Note: Values represent cluster-specific game metrics averaged within each prototype exergame-specific session. A = Airplane, S = Ski, G = Gym, SK = Skewer, T = Tile; S = score, L = level, E = errors, T = time. TT is tile time, not total tile game time.
Cluster 0: Initial & Setup; Cluster 1: Intermediate; Cluster 2: Advanced.

Tables 3.8 and 3.9 present cluster-averaged metric profiles for all exergames, employable to describe a cluster-specific prototype session, derived respectively from the KMeans ($k = 2$) and AgglomerativeClustering ($k = 3$) methods. These values serve as reference patterns for session classification across exergames.

Chapter 4

Discussion

This chapter provides synthesis of the principal findings presented in chapter 3 and offers a critical interpretation of their implications. The results discussed are derived from extensive analyses of heterogeneous data collected continuously during Phase 2 of the project. This discussion is structured into four main sections:

- Usability and patient engagement.
- Correlation between game data, survey answers and clinical evaluations.
- Implications on telemedicine
- Limitations

4.1 Usability and Patient Engagement

Analysis of session-aggregated game results highlights clear trends that demonstrate the usability of the exergames and the progressive learning of participants. Across all exergames, performance-related metrics designed to be maximized generally displayed an increasing trend, while those intended to be minimized showed a decreasing pattern. Some parameters values like Tile's tile time and errors (Figures 3.16, 3.17) or Ski's game time and errors (Figure 3.20 and 3.21) remained relatively stable across sessions despite the progressive increase in difficulty, suggesting overall performance improvements.
Session-aggregated questionnaire responses showed no major directional trend for individual items. However, their consistency provides valuable insight into how the study and exergames were perceived by participants. Average per session values of satisfaction and repeatability remained high (respectively 4.22/5 and 4.53/5, see Figures 3.49 and 3.50), while fatigue and difficulty low (respectively 2.89/5 and 2.48/5, see Figures 3.51 and 3.52). Notably, the overall mean perceived difficulty

remained close to 50% of the maximum possible, indicating that progression in game levels maintained sufficient cognitive challenge, promoting more an increment of fatigue rather than of difficulty.
Games linked with self-reports evolution trends, steeper or not depending on the exergame-specific cognitive and physical stimuli, led to the definition of up to three (depending on the method and number of clusters) "prototype" sessions, reflecting the stage of study: Initial & Setup, Intermediate or Advanced. This confirms an evolution from familiarization to mastery for the games and is useful to define a personalized strategy for a new patient, able to start the rehabilitation process with the most suitable game level for his condition.

Airplane

The Airplane exergame emerged as the most challenging overall: only 8 out of 14 participants reached the maximum level (see Figure 3.24), typically requiring on average more than seven sessions (Figure 3.25). Considering the cumulative distribution in Figure 3.28, this exergame can be considered pretty versatile, with a peak of patients dealing with 7 out 8 levels in the same session.
Airplane's metrics showed a monotonic trend and great variability among participants. Particularly, inter-tertile learning curves throughout sessions differ significantly (Figure 3.26): less severe patients had a steeper curve ending in an on average maximum level plateauing since third to last session. Medium severity group's curve reflects an overall slower learning process (with average differences of group peak level reached throughout session between 2.5 and 5 with respect to low severity cohort, starting from third session) ended still on ceiling level. Most severe group mean advancement was characterized by relevant standard deviations and was notably slower, never surpassing medium difficulty level (4/8). In conclusion, this solution, given the 8 levels division, is suitable for heterogeneous patients: can be considered an easy to learn, hard to master exergame.

Tile

This exergame was labeled as "medium" difficulty: 11 out of 14 patients succeeded in reaching maximum level (see Figure 3.24). Hidden complication was precision, which directly involved upper arms ability to move at a controlled pace. This is why, combining Figures 3.25, 3.3 and 3.7, results indicate a clear tendency for less severe patients to master exergame in 5 sessions on average, while more severe cases are unable to reach peak level. Mean tile-press times declined steadily after session 3 (Figure 3.16) and error rates stabilized at end near the group average (Figure 3.17). In conclusion, despite the performance gaps inter-tertile, all patients showed progression, proving Tile usability and clinical validity.

