Doctoral theses at NTNU, 2022:124

Elise Klæbo Vonstad

Improving Exergame Technologies for Older Adults Using Machine Learning

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Information Technology and
Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Elise Klæbo Vonstad

Improving Exergame Technologies for Older Adults Using Machine Learning

Thesis for the Degree of Philosophiae Doctor

Trondheim, May 2022

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering'
Department of Computer Science

ISBN 978-82-326-5908-1 (printed ver.)
ISBN 978-82-326-6623-2 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2022:124

Printed by NTNU Grafisk senter

Abstract

Facilitating exercise for the aging population is an important focus area in the years to come. Exercise is one of the keys to healthy aging, and the most effective measure for preventing disease and loss of independence. Technology is already an important tool in healthy aging, and in recent years exercise games (exergames) have been shown to be a motivating, fun and efficient method of exercising. However, the existing technologies that facilitate use of exergames have some drawbacks that could decrease usability and accessibility of exergames for older adults. Advances in artificial intelligence have provided tools and methods that might be useful for improving exergame technologies, but it is not known how well these work in the context of balance exergames. The overall aim of this thesis is to explore how use of machine learning can improve existing solutions of core elements of exergaming systems used for balance training in elderly.

Three research papers have been published as a result of the work in this thesis. These address three core aspects of exergame technologies: motion capture technology, movement pattern assessment, and force estimation. These papers together provide the following key findings:

- A deep learning image analysis system is a viable option for accurately extracting joint center locations from digital video for use in in-home exergame settings.
- A machine learning model can classify correctly performed medio-lateral weight-shifts in >9 out of 10 repetitions, without using pre-determined rules or thresholds.
- Weight-shifting performance can be reliably estimated from joint center kinematic data using a recurrent neural network.

In conclusion, we show that using machine learning models can make exergames more available and easy to use by eliminating possible barriers of use related to technological tools.

Preface

This thesis is submitted to the Norwegian University of Science and Technology (NTNU) in partial fulfillment of the requirements for the degree of Philosophiae Doctor. The Ph.D. work was performed at the Department of Computer Science, NTNU, Trondheim, under the supervision of Associate Professor Jan Harald Nilsen as primary supervisor, and Associate Professor Kerstin Bach and Professor Beatrix Vereijken as co-supervisors.

Acknowledgments

I extend my sincere gratitude to my fantastic supervisors. Jan, you have supported me, and pushed me, throughout these four years with great warmth and wisdom. You have listened to my more or less sane suggestions and gently guided me towards this end goal. Beatrix, your enthusiasm and positive attitude, combined with a wealth of knowledge is, and has always been, truly inspiring and motivating. And Kerstin, our great talks, your guidance, and patience, in my endeavor into the machine learning world have been invaluable for me throughout this journey of merging technology and health science.

I also give a huge thank you to Xiaomeng Su for your mentoring and companionship. I truly appreciate all our talks and discussions and look forward to many more in future collaborations.

During these four years, especially the 1.5 years in home office, Gunhild Lundberg, Beate Eltarvåg Gjesdal, Madeleine Lorås, and Emanuel Lorentz have been essential for me. Our daily talks and writing sessions have been keeping me going, and I am certain that I would not have finished any time soon without your support. Gunhild, I have learned so much from you, and you inspire me with your positivity, structure, and never-ending compassion. Beate, your humor, knowledge, and enthusiasm make me happy. Madeleine, your skills, confidence, and fearlessness is inspiring. And Emanuel, your wit and meticulousness are amazing, and I really enjoy our nerdy chats.

To the amazing AiT group at IDI: I have learned so much from you guys, and truly enjoyed being your colleague. Thank you for taking me in with open arms. And Monica Storvik, you are simply the best.

To my fellow PhDs at AiT, thank you for all our great laughs, cleansing rants, and long lunches at the office. I wish you all the best.

To the Sunnaas gang, and Arve Opheim and Linda Rennie in particular, your positive attitude and openness was the reason I discovered my passion for health tech-

nology, which has brought me to this point. You will always have a special place in my heart.

To mom, dad, Frida, and Sofie, your cheers, support, and love have meant the world to me during these four years. You have been my "Fyr" when I have been searching for direction, which I am forever grateful for.

I want to thank all my other friends for listening to me talk about this project, cheering me on, and asking when I'm done, for four years. I am now happy to announce that I am, in fact, done.

And, Malin and Even. This would not have been possible without you. Thank you for keeping me sane and grounded to earth, and for being my true source of motivation and perseverance.

Last, but not least, I want to thank myself for keeping going and for making it through all the ups and downs, all the way to the finish line.

to Even

Abbreviations

- **% BW** percent bodyweight
- **3DMoCap** 3-dimensional motion capture
- **ADL** Activities of daily living
- **AI** Artificial intelligence
- **AGI** Artificial general intelligence
- **CV** Cross-validation
- **ERC** European Research Council
- **CoM** Center of mass
- **CoP** Center of pressure
- **CoV** Coefficient of variance
- **CNN** Convolutional neural network
- **DL** Deep learning
- **DLC** Deep lab cut
- **DV** Digital video
- **GRF** Ground reaction force
- **IMU** Inertial measurement unit
- **kNN** k-Nearest Neighbor
- **LinReg** Linear Rrgression
- **LOGO** Leave one group out

- **LSTM** Long-short term memory
- **MAE** Mean average error
- **ML** Machine learning
- **MLP** Multi-layer perceptron
- **MPJPE** Mean per joint position error
- **MSE** Mean square error
- **PCA** Principal component analysis
- **PCK** Percentage correct keypoints
- **PiG-FB** Plug-in-gait full-body
- **ReLU** Rectified linear unit
- **ResNet** Residual neural network
- **RNN** Recurrent neural network
- **RMSE** Root mean square error
- **RFC** Random forest classifier
- **SD** Standard deviation
- **SVM** Support vector machine
- **ToF** Time-of-flight
- **UN** United Nations
- **WHO** World Health Organization
- **XGBoost** Extremely boosted gradient trees

Figures

Table of Contents

Part I

Research overview.

Chapter 1

Introduction

1.1 The Potential of Games to Facilitate In-Home Exercise for the Aging Population

1.1.1 Exercise and Aging

FOR many of us, aging in good health is a major focus to preserve quality of life in older age. We know that physical activity has a preventive effect on many conditions and diseases that require help and assistance from our health care systems [1]. The ability to live at home independently, manage activities of daily living, and participate in the local community depends on the prevention of such diseases and conditions. Even though information about the benefits of physical activity abounds, only a small proportion of us actually reach the recommended level of physical activity to achieve a preventive effect.