Skewer

Skewer was ranked as easiest game, with all but one (with severe upper-limbs movement limitations that caused impracticability to deal with the exergame) participants able to reach mastery and in just 4 sessions at most. Rapid attainment of peak levels by session 3 (Figure 3.34) and low intra-patient variance (Figure 3.32) indicate that Skewer posed minimal barriers, making it highly accessible even to more severely affected individuals. Despite the early in sessions level mastery, learning and adaptation process continued until the end, with a negative trend for completion time (Figure 3.18).

Ski

This exergame resulted to be the second most difficult, with 10 out of 14 participants mastering it. As for the Airplane, the total number of levels available granted adaptability for all profiles, with at most 7 modes played in the same session. Learning curve was monotonic for the majority of the study, considering participants individually and grouped by clinical condition similarities. This solution was the only one which required full body activation, requiring effective motor coordination. For this reason, game times and errors standard deviations in each session were high and fluctuant (Figures 3.20 and 3.21). Variability among high-severity patients suggests that Ski effectively discriminates clinical stages, which top performers in game metrics being the subjects with lower criticality profiles. In conclusion, this solution can be applied to the examined group in total. Ski results, however, more suitable as a mid-phase exergame within a progressive rehabilitation protocol, since mostly only less severe participants are able to start dealing with it since beginning.

Gym

Gym was ranked second to last, with a final number of 11 out of 14 players that reached maximum level, taking, on average, 5 sessions (Figures 3.24 and 3.25). Progression was consistent among each UPDRS tertile, showing a clear upward trend. Only most severe group members were not able to reach mean ceiling level in last session (Figure 3.38) but, however, successfully surpassed basic mode by half of the study (Figure 3.40). Metric outcomes from Figure 3.22) and Figure 3.23 confirm that Gym provided a suitable motor-cognitive workload without impeding patient engagement, more indicated for advanced disease stages. For this solution, additional high performers (patients 10 and 21) emerged (Figures 3.12 and 3.13), showing how challenge mostly regarded resistance and willing of the individual to execute a more fatiguing modes.

4.2 Insights on Correlation Between Game Performances, Clinical Scores and Survey Answers

Spearman correlation analyses revealed robust associations between game level/score and UPDRS total and sub-scale scores (Figure 3.53). Airplane and Ski games reported the higher correlations with clinical scores, in particular the total UPDRS value, the sub feature regarding the limb ability to move correctly, bradykinesia and, specifically for the second, leg agility (Figures 3.54 and 3.57). Patients in lower UPDRS tertiles consistently achieved higher maximum levels with less variability, whereas higher-severity tertiles exhibited wider dispersion in both objective metrics and self-reports. Composite patient ranking based on multi-modal metrics (self-evaluations and game performances, Table 3.2) aligned closely with clinical evaluations: the top six ranked patients, with the exclusion of 18 (the "outlier") and 1 (PD severity mainly derived from vocal impairments), belonged to lower UPDRS stages, confirming that game-based performance metrics serve as valid proxies for functional severity. In addition, these participants reported above overall average mean satisfaction and replayability scores and below overall average mean difficulty and fatigue scores, demonstrating alignment between objective performance, clinical severity, and subjective self-assessment.

4.3 Implications for Telemedicine and Telerehabilitation

These findings support the deployment of webcam-based exergames as scalable, home-based rehabilitation tools. Prototype session profiles enable the definition of adaptive training phases (from initial to advanced) that can be prescribed according to patients' clinical characteristics. High satisfaction and replayability scores imply good patient acceptance, while metric-clinical correlations justify use of in-game data for remote monitoring and tailoring of therapeutic intensity. In addition, games' levels division ensures a proper adaptability for all necessities. The implicit data collection is crucial both for following evolution of disease in patients and for having in availability a an amount of data that grows time by time, facilitating more robust modeling and enabling the use of predictive algorithms to anticipate future session performance.