Among such preventive measures, those that target falls are essential. In elderly living at home, 40% experience a fall each year, and 1 out of 40 falls lead to hospitalization. Falls are the fifth most common cause of death in people above 65 years old [2]. Falls can also lead to increased fear of falling (again), thereby leading to reduced quality of life and reduced community participation, and increased mortality [3]. However, research has shown that physical activity is an effective preventive measure of falls [4].

Promoting and finding ways to deliver physical exercise to the aging population is one of the key areas of focus in our society in the coming years. We are in a demographic shift towards a higher proportion of elderly, which will require more people to *live at home for longer* to maintain a sustainable health care system. The question, then, becomes: **How can we as a society enable and facilitate physical**

activity for the aging population? One of the keys to a possible solution is tapping into how and why we choose to engage in physical activity - our motivation. One exciting avenue to explore in this regard is *gamification*.

1.1.2 Technology and Gamification of Exercise

Gamification is a term coined to describe the uptake of aspects from gaming into other areas of practice [5]. Here, elements that contribute to making games fun, engaging, and motivating are adapted and implemented into activities that previously were perceived as tiresome, boring, and repetitive. Storytelling, competition, and rewards are elements from games that have been adopted into serious settings [6, 7], and a digital platform provides an opportunity for creating customized virtual worlds. Gamification elements can be found in for example teaching, professional training, sales work, and also in apps and solutions meant for physical exercise. Ranking and competition is not a new concept in exercise, but implementing the dimension of rewards using logged or tracked data from the performance during the exercise itself has been shown to improve exercise adherence and motivation [8]. This is also reflected in the number of exercise apps that implement gamification, and in the creativity shown in the manner of implementing these elements. You can for example choose to run from zombies (Zombies,Run!)[1], walk round and find, catch and compete with monsters (Pokémon GO!)[2], explore the universe by walking (Walkr: Fitness Space Adventure)[3], build a superhero and protect Earth through strength-training (Superhero Workout)[4], and much, much more.

This crossover from casual gaming to more serious activities was, and still is, driven by the advancement of technological solutions in the past two decades. Both hardware and software developments have provided high levels of immersion and interaction between a person and a computer, paving the way for improved workflows, control, and feedback. Miniaturization makes devices small and low cost, improving accessibility for a wider audience. We can use virtual reality to immerse ourselves in a three-dimensional world, we can talk to our computers and receive a more or less coherent response, and we can control devices using hand gestures or other bodily movements. Motion-based control is what has been driving the development of games that can be used for exercise purposes - commonly referred to as exergames [9].

[1] Six to Start, https://zombiesrungame.com/

[2] Niantic, https://pokemongolive.com/en/

[3] Fourdesire, http://walkrgame.com/en/

[4] Six to Start, https://www.sixtostart.com/superhero-workout

Figure 1.1: The Nintendo Wii controller and console (left), and Xbox 360 console and Kinect camera (right).

1.1.3 Exergames

The use of exergames in serious health settings can be largely attributed to two devices that were originally designed for casual gaming: The Nintendo Wii (Nintendo Co., Ltd., Japan) Motion Controller, and the Kinect (Microsoft Corporation, USA) camera system for Xbox and Windows (Figure 1.1). These devices in combination with sports game collections such as "Wii Sports" (Nintendo Co., Ltd.) and "Kinect Adventures" (Microsoft Game Studios), created a consumer market for motion-controlled casual games, as the technology enabling this type of interaction was emerging. Here, a variety of sports could be played either solo or in multiplayer mode, such as tennis, bowling, or imaginary games performed for example under water or while flying. Casual games often have in common tasks that require a player to move in a specific manner to avoid in-game obstacles, hit targets, lean sideways, or step in specific directions. This was quickly recognized as being a possible facilitator for making people perform exercises in more serious settings, such as in fitness training or rehabilitation after injury. The motivational aspect is obvious, playing a game by moving in a specific manner to regain, maintain or improve physical function has huge potential, which is reflected in a rapidly expanding body of research on exergames and motivation [10, 11, 12].

For the aging population, gaming through digital devices is not yet seen as a typical or common activity. Nevertheless, both health care workers and elderly living in assisted care facilities have picked up on the potential for exergames as a fun and at the same time possibly fruitful tool. Research has documented that this demographic enjoys playing exergames as well [12], both for general activity and in more targeted exercise situations such as rehabilitation after injury [13, 14]. The exercise efficacy of exergames has been documented in many geriatric patient populations as well, such as stroke survivors [15], frailty [16], and patients with neurological disease [17]. In healthy older adults, exergames have been shown to be as effective or even more effective than the traditional exercise programs that would have been prescribed for balance training [18, 19] and falls prevention [20].

Additionally, using digital devices to facilitate exercise allows for individually tailored exercise programs. Tasks and objectives in the game can be created to elicit specific movement patterns, and the difficulty level can be adjusted to the person playing to facilitate mastering the task. This makes it possible to target specific physical functions, and the game can be set to elicit repeated performances of that specific movement in a fun and motivating manner, thus providing an opportunity for increased time spent exercising that specific function [7]. This is highly relevant for exercising physical function, as repetition and specificity are crucial for progress and efficiency in such training [21, 22]. The game can also be changed visually or aurally to reflect the preferred scenes and interests of the player. For example, you can receive in-game rewards by reaching for apples in a French orchard, or by squatting and leaning sideways to ski down your favorite slope in the Alps.

The combination of effective movement patterns for targeted, guided exercise and a highly motivational component addressing the lack of adherence typically seen in traditional exercise makes exergames an especially promising avenue to pursue. This has major potential for becoming a highly useful tool, especially as you can receive real-time feedback on the performance of movement patterns during training. This feedback can be delivered both by being rewarded for good performances and by being guided on how to improve improper performance using visual and auditory information.

1.1.4 Filling the Gaps with Machine Learning

Despite this potential of using exergames for targeted exercise, there are some aspects of exergame systems that need to be addressed before they can be employed in widespread use. Current technological solutions are based on casual gaming situations for entertainment, where the accuracy of movement tracking and feedback are not essential. This resulted in technology that might not represent or analyze the players' movement patterns in a manner adequate for a serious setting such as exergaming for physical training. Studies have shown that in a clinical setting, trust and perceived usefulness are essential for successful implementation and actual usage of technology [23, 24]. This means that the shortcomings of existing systems for casual gaming might be detrimental to trust and perceived usefulness, for example, if the users experience a poor match between their real-world movements and the movements captured by the system. Based on these considerations we identified three major features of exergame technology that could potentially be improved using modern data-analysis methods within machine learning.