4.4 Limitations

- **Sample size and dependence.** The inclusion of fourteen participants, predominantly classified as H&Y stages 2-3, limits the extrapolation of findings to more advanced or diverse Parkinson's Disease cohorts. In addition, the small sample size and the dependence among repeated sessions for the same individual rendered the application of supervised learning models impractical for classifying patients by disease stage solely on the basis of game-derived metrics and self-evaluation data. Likewise, unsupervised grouping of patients into clinically meaningful clusters was infeasible for the same reasons. While correlations among features are desirable to some extent, high inter-correlation complicated attempts to identify the most informative metrics. The dependence among consecutive sessions, combined with high inter-individual variability and challenges in robust feature selection, contributed to silhouette scores near 0.2 for session clustering, indicating only modest cohesion.

- **Absence of game-specific questionnaires.** The surveys designed for this study did not include exergame-specific items on usability or perceived difficulty, potentially reducing the precision of subjective feedback related to each game modality. This limitation constrains the interpretability of per-game correlation heatmaps and reduces the value of subjective perception as an input for personalized exergame-based intervention planning.

- **Lack of per-session clinical evaluations.** The study design did not incorporate repeated clinical assessments across the exergame sessions. Consequently, short-term fluctuations in motor status could not be directly aligned with session-level performance data. All clinical evaluations were conducted only once, prior to the start of the data collection phase. This approach likely contributed to the outlier status of Patient 18, whose clinical assessment occurred approximately 210 minutes after the last Levodopa dose, plausibly coinciding with an "off" state characterized by more pronounced motor symptoms and potential dyskinesia. In contrast, all recorded exergame sessions were completed predominantly in the "on" condition, within 66 to 100 minutes of medication intake, which may explain the observed discrepancies between clinical severity scores and in-game performance.

- **Highly variable number of attempts.** The number of trials completed per exergame varied substantially across patients and sessions, which may introduce bias in aggregated metrics and reduce comparability. For example, Gym, Skewer and Ski were typically attempted only a few times per session per patient, whereas Airplane and Tile were often repeated multiple times, either until progression was achieved or the participant requested to advance to the next task. This variation was partly due to structural differences among the exergames themselves, as shorter-duration games inherently required more repetitions to provide adequate practice and learning opportunities.

- **Impossibility to track failed Airplane attempts.** The number of failed attempts in the Airplane game could have represented a meaningful performance indicator. However, efforts to estimate this metric by subtracting the sum of individual execution times from the total session duration produced inconsistent results due to idle periods, despite attempts remove them by inter-attempt intervals subtraction using timestamp differences.
- **Differences in inter-patient attendance.** The frequency and consistency of session attendance varied among participants, with some individuals completing all planned sessions and others discontinuing earlier. This variation further limits the generalizability of progression trends and affects overall means for key metrics. The reduction from an initial 14 participants to 9 in the final session amplified the sample size and dependence limitations noted above.
- **Limited observation window.** The duration of the data collection phase, covering at most ten sessions per patient within a relatively short timeframe, did not allow for the investigation of longer-term adaptation, sustained adherence, or possible fluctuations in motor condition over several months. A longer observation window, ideally extending up to six months, would be necessary to capture more gradual disease progression trends and potential effects that may affect exergame engagement and rehabilitation outcomes.

Chapter 5

Conclusion and Future Work

This work investigated whether webcam and markerless tracking-based exergames can offer a usable, engaging and clinically meaningful solution for telerehabilitation in Parkinson's Disease. Through the continuous monitoring of five custom exergames, supported by detailed session-level gameplay metrics, patient survey responses and clinical scores, this study demonstrates the technical feasibility and patient acceptance of this approach in a cohort with moderate disease stages.

The progression patterns found in objective metrics, alongside consistently high satisfaction and repeatability scores, confirm that the exergames maintained usability and engagement over repeated sessions. In addition, the fact that perceived difficulty remained slightly lower than perceived fatigue, which was overall around 50% of the scale, indicates that the games sustained an appropriate cognitive challenge while progressively increasing motor load through repetition: a balance well aligned with home-based rehabilitation solutions.