Firstly, the quality of the motion capture is essential for an exergame system to function properly, as this is the basis for interaction with the exergame. The motion capture system has to accurately measure player movement to be able to control the game, assess the movement patterns, and subsequently provide rewards, guidance, and feedback. Two factors make the typically used depth-sensing cameras such as the Kinect sub-optimal for in-home exergaming. This is a separate, specialized device that needs to be acquired, set up properly, and maintained, and the quality of the data provided by these cameras is varying, resulting in the tracking of the player being unsatisfactory for serious settings [25].

Secondly, the assessment of movement pattern quality is typically based on rules and thresholds. Here, a player receives points based on a comparison between the performed movement pattern and a pre-determined template, or set of rules. However, if the player has a body shape that does not fit this template or rule set, or is unable to perform the movement according to the template or rule set because of physical limitations, they will not receive an in-game reward even though they have performed the movement correctly based on their individual prerequisite. Because the older adult population is a heterogeneous group, with widely varying levels of physical function, exergames that assess movement patterns according to templates or rules might exclude some people from using the systems in a useful manner.

Thirdly, there are several shortcomings in existing systems that specifically target balance training for older adults. Balance training is a vital part of physical exercise for fall prevention [26], and being able to perform guided balance training using exergames in an in-home setting would be a major advantage for both users and health care providers. The issue is that accurately assessing performance, and providing feedback, during weight-shifting or leaning exercises requires equipment that measures how much force you exert on the ground. Devices that do this with sufficient accuracy are typically found in specialized laboratories, as they are very costly and not feasible to use in an in-home setting. An exergame system capable of providing force information in a simple, accessible manner would facilitate guided exergaming that enables balance training for fall prevention.

Exploring advancements in machine learning (ML) that have occurred in the past two decades is an avenue that holds great potential for finding solutions addressing the above-identified gaps, thereby more useful and available exergame systems for balance exercise. Image analysis methods, classification algorithms, and estimation models are all ML tools that have been used previously in similar tasks as the ones mentioned above. However, the knowledge of how these ML tools would perform in motion tracking, assessment, and enabling exergaming is still missing, particularly with the ease of use and accuracy in motion tracking and feedback as the overall goal of using such methods for analyzing exergaming data.

1.2 Aim of the thesis

The overall aim of the thesis is to explore how the use of machine learning and deep learning could improve existing solutions of core elements of exergaming systems used for balance training in the elderly. Specifically, the thesis aimed at investigating the use of image analysis methods for more accessible motion capture, classification models for more accurate movement pattern evaluation, and recurrent neural networks for enabling weight shifting analysis using only kinematic data.

1.2.1 Research questions

The main research questions addressed in the papers are the following:

- **RQ1** How does a state-of-the-art deep learning 2D image analysis system perform with regard to segment length variability compared to a 3D motion capture (3DMoCap) system and Kinect system?
- **RQ2** To what extent can machine learning classification models identify correctly performed movement patterns during balance exergaming?
- **RQ3** What accuracy does a recurrent neural network model achieve on estimation of force data, using kinematic data from 2D and 3D motion capture systems?

1.2.2 Thesis overview

This thesis is based on the work conducted during four years of study, from 2017 to 2021. Three studies were undertaken, with the main data collection supplying data for all three studies. From each study, original research papers were produced and published as detailed below. The papers can be found in full in part II.

Paper I - *Comparison of a Deep Learning-Based Pose Estimation System to Marker-Based and Kinect Systems in Exergaming for Balance Training* (referred to as "Pose Estimation" hereafter) details the study conducted to answer RQ1. The paper from this study was submitted in October 2020 and published in Sensors in December 2020.

Paper II - *Assessment of Machine Learning Models for Classification of Movement Patterns during a Weight-Shifting Exergame* (referred to as "Movement Classification" hereafter) details the study conducted to investigate RQ2. The paper from this study was submitted in May 2020 and published in IEEE Transactions of Human-Machine Systems in the spring of 2021.

Paper III - *Estimation of Ground Reaction Force from Kinematic Data in Balance Exergaming* (referred to as "Force Estimation" hereafter) details how we conducted the research to answer RQ3. The paper written on this study was submitted to the Journal of NeuroEngineering and Rehabilitation in June 2021 and is currently under review.

The studies are ordered in this manner due to the themes in each study: pose estimation comes first, as motion capture is the basis for using exergames; assessment of the captured movement comes next; followed by the use of kinematic data to enable force feedback during exergaming.

The remainder of this thesis is structured as follows. Chapter 2 outlines the background and context in which this thesis is situated. Chapter 3 details the technical frameworks the thesis is based on, along with the related work, positioning the thesis in the current literature. Chapter 4 gives an overview of the methods employed in the thesis, including ethical considerations, participant recruitment, practical details of data collection, and the analysis and evaluation methods used. In chapter 5, an overview of the results from the three studies is presented. These are then discussed in chapter 6, along with contributions and implications. The conclusion of the thesis can be found in chapter 7. Part II of the thesis contains the full papers this thesis is based on.

Chapter 2

Background

This chapter provides an overview of the setting in which this thesis was conducted. It provides a general description of the status and challenges within the relevant fields, meant to inform the reader about the rationale for the research questions posed in chapter 1.

2.1 Elderly, Balance and Exercise

In daily life, being able to successfully navigate and overcome challenges of balance (or postural control) is a key factor in health and well-being. We climb stairs, avoid obstacles, lean over to one foot to reach something, turn, and walk - among other activities - all common movements that are part of our activities of daily living (ADL). As we age, our capacity to complete these movements gradually deteriorates, as our physical and mental functions are affected by the processes of aging [27]. The effect aging has on our ability to maintain postural control has been well documented. Our reaction time gets slower [28], muscles become weaker [29], and the processing demands for actions increase [30], all contributing to a reduced capacity to successfully navigate the world around us. One of the most serious consequences of deteriorated balance is the increased risk of experiencing a fall. Even though the immediate effects of a fall such as physical injury can be severe for an elderly person, the long-term effects can be even worse. Quality of life can deteriorate because of reduced independence, community participation, and inability to perform activities of daily living safely [31]. The mechanisms causing falls are complex and highly variable, but one commonality in people experiencing fear of falling is decreased physical capacity [32]. Exercises aimed at improving muscle strength and balance function are one of the best prevention tools against falls [4].

While becoming aware of these factors, we find ourselves in a demographic transition towards a higher proportion of elderly in our societies, and at the same time life expectancy keeps increasing [33]. This means that not only will there be a larger population of elderly in our society, but also that these elderly will live longer. Elderly people have a higher rate of diseases and conditions that require assistance from health care personnel. To have a sustainable health care system in the coming years, it is therefore essential that elderly persons stay healthy and independent, and are able to live at home for longer than they do now.

To achieve this, maintaining physical and mental function - and thus quality of life - will be an extremely important effort to undertake. As research has shown, the best way to achieve this is through a healthy and active lifestyle [31]. Exercise aimed at balance or postural control is one of the cornerstones for achieving this, along with other types of physical activity [26, 34]. This is because balance training maintains and improves our ability to avoid a potentially hazardous fall by responding quickly enough to destabilizing conditions and thus staying in control of our balance [35, 36], and keeping a steady and stable gait pattern [37]. Because of the importance of balance exercise for maintaining balance function and preventing falls, weight-shifting exercises are recommended in exergames for older adults.

As exercise can prevent or mediate the decline in physical and mental functions, focus on exercise for elderly persons has increased in recent years. Healthy and active aging is a major topic in both national and international forums, such as the World Health Organization (WHO[1]). This is also reflected in the vast amount of resources being funneled towards finding methods, models, and pathways of delivering efficient, safe, and motivating exercises to this population, as seen by the European Research Council (ERC[2]). An important outcome from these studies has been that delivering exercise through technological solutions is one of the most promising avenues to pursue (e.g. [38, 39, 40]).

2.2 The Use of Technology in Exercise

Elderly Also Cheat in Videogames

Technological advances in the past two decades have driven the development of health care to what can be defined as a new era [41]. Specifically, regarding exercise for physical and mental functions, technology is now commonly used to track, evaluate and suggest exercises in all age groups, and from (re)habilitation to

[1] https://www.euro.who.int/en/health-topics/Life-stages/healthy-ageing

[2] https://ec.europa.eu/info/research-and-innovation/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe/cluster-1-health$_en$

world-class athletics. This is facilitated in particular by technology that allows for devices that have long battery life, are small and light in size and that can provide *accurate enough* tracking of human activity. This is combined with advances in software development that provide possibilities for individually tailored feedback and guidance.

In its most basic definition, physical exercise is the movement of our limbs performed by a complex interplay between our muscular, skeletal, and neurological systems. The purpose of physical exercise is to improve our strength, balance, endurance, coordination, and/or motor control, and the movement patterns we perform during exercise are specifically aimed at training for example leg muscle strength or endurance capacity. The movement patterns are designed to make these exercises as efficient as possible, by tailoring the range of motion over the joint, movement speed, and movement direction. Although there is some leniency, the most efficient exercise is where the movement patterns are performed as close to the intended movement as possible. This is also why guided exercise is more efficient than non-guided exercise [42]. Getting feedback on your movement patterns makes it easier to perform them correctly, as you can correct mistakes and are more incentivized to keep performance at a high level.

Being able to provide access to feedback and guidance to elderly exercising at home is a core issue, because of the major positive effect independent at-home exercise could have in mediating the coming strain on our healthcare system. If you do not have to see a clinician or a physical therapist for exercise, but instead can perform your exercises at home with a therapist-approved, safe and motivating system, each therapist can help a larger number of patients without sacrificing the quality of care. Furthermore, elderly persons can then exercise independently, while knowing that they are efficiently performing exercises. They could also, for example, choose from a set of pre-determined exercises, which can encourage people to take more responsibility for their exercise. This is another interesting aspect, as we know that agency, control, and participation in one's progress also improve adherence [43]. As we perceive the world around us mostly using vision, having a visual representation that provides feedback on performance while exercising can be very useful in making sure that we are performing the intended movement patterns correctly. Using visual biofeedback during exercise leads to more effective exercise than not using it, as shown in for example [44]. In addition, the rise of gamification has inspired the further development of screen-based visual feedback systems. Here, elements from games and gaming that could be useful for motivation and feedback have been identified and utilized, which has driven the development of screen-based exercise games.

Screen-based exergames typically present the user with an avatar that represents the position and movements of the user's body in the virtual world [9]. The player then has to move their body to make the avatar perform tasks such as hitting targets or avoiding obstacles in the game, and receive rewards in the game based on the completion of these tasks. This means that the accuracy with which a users body and movements are represented in the virtual world is vital for gaining rewards - and by extension, for the user's motivation. Two facets are essential for exergames to be useful and motivating. One is the technical aspect, i.e., the ability of the system to successfully track and represent the player. The other is the assessment of movement pattern performance in relation to the intended task. *These two are closely related, as the assessment of performance is based on the body representation from the motion tracking system.* These two facets can produce two separate ways in which the player is not appropriately rewarded. If the motion tracking system is inaccurate, the actual performance of the user might not be captured by the system. This can lead to the system believing the user for example extended their arm properly, while they did not in real life, resulting in an undeserved reward. Or they did in fact extend their arm properly, but the system did not capture the movement properly and failed to deliver a deserved reward. Similarly, if the assessment of the movement pattern is inaccurate, the same two scenarios might happen: the user properly extends his arm, relative to his body and ability, but the system classifies it as wrong because of an internal rule that was not able to adapt to the individual player.

Such discrepancies between real-world movement patterns and what the system captures and assesses have resulted in situations where users, deliberate or not, find the easiest or least amount of movement necessary to receive a reward. In casual gaming settings, this is exemplified in game systems like the Nintendo Wii Sports (Nintendo, Japan), where a mere fine-tuned flick of the wrist while sitting on the couch can give you a bowling strike. In more serious settings, players have been shown to do this as well, by performing what is described as a "low effort" movement pattern [45, 46]. The reward is provided undeservedly, making the exercise effect much less than it could have been.

To call this behavior "cheating" might be too strong a term if it is done unwittingly, but it does present a shortcut that provides rewards without the player having to do the work that the game was initially designed for. The result is a situation where the technical side of an exergame system is not robust enough to capture and assess the actual movement pattern that was performed in real life. Whether or not the user realizes this is happening, it is likely detrimental to progress or maintenance of physical function, as they are not made aware that they are not performing the exercises correctly. In addition, it can negatively affect motivation if the exergaming system is perceived as unreliable and not trustworthy in the rewards that are provided.

Another facet of this is the issues related to using consumer-grade products intended for entertainment or casual gaming for serious settings, like balance exercise to improve or regain function. For example, the Wii Balance Board (Nintendo, Inc) is a force sensing device that reports center of pressure (CoP) while a person is standing on it, and the promise of being able to report force data in a resource-friendly manner has piqued the curiosity of using this for serious settings, as shown in systems being developed for this purpose (e.g., [47, 48, 49, 50]). However, it has also been demonstrated that the Wii Balance Board is not able to represent the CoP to an accurate enough degree to be used in such serious settings [51, 52, 53]. Its usefulness has also been shown to be limited due to experiences of technical difficulties in setting up the system, which was happening for both health care personnel and patients [54].