The correlations observed between in-game metrics and UPDRS scores, as well as the game trends of patients divided by their corresponding UPDRS tertile, further indicate that gameplay performance can reflect underlying motor status, supporting the potential for remote monitoring.

Unsupervised clustering highlighted distinct prototype session phases, from familiarization to mastery, providing a practical basis for tailoring adaptive telerehabilitation plans to individual motor profiles and disease stages.

However, several limitations must be acknowledged. The small sample size constrained the use of supervised learning methods for classification tasks and reduced the generalizability of correlation and clustering results. The absence of repeated clinical assessments across the study sessions limited the ability to align short-term motor fluctuations with game performance, also likely contributing to outlier effects due to timing discrepancies regarding last medication assumptions between clinical assessments and gameplay sessions. The lack of exergame-specific usability items in the surveys narrowed the precision of subjective feedback at the game level, limiting the possibility for patients-oriented exergames patches. Variability in the number of attempts per session and differences in patient attendance further contributed to inconsistencies in aggregated metrics and modest unsupervised clustering results.

Future work should aim to address these constraints. Scaling up the participant cohort and ensuring consistent session completion would allow for more robust statistical modeling and validation of results, especially if combined with a longer observation window (ideally extending up to six months) during which progressive changes in motor status could become more observable.

A longer timeframe would also increase the total number of sessions per patient, expanding the available dataset and enabling more reliable detection of individual progression trends, sessions clustering and potential adaptation or dropout patterns, offering valuable evidence for refining exergame content and session scheduling to sustain long-term adherence.

Incorporating repeated clinical evaluations per session would enable finer alignment between objective gameplay indicators and motor status. Designing and deploying exergame-specific survey items could refine usability insights and strengthen subjective-objective metric correlations.

From a technical perspective, in order to support completely a telerehabilitation protocol, could be useful to develop new solutions in order to target all PD stages. With larger datasets, obtained by longer observation windows and more participants (since it is a remote solution, this should be a natural consequence once platform is released and accepted as one of the standard rehabilitation solutions), supervised machine learning models could be explored to predict session-level outcomes or dynamically adapt difficulty to maximize benefit.

In addition, as described in chapter 2, the data collection included also gameplay recordings and UPDRS tasks executions, which were not used in this study.
The first could be used in the future to qualitatively assess user engagement, any discrepancies with tracked motion data and specific clinical items like dyskinesia. The recorded sessions may be reprocessed by third parties to extract motion data from the video material without requiring patients to undergo the rehabilitation protocol again.
The latter included video captures of patients performing selected items of the

MDS-UPDRS scale, recorded both before and after each session for clinical comparison. Tasks included Sit-To-Stand, Leg Agility, Gait (from two perspectives, frontal and horizontal), Finger Tapping, Toe Tapping and Hands Movement. This material could be useful for future works to enrich the analysis done by including partials per session clinical evaluations to ensure correct initial medical profiles and track eventual motor differences generated by the rehabilitation process.

Finally, also motion data (frame-by-frame tracking of joint positions using 3D Cartesian coordinates *x, y, z*) analysis represents a crucial next step to build on the correlations and performance-based findings already obtained. Integrating detailed kinematic features could uncover some "micro-level" movement patterns, such as compensatory strategies, tremor fluctuations or bradykinesia markers, that are not fully captured by score and level metrics alone.

In particular, analyzing raw motion trajectories may reveal movement interruptions and smoothness profiles, providing direct quantification of motor hesitations or freezing episodes. It would also allow for a detailed comparison of limb-specific motor capacity, highlighting asymmetries typical of Parkinson's Disease and enable precise estimation of maximum ranges of motion (ROM), which could be correlated with disease severity and progression. Postural stability could be assessed through tracking trunk and head coordinates, revealing compensatory balance strategies not visible in task scores.

Furthermore, motion data would support the computation of trial-to-trial consistency indicators, describing how stable or variable a patient's motor execution is across repeated attempts, which is particularly relevant for the Gym exergame. These new kinematic variables could be correlated with both clinical scores and self-reported outcomes, providing a richer basis for unsupervised clustering and refinement of prototype session phases.