Thus, the key challenge in the development of balance exergame systems is to develop accessible technological solutions for motion tracking and feedback, as proper reward systems are based on this information. This challenge is arguably the very backbone for the usefulness and motivational aspects of exergames.

2.3 Artificial Intelligence and Machine Learning

Artificial Intelligence (AI) is a term coined to describe the efforts of making computers *think*. Since computers were invented, we have wondered if this was possible - can computers be programmed to *reason* the way humans do [55]? The word *"artificial"* in this context refers to the fact that computers are human-made tools, and *"intelligence"* is the ability to discern patterns and make decisions based on available information - i.e., "thinking". The goal is to mimic the way humans understand and reason by elucidating intricate relationships in data by remembering past experiences, and learning from them [56]. The increase in attention and research on AI has been driven by the increased possibility to collect and store so-called big data due to technology being increasingly used in all areas of society. This is also why applications using AI in some cases are referred to as *data-driven* or *big data* applications [57]. There are several different avenues within artificial intelligence, some of which are out of the scope of this thesis. There are branches specifically focused on, for example, speech, or text, recognition, image analysis, or robotics. The current thesis, however, is concerned with using machine learning models, a versatile branch of AI, that is being used in a wide variety of areas and tasks.

2.3.1 Machine learning

Machine learning (ML) is an area where the focus is creating and developing models that are tailored to specific tasks and specific areas of research. Here, the goal is more geared towards improving with experience, as defined my Mitchell [55](pp.2): "A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." There are two major sub-fields within machine learning: supervised and unsupervised learning. In supervised learning, the models are trained on data sets with the correct answers, or targets, available (while keeping a small part of the data hidden for validation purposes). In unsupervised learning, there are no target labels, so the models try to find patterns or groups in the input data. In the current thesis, the task is to model relationships between the data and known target variables, which places this work in the supervised learning realm.

Because of the huge potential for improving workflows and supporting decision-making processes, ML is currently used, or in the process of being used, in many areas of business and research. This has led to a proliferation of different model types and configurations, each tailored to a specific task at hand. Despite this, the same main approaches in supervised learning provide the foundation for all these models. The main tasks are regression and classification, which are commonly used for forecasting, estimation, and prediction. There are several models within each approach, but each machine learning model is an approximation of the true function f, the hypothesis that best represents the relationship in the data. It is an *approximation* because we do not have enough training data available to model the true relationship [55]. This makes it essential to evaluate the performance of an ML model on unseen data, a test of how well the model can generalize to other data, i.e., how close the approximation is to the true function.

Thus, training a machine learning model is the task of finding the hypothesis in the *hypothesis space* that yields the highest accuracy on the test set. To do this, we need to provide the model with training data that represents, as best as possible, the true relationship between the *features* and the *target* data. The features are the input data provided to the model, for example, the color values of each pixel in an image or stock information, and the targets are the true identification or class related to the input data, for example, "car" or "dog", or actual closing stock price [58]. Since the input data is the only information given to the model, the quality of the input data is essential for the model to be able to approximate the true relationship in the data. The term "GIGO" captures this well, which is an acronym for "Garbage in, garbage out", referring to the generalization inability of a model trained on a data set of poor quality. Hence, feature selection (finding a subset of the original feature space that represents well the relationships in the data) and feature engineering (creating new features based on the original features, e.g. taking the sum, or mean, of two features) is crucial for model performance.

Classification algorithms are designed to be able to use a set of training data to decide what class, or label, a new, unseen input belongs to. Even though neural networks (described below) often outperform other classification models, the latter is often preferred in many settings because of the lower demand for training data volume, lower demand for computational resources, and transparency in the decision-making processes. In many situations, these properties are more important than achieving slightly higher accuracy in a given task.

2.3.2 Deep learning

One subgroup of machine learning is deep learning, and the current ML hype is the neural network. Neural networks are designed to mirror the way a human brain works, with interconnecting neurons in (sequential) layers that process the information from a data set [59](illustrated in Figure 2.1). This is the area of machine learning that is commonly referred to as *deep learning*, because of the architecture that allows for hundreds or thousands of neurons in each inter-connected layer. The last layer of the network is called the *output* layer, where the actual prediction is made. The layers between the input and the output layer are called the *hidden layers*.

Neural networks can be used for both classification and regression, and are powerful in many applications. The drawback of neural network-based models is a high computational demand in training, combined with requiring large data sets for learning the relationship between the input data and the target data.

The application of deep learning in highly valuable and diverse domains in our society demonstrates the agility and power of deep learning methods. Research and development of a wide variety of architectures and configurations of neural networks, each designed for a specific domain or task, make it possible to create deep learning models that achieve human-level accuracy in any field. Even though there are drawbacks to neural networks, deep learning makes it possible to improve efficiency, reliability, and resource demand in data analysis and decision-making processes. This is achieved by automating systems that previously required a human to perform them, or involving traditional numerical analysis or statistical methods. The unique ability of neural networks to learn complex, multi-faceted relationships and patterns, and successfully apply this to large amounts of data (in real-time), has been used to develop new and better applications in for example automated driving [60], electrical utility prediction [61], manufacturing [62], and fraud detection [63]. Specifically within health care, deep learning is being used

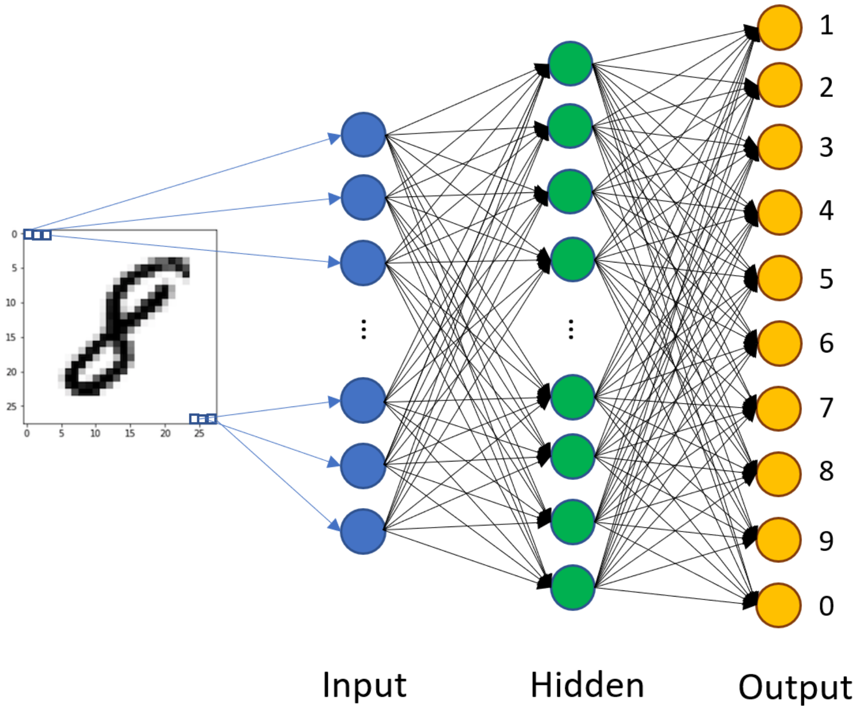


Figure 2.1: A simple neural network where the input neurons are the pixel values of a grayscale image that makes up an image of a number. There is one hidden layer, and the output layer consists of the numbers 0-9.

for decision-making support through, for example, medical imaging [64], prediction of health risk [65], and large-scale analysis of electronic health record data [66].

Within human movement science, deep learning applications are being used and developed in three major areas: 1) simplify and improve data capture methods, herein making data capture available outside of the laboratory, e.g. [67, 68, 69], 2) predicting, or estimating, for example gait events [70], joint kinematics [71], and joint loading [72], and 3) activity classification, e.g. [73], and movement assessment [74, 75]. These examples show that the aforementioned issues with existing exergame systems can be addressed by utilizing different aspects of deep learning, as the algorithms can be tailored to pose estimation, movement quality assessment, and weight-shifting estimation during exergaming.

Chapter 3

Scientific Framework

This section details the technical frameworks that the thesis is based upon, as well as extant work related to each of the three research questions. The section is organized to reflect the workflow in a balance exergame system. First, the pose estimation needs to be performed, which in this thesis is done by employing a residual neural network for 2D image analysis. Then, a complete movement is identified, which then can be assessed for correctness of the movement pattern, using traditional classification methods here. Or, the weight-shifting data can be extracted by using an LSTM recurrent neural network. This is illustrated in Figure 3.1

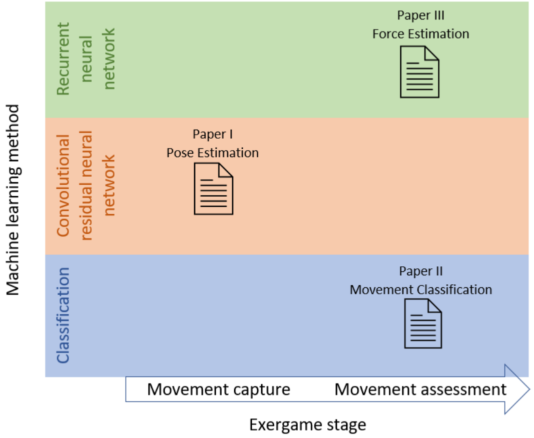


Figure 3.1: Illustration of the exergames stage addressed and machine learning method used in each of the three papers.

3.1 Paper I - Pose Estimation

As the inceptive task of an exergame system, pose estimation is essential for all subsequent data analysis steps in the exergame system. The representation of the bodily positions of the player must therefore be as accurate as possible. Deep convolutional neural networks have been shown to perform very well in image analysis settings where pose estimation is the end goal, and several different methods have been developed for various pose estimation problems. We therefore hypothesized that deep convolutional neural networks with a residual component might perform well in the current context of exergaming for balance training.

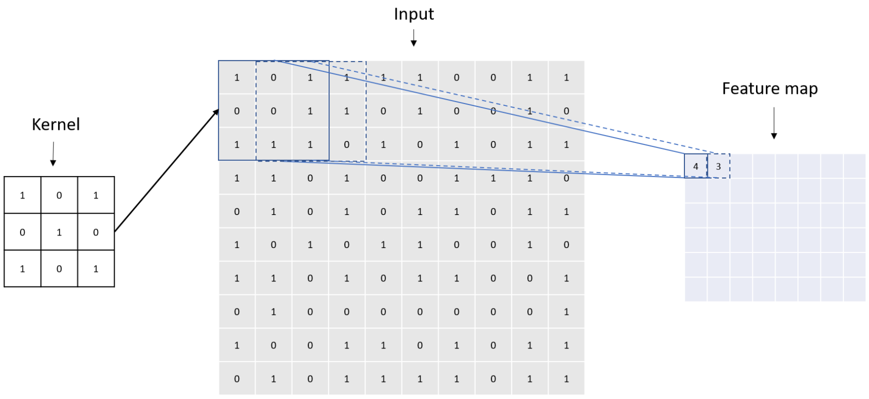
As indicated above, deep learning refers to architectures using multiple layers of neurons between the input and the output layers, with each layer containing multiple neurons. A deep neural network could for example be constructed as seen in Figure 2.1, but with more hidden layers. If all neurons in one layer are connected to all neurons in the next, the layer is called a *dense* or *fully connected* layer. The neurons in each layer compute the weighted sum of the input from all the connected neurons in the previous layer, and this information is propagated to all connected neurons in the next layer. This information that a neuron contains is a number, called a *weight*. The weights of all the neurons are randomly initiated with small values, which are then applied to the input values from the previous layer.

To adjust the activity, or the relevance, of a weight, a *bias* is often applied to each layer. In this context, bias is a value added to the weighted sum of the values in the input, adjusting the information propagated to the next layer, resulting in the output function. In the output layer, the neuron that is activated (i.e., the prediction that is made) depends on the weights that are received from the last hidden layer. In supervised learning, the output is then compared to the actual value that the specific input should have provided. If the output is wrong, the error is propagated backward down the network again to adjust the weights and biases of the neurons. This iterative process is how neural network models derive the optimal weights and biases of each neuron and is called *backpropagation* [59].

Activation functions are a vital part of neural networks. There are three main types of activation functions: the hyperbolic tangent (tanh), rectified linear unit (ReLU) and logistic sigmoid [76], where the data is passed through a "squeezing" based on the value of the data and the function applied, before being sent to the next layer. In the sigmoid activation function, for example, the more negative values result in weights closer to 0, while the more positive values result in weights closer to 1; in tanh, more negative values result in values closer to -1, while more positive result in values closer to 1. These functions can be illustrated as *squeezing* the data, as the input value of the data is assessed, and the output from the gate is constrained between -1,0, and 1 based on the function applied. ReLU has been the most popular variant in recent years, which is a function that sets all values <0 to 0, and a linear function to the input data >0. This has been shown to improve network performance without sacrificing accuracy [77].