Finally, as UPDRS tasks video recordings, motion tracking data can be reprocessed independently, eventually by third parties.

In summary, this thesis demonstrates that accessible webcam-based exergames represent a promising, scalable approach for remote motor-cognitive rehabilitation in Parkinson's Disease. By combining usability, patient engagement and meaningful correlation with clinical measures, they form a robust basis for developing adaptive, individualized care pathways that complement traditional treatments. Moreover, by reducing the need for travel and in-person sessions, such systems can ease the burden on clinics and improve access for patients with mobility limitations.
In conclusion, the combination of telemedicine with exergames and related data

collection also ensures individuals correctly follows their at-home rehabilitation therapy, crucial for a degenerative disease like Parkinson's Disease.

Bibliography

[1] Jennifer D. Sciubba. «Population Aging as a Global Issue». In: *Oxford Research Encyclopedia of International Studies* (2020). Accessed: 2025-06-24. DOI: 10.1093/acrefore/9780190846626.013.559. URL: https://oxfordre.com/internationalstudies/abstract/10.1093/acrefore/9780190846626.001.0001/acrefore-9780190846626-e-559 (cit. on p. 1).

[2] Pablo Alvarez. *Charted: The World's Aging Population from 1950 to 2100*. Accessed: 2025-06-24. 2023. URL: https://www.visualcapitalist.com/cp/charted-the-worlds-aging-population-1950-to-2100/ (cit. on p. 2).

[3] United Nations, Department of Economic and Social Affairs, Population Division. *World Population Prospects 2022: Summary of Results*. https://www.un.org/development/desa/pd/content/World-Population-Prospects-2022. Accessed: 2025-06-24. 2022 (cit. on p. 2).

[4] SynArea Consultants S.r.l. *ELEVATOR Project Overview*. https://www.synarea.com/project-elevator.php. Accessed: 2025-06-24. 2025 (cit. on p. 1).

[5] NODES – Nord Ovest Digitale e Sostenibile. *NODES: Nord Ovest Digitale e Sostenibile*. https://ecs-nodes.eu/. Accessed: 2025-06-24. 2025 (cit. on p. 1).

[6] SynArea Consultants S.r.l. *SynArea Consultants – ICT Solutions and Innovation*. https://www.synarea.com/index.php. Accessed: 2025-06-24. 2025 (cit. on p. 2).

[7] SynArea Consultants S.r.l. *REHOME Project Overview*. https://www.synarea.com/project-rehome.php. Accessed: 2025-06-24. 2025 (cit. on p. 2).

[8] Claudia Ferraris et al. «Usability of the REHOME Solution for the Telerehabilitation in Neurological Diseases: Preliminary Results on Motor and Cognitive Platforms». In: *Sensors* 22.23 (2022), p. 9467. DOI: 10.3390/s22239467. URL: https://www.mdpi.com/1424-8220/22/23/9467 (cit. on pp. 2, 8).

[9] Okun MS. Armstrong MJ. «Diagnosis and Treatment of Parkinson Disease: A Review.» In: *JAMA* (2020) (cit. on p. 3).

[10] Parnetti L. Paolini Paoletti F Tambasco N. «Levodopa treatment in Parkinson's disease: earlier or later?» In: *Ann Transl Med* (Sept. 2019) (cit. on p. 3).

[11] Carroll DG Halli-Tierney AD Luker J. «Parkinson Disease». In: *Am Fam Physician* (Dec. 2020) (cit. on p. 3).

[12] Parkinson's Disease Team. *Rating Scales and Staging*. `https://parkinsonsdisease.net/diagnosis/rating-scales-staging`. Accessed: 2025-06-12. 2025 (cit. on pp. 3, 4).

[13] Movement Disorder Society. *MDS-Unified Parkinson's Disease Rating Scale (MDS-UPDRS)*. `https://www.movementdisorders.org/MDS/MDS-Rating-Scales/MDS-Unified-Parkinsons-Disease-Rating-Scale-MDS-UPDRS.htm`. Accessed: 2025-06-12. 2025 (cit. on p. 4).