Analyzing images - and videos - has been a major focus area within the deep learning research community. A visual representation of an object of interest can contain highly valuable information that other types of technological tools cannot represent in an equally compact and efficient manner. Identification of an object or context, or detecting whether these are present in an image, is the most common tasks for deep learning models. This is reflected in that the major AI challenges and competitions are tasks where the goal is to reach the highest possible accuracy on image recognition data sets such as MPII [78], ImageNet [79], and MS Coco [80].

Pose estimation can be seen as a sub-field of object identification in images. The task in pose estimation is to identify where in an image the different body parts of a person are located. To achieve this, convolutional neural networks (CNNs) have traditionally been the most popular alternative, because of their historically excellent performance in a wide variety of image recognition areas. The core of CNNs is the many iterations of *convolutions* over the data. A convolution operation consists of three parts: the input (image pixel coordinates (X, Y) and pixel value Z, where the Z-dimension can be more than 1 in cases where there are for example three or more color channels), a *kernel*, which is the function applied to the data at each iteration, and the output, often referred to as the *feature map* [59]. As shown in an example in Figure 3.2, an input can be a monochrome image of size 10×10×1. The kernel, in this example of size 3×3×1, is applied over an area of the input of the same dimensions, and the sum of the kernel operation is extracted to the resulting 8×8×1 feature map. The kernel can be of any size, and the stride size (i.e., how many pixels the kernel moves in each iteration) can also be adjusted. The next step is a pooling operation performed on the feature map. Here, the size of the feature map is further reduced, for example by taking the average or maximum value of a 2×2×1 window (Figure 3.3). This results in a higher-level representation of the image, where for example the edges of the object in the image are highlighted while more detailed information is suppressed, while also reducing the computational demands of further processing of the image [76].



1	0	1
0	1	0
1	0	1

1	0	1	1	1	1	0	0	1	1
0	0	1	1	0	1	0	0	1	0
1	1	1	0	1	0	1	0	1	1
1	1	0	1	0	0	1	1	1	0
0	1	0	1	0	1	1	0	1	1
1	0	1	0	1	1	0	0	1	0
1	1	0	1	0	1	1	0	0	1
0	1	0	0	0	0	0	0	0	1
1	0	0	1	1	0	1	0	1	1
0	1	0	1	1	1	1	0	1	1

Figure 3.2: Illustration of the convolution operation.

A convolutional *network* consists of several of these elements in sequence, resembling the sequential processing of information in deep neural found in deep neural networks, with some well-known examples in AlexNet [81], GoogLeNet [82], and ResNet [83]. These perform very well on large-scale public data sets for image recognition tasks such as ImageNet [84]. The latter model, ResNet, is a version of a CNN that includes a residual component. This means that the weights and biases from one layer skip a connection and are therefore propagated deeper into the network than to the subsequent layer, as depicted in Figure 3.4. This might reduce the issue of vanishing/exploding gradients that typically occurs in deep networks. Here, the gradient used to update the weights in the network exponentially increases or decreases at each derivative, resulting in weight updates that are extremely large or extremely small [85]. As a result, the network is unable to learn properly. By skipping connections with one or more layers, the gradient does not explode or vanish to the same degree as non-residual networks. The ResNet architecture thus allows for deeper networks, which can achieve better performance

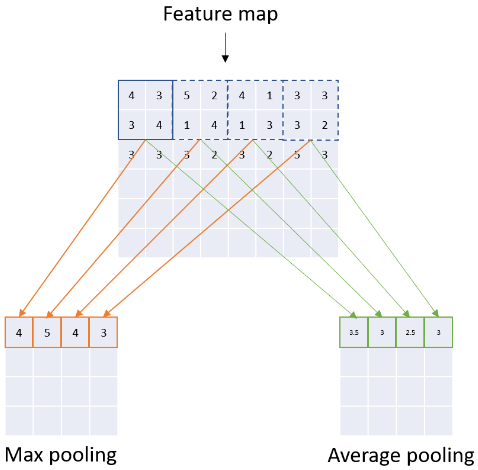


Figure 3.3: Illustration of a max and average pooling layer.

than networks without the residual component [83].

The ResNet architecture is the basis for a pose estimation algorithm called DeeperCut that produces state-of-the-art accuracy on body part detection tasks [86]. This inspired the development of another algorithm, DeepLabCut [87](DLC), which implements a ResNet architecture pre-trained on the MPII dataset [78] and applies an additional layer of training in their framework, namely a specialization layer that is trained using annotated images specific for the context in which the user is employing it. By doing this, the size of the training data set required for accurate identification of body parts in an image of a specific animal is very small, around 200 images [87], by providing a network that is already trained to detect body parts from the MPII data set. DeepLabCut is available as an open source tool (github.com/DeepLabCut/DeepLabCut), which makes pose estimation available for a wide range of users.

Different adaptation methods have been developed to improve performance or reduce computational demand in neural networks. These are commonly referred to as *regularization* and *optimization* and are regarded as the most important areas of research in the machine learning community [59]. Optimization is the process of minimizing the *loss* or *cost* function, which is the error the network makes in its predictions [59]. This process is called *gradient descent*, which refers to making adjustments to the model iteratively in the direction that reduces the cost function. Regularization is any adjustment of a model aimed at improving the generalization accuracy of the model, without necessarily improving training error. This is a

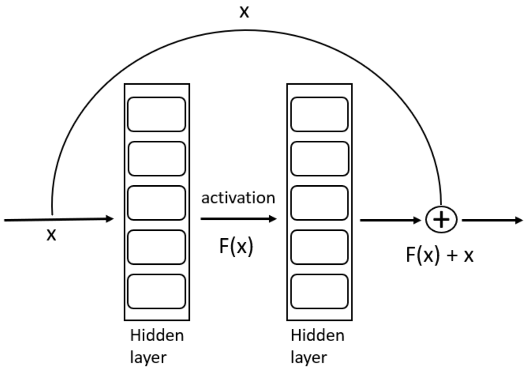


Figure 3.4: A residual block showing how the information flow in a neural network uses "shortcut connections" to skip layers in and propagate weights deeper into the network. F(x): activation function. Adapted from He et al 2016.

crucial adjustment, as it increases the accuracy when making predictions based on previously unseen data, and thus making the model more suited for generalizing to a real-world situation where the trained model has to make predictions exclusively on unseen data.

3.1.1 Previous Work related to Human Pose Estimation from Video

To develop human pose estimation algorithms, large-scale public data sets have been created and curated for benchmarking purposes. These have different modalities for motion capture, and different activities performed by the people being filmed. For example, the MPII data set [78] contains videos downloaded from YouTube, with a wide variety of activities and postures. Sixteen hand-labeled body joint positions and several other annotations are provided in the around 30 000 images and videos. Commonly used data sets are created in similar manners, such as the COCO data sets [80] and the J-HMDB database [88]. Popular data sets that contain 3D data from gold-standard systems (3DMoCap) for validation purposes are the HumanEva I & II [89], Human3.6M [90], and TotalCapture [91] databases.

Even though the purpose of this thesis is not to validate a new method of pose estimation, previous works on comparing new pose estimation algorithms to 3DMoCap systems are relevant to our work. Performance of pose estimation algorithms is typically reported as mean per joint position error (MPJPE, in mm) or percentage of correct keypoints (PCK). Although not directly comparable to our metrics of variability, these papers provide a good indication of the performance level that can be achieved in adjacent work. Because the only methods that compare their estimation performance to 3DMoCap data are the ones attempting to estimate 3D joint positions, we will focus on these, despite our focus being on 2D segment length from a single camera.

Previous research has been conducted for the purpose of pose estimation using single camera setups, with one of the most popular pose estimation algorithms being the OpenPose network [92]. OpenPose popularized a technique where the joints and body part positions are predicted using part affinity fields, using a customized CNN model, which has been shown to be an efficient and high-performing framework. More recent models, such as EfficientPose [93], build on this framework to produce even more efficient and robust models. A survey on pose estimation methods [67] showed that a large portion of the earlier proposed methods are using some type of ResNet architecture. These typically achieve 45-90 MPJPE, similar to other methods such as hourglass and CNN-based methods which achieve around 50-80 mean per-joint position error (MPJPE, in mm). The most similar approach to DeepLabCut is the one proposed by Sun et.al. [94], in which a ResNet model is jointly trained on MPII and the Human3.6M data set, and achieves an MPJPE of 48.3 mm. Two other studies also compared their respective models' performance to the ground-truth data in the Human3.6M dataset, Arnab et al. [95] and Mehta et al. ([96], based on the OpenPose architecture), and achieved MPJPE of 54.3 mm and 63.6 mm, respectively. No data on variability is reported in any of these studies. Previous research on segment length variability in monocular image data is scarce, but Bonnechère et al. [97] found that the Kinect had higher variability in the thigh, shank, upper and lower arm compared to that of a 3DMoCap system.

For use in a real-world setting, these models and frameworks need to be built and trained on such large-scale data sets as the ones described above. This requires computational power and expertise to carry out, which is not commonly available in settings where exergames for weight-shifting exercises could be used. Furthermore, the models that *are* trained and tested on the 3DMoCap data sets, typically contain a limited set of persons with a limited set of movements, which could make the models less accurate in situations where the movement being performed is different from those in the training data set. Also, some of the methods provide low-resolution output, which can make them too inaccurate for situations like monitoring exercise performance since the joint center locations will not be placed at a specific enough location. In contrast, the DeepLabCut (DLC) framework allows the user to add contextual information that can specialize the model to the desired application, for example, providing training data about weight-shifting movements during exergaming. Even though the model in the framework is pre-trained on a different data set, this specialization layer allows the DLC model to learn context-specific movement patterns with a relatively sparse data set (<200 images). This feature makes it feasible to adopt this framework in a setting where computational power and knowledge of building complex deep learning models might be limited, such as in the context of exergame development situations.

As can be seen in the previous work, research on pose estimation is focused on each joint center separately, using an average error metric to evaluate the estimation algorithm at hand. This provides an overall evaluation of the performance, which might be sufficient in some contexts. The issue here is that in a real-world gaming situation, the joint center estimation must be stable *over time*. Low temporal variation, or variability, of the estimated joint positions is essential to be able to reward movement pattern performance appropriately. When playing a game, jittering of the joint positions makes the avatar, or other on-screen representations, move incorrectly even though the person playing might be performing the correct movement. This has been a drawback of for example the Kinect camera, where especially the hands, knees, and feet have been prone to unstable motion tracking and jittering [25]. Variability can be represented using metrics such as standard deviation (SD), coefficient of variance (CoV), or other measures of the spread around a central measure [98]. To our knowledge, the comparison of segment length variability between a depth-based camera system, a 3DMoCap system, and a single-camera DL-based pose estimation system has not been conducted prior to the work in this thesis.

3.2 Paper II - Movement Classification

Activity classification, monitoring, or recognition can be seen as different perspectives of assessing the movement pattern a person performs over a given period of time. In this current context of exergaming for balance training, recognizing correctly performed movement patterns in a repetition of an exercise is a core task for the system. Importantly, this task is subsequent to the previously described pose estimation task, as pose estimation data is typically used for activity classification. This assessment provides the basis for feedback to the player, and can therefore directly impact both motivation and effectiveness of the exercise. This is a suitable task for machine learning models, as the input data can be represented as statistical features such as range in movement pattern in a single limb, highest and lowest acceleration, change in orientation of a limb, and so on, and the activity class (e.g. correct or incorrect movement pattern) can be represented as a target value.

One such ML model is called decision trees. Single decision trees are what is called weak learners, because of their poor ability to generalize to unseen data. However, combining many decision trees into an ensemble of trees improves classification performance significantly. The random forests classifier (RFC) is an example of such models. Here, input features are split into random subsets, and the different trees use the features to create splits that best represent the different classes. This is illustrated in Figure 3.5, where the task is to classify inertial measurement unit (IMU) data as being the activity "sitting" or "walking". There are three trees, and each tree has a random set of features around which it constructs the decision making process. At each split, the feature is evaluated on a specific threshold, and the result of that evaluation moves the decision further down the tree. The last node is the leaf node (red outlines), which is the classification that the tree predicts the input most likely belongs to. The final class predicted by the forest is decided using the majority vote from all trees [99]. Decision tree models can also be used to estimate continuous data, such as time series data, in regression tasks. In this case, the leaf nodes predict a continuous value, and the mean of these values over all trees is used as the estimated output [100]. Decision trees try to find the tree with the lowest number of splits necessary to model the output with sufficient accuracy, thereby favoring shallow trees over deeper ones [55]. This might impact prediction performance in data sets of limited size, as there might not be enough information to model the relationship between the input data and output targets. Several methods have been developed to improve the performance of decision-tree based models. One of the most common ones is *bagging*, a shorthand for *bootstrap aggregation*. Bagging is a technique developed to counteract decision trees being very sensitive to changes in input data. The trees are shown training data of the same dimension as the original data, but with some values replaced with random data drawn from the original data set. Here, about one third of the real input data is used [100]. Another method often used is *boosting*, which is a term describing a family of methods for combining features in order to form a strong learner[99]. One technique called Adaptive Boosting (AdaBoost) is popular because of its powerful way of weighing inputs in a feature set [99]. Here, the classifier finds the inputs that are wrongly