[14] A. Rana, A. Dumka, R. Singh, M.K. Panda, N. Priyadarshi, and B. Twala. «Imperative Role of Machine Learning Algorithm for Detection of Parkinson's Disease: Review, Challenges and Recommendations». In: *Diagnostics* 12 (2022), p. 2003. DOI: `10.3390/diagnostics12082003`. URL: `https://doi.org/10.3390/diagnostics12082003` (cit. on p. 6).

[15] G. Güney, T.S. Jansen, S. Dill, J.B. Schulz, M. Dafotakis, C. Hoog Antink, and A.K. Braczynski. «Video-Based Hand Movement Analysis of Parkinson Patients before and after Medication Using High-Frame-Rate Videos and MediaPipe». In: *Sensors* 22.20 (2022), p. 7992. DOI: `10.3390/s22207992`. URL: `https://doi.org/10.3390/s22207992` (cit. on p. 6).

[16] B.M. Bot et al. «The mPower study, Parkinson disease mobile data collected using ResearchKit». In: *Scientific Data* 3 (2016), p. 160011. DOI: `10.1038/sdata.2016.11`. URL: `https://doi.org/10.1038/sdata.2016.11` (cit. on p. 6).

[17] A. Tsanas, M.A. Little, C. Fox, and L.O. Ramig. «Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease». In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 22.1 (2014), pp. 181–190. URL: `https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-1597?show=full` (cit. on p. 6).

[18] H. Jaeger, D. Trivedi, and M. Stadtschnitzer. *Mobile Device Voice Recordings at King's College London (MDVR-KCL)*. Zenodo, 2019. DOI: `10.5281/zenodo.2867215`. URL: `https://doi.org/10.5281/zenodo.2867215` (cit. on p. 6).

[19] M.E. Isenkul, B.E. Sakar, and O. Kursun. «Improved spiral test using digitized graphics tablet for monitoring Parkinson's disease». In: *2nd International Conference on e-Health and Telemedicine (ICEHTM)*. 2014, pp. 171–175 (cit. on p. 6).

[20] S. H. J. Keus, B. R. Bloem, E. J. M. Hendriks, A. B. Bredero-Cohen, and M. Munneke. «Evidence-based analysis of physical therapy in Parkinson's disease with recommendations for practice and research». In: *Movement Disorders* 22.4 (2007), pp. 451–460 (cit. on p. 6).

[21] I. H. W. M. Sturkenboom, M. J. L. Graff, J. C. M. Hendriks, Y. Veenhuizen, M. Munneke, B. R. Bloem, M. M. A. Derix, and M. W. G. Nijhuis-van der Sanden. «Efficacy of occupational therapy for patients with Parkinson's disease: A randomised controlled trial». In: *The Lancet Neurology* 12.6 (2013), pp. 557–565 (cit. on p. 6).

[22] Lorraine O. Ramig, Shimon Sapir, Susan Countryman, Laura Thompson, Yoshiyuki Horii, and Cynthia Fox. «Intensive speech treatment for patients with Parkinson's disease: short- and long-term comparison of two techniques». In: *Neurology* 47.6 (2001), pp. 1496–1504 (cit. on p. 6).

[23] R. D. Dobkin et al. «Cognitive-behavioral therapy for depression in Parkinson's disease: A randomized, controlled trial». In: *American Journal of Psychiatry* 168.10 (2011), pp. 1066–1074 (cit. on p. 6).

[24] B. Pickut, S. Vanneste, M. A. Hirsch, W. Van Hecke, E. Kerckhofs, P. Mariën, P. M. Parizel, D. Crosiers, and P. Cras. «Mindfulness Training among Individuals with Parkinson's Disease: Neurobehavioral Effects». In: *Parkinson's Disease* 2015 (2015). Epub 2015 May 26, p. 816404. DOI: 10.1155/2015/816404 (cit. on p. 6).

[25] F. Li, P. Harmer, K. Fitzgerald, E. Eckstrom, R. Stock, J. Galver, G. Maddalozzo, and S. Batya. «Tai Chi and Postural Stability in Patients with Parkinson's Disease». In: *New England Journal of Medicine* 366.6 (2012), pp. 511–519. DOI: 10.1056/NEJMoa1107911 (cit. on p. 6).

[26] Jenny Clissold et al. «KICK OUT PD: Feasibility and quality of life in the pilot karate intervention to change kinematic outcomes in Parkinson's Disease». In: *Pilot and Feasibility Studies* 6.1 (2020), pp. 1–12. DOI: 10.1186/s40814-020-00644-3 (cit. on p. 6).

[27] C. Schlick, S. Pilge, C. Jurkat, M. A. Hobert, C. van Schie, K. Bötzel, R. Krüger, and G. Deuschl. «A pilot study of a 12-week community-based boxing program for Parkinson's disease». In: *Clinical Parkinsonism & Related Disorders* 7 (2022), p. 100080. DOI: 10.1016/j.prdoa.2022.100080 (cit. on p. 6).

[28] Madeleine E. Hackney and Gammon M. Earhart. «Community-based Argentine tango dance program is associated with increased activity participation among individuals with Parkinson's disease». In: *Archives of Physical Medicine and Rehabilitation* 94.2 (2013), pp. 240–249. DOI: 10.1016/j.apmr.2012.08.499 (cit. on p. 6).

[29] Jeanette Tamplin, Meg E. Morris, Caterina Marigliani, Felicity A. Baker, and Adam P. Vogel. «ParkinSong: A Controlled Trial of Singing-Based Therapy for Parkinson's Disease». In: *Neurorehabilitation and Neural Repair* 33.6 (2019), pp. 453–463. DOI: 10.1177/1545968319847948 (cit. on p. 6).

[30] A. E. Dias, J. C. Limongi, E. R. Barbosa, and W. T. Hsing. «Voice Telerehabilitation in Parkinson's Disease: Perceptual and Subjective Measures». In: *Journal of Voice* 30.6 (2016). Evaluated via Adobe Connect videoconferencing (cit. on p. 7).

[31] Kei Sekimoto, Shin Kawaguchi, Yasushi Yabe, et al. «A Randomized Crossover Pilot Study of Telemedicine Delivered via iPads in Parkinson's Disease». In: *Parkinson's Disease* (2019). DOI: 10.1155/2019/9403295 (cit. on p. 7).

[32] Carla Silva-Batista, Filipe O. de Almeida, Jennifer L. Wilhelm, Fay B. Horak, Martina Mancini, and Laurie A. King. «Telerehabilitation by Videoconferencing for Balance and Gait in People with Parkinson's Disease: A Scoping Review». In: *Geriatrics* 9.3 (2024), p. 66. DOI: 10.3390/geriatrics9030066 (cit. on p. 7).

[33] C. Pazzaglia, I. Imbimbo, E. Tranchita, C. Minganti, D. Ricciardi, R. L. Monaco, A. Parisi, and L. Padua. «Comparison of virtual reality rehabilitation and conventional rehabilitation in Parkinson's disease: A randomised controlled trial». In: *Physiotherapy* (2020) (cit. on p. 8).

[34] Pablo Cornejo Thumm, Nir Giladi, Jeffrey M. Hausdorff, and Anat Mirelman. «Tele-rehabilitation with virtual reality: A case report on the simultaneous, remote training of two patients with Parkinson disease». In: *American Journal of Physical Medicine and Rehabilitation* (2021) (cit. on p. 8).

[35] M. Gandolfi et al. «Virtual reality telerehabilitation for postural instability in Parkinson's disease: A multicenter, single-blind, randomized, controlled trial». In: *PubMed* (2017) (cit. on p. 8).

[36] Jos Hardeman, Daisy Geerse, Simone Hoogendoorn, Joyce Nonnekes, and Melvyn Roerdink. «Remotely Prescribed, Monitored, and Tailored Home-Based Gait-and-Balance Exergaming Using Augmented Reality Glasses: A Clinical Feasibility Study in People with Parkinson's Disease». In: *Journal of NeuroEngineering and Rehabilitation* 21.1 (2024). ClinicalTrials.gov: NCT05605249. DOI: 10.1186/s12984-024-1373740 (cit. on p. 8).

[37] Gianluca Amprimo, Giulia Masi, Lorenzo Priano, Corrado Azzaro, Federica Galli, Giuseppe Pettiti, Alessandro Mauro, and Claudia Ferraris. «Assessment Tasks and Virtual Exergames for Remote Monitoring of Parkinson's Disease: An Integrated Approach Based on Azure Kinect». In: *Sensors* 22.21 (2022), p. 8173. DOI: 10.3390/s22218173. URL: https://www.mdpi.com/1424-8220/22/21/8173 (cit. on p. 9).

[38] M. Gandolfi et al. «Immersive Virtual Reality Exergames for Gait and Balance Rehabilitation in Parkinson's Disease». In: *Frontiers in Neurology* 15 (2024), p. 1347755. DOI: 10.3389/fneur.2024.1347755. URL: https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2024.1347755/full (cit. on p. 9).

[39] Microsoft. *Azure Kinect DK*. https://azure.microsoft.com/it-it/products/kinect-dk. Accessed: 2025-06-20 (cit. on p. 9).

[40] Google AI. *MediaPipe Pose Landmarker*. https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker. Accessed: 2025-06-20 (cit. on pp. 9, 12).

[41] Ameur Latreche, Ridha Kelaiaia, Ahmed Chemori, and Adlen Kerboua. «Reliability and validity analysis of MediaPipe-based measurement system for some human rehabilitation motions». In: *Measurement* (2023), p. 112826. DOI: 10.1016/j.measurement.2023.112826. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04055464v1/file/Article_Latreche_Final.pdf (cit. on p. 9).

[42] Carlalberto Francia. «Validation of a MediaPipe system for markerless motion analysis during virtual reality rehabilitation». MA thesis. Politecnico di Milano, 2024. URL: https://www.politesi.polimi.it/retrieve/04940262-6340-43d5-a2ab-222e497d5336/2024_04_Francia_ExecutiveSummary_02.pdf (cit. on p. 9).

[43] Muhamad Mustar, Bintang Surya Tryatmojo, Maulana Hidayat, Rama Wiyagi, Maria Maria Ulfa, Meiky Fredianto, and Muhammad Ihsan. «Range of Motion Detection System in Humans Based on MediaPipe Special Flexion-Extension and Abduction-Adduction Movements». In: Aug. 2023, pp. 486–491. DOI: 10.1109/ICE3IS59323.2023.10335332 (cit. on p. 10).

[44] F. Roggio, S. Di Grande, S. Cavalieri, D. Falla, and G. Musumeci. «Biomechanical Posture Analysis in Healthy Adults with Machine Learning: Applicability and Reliability». In: *Sensors* 24 (2024), p. 2929. DOI: 10.3390/s24092929. URL: https://doi.org/10.3390/s24092929 (cit. on p. 10).

[45] J. Bertram, T. Krüger, H.M. Röhling, A. Jelusic, S. Mansow-Model, R. Schniepp, and et al. «Accuracy and repeatability of the Microsoft Azure Kinect for clinical measurement of motor function». In: *PLOS ONE* 18.1 (2023), e0279697. DOI: 10.1371/journal.pone.0279697. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0279697 (cit. on p. 11).

[46] C. Brambilla, R. Marani, L. Romeo, M.L. Nicora, F.A. Storm, G. Reni, M. Malosio, T. D'Orazio, and et al. «Azure Kinect Performance Evaluation for Human Motion and Upper Limb Biomechanical Analysis». In: *Heliyon* 9 (2023), e21606. DOI: 10.1016/j.heliyon.2023.e21606. URL: https://cris.unibo.it/handle/11585/946976 (cit. on p. 11).

[47] Qiurong Xie, Bo Sheng, Jia Huang, Qi Zhang, and Yanxin Zhang. «A Pilot Study of Compensatory Strategies for Reach-to-Grasp-Pen in Patients with Stroke». In: *Applied Bionics and Biomechanics* 2022 (Nov. 2022), pp. 1–13. DOI: 10.1155/2022/6933043 (cit. on p. 12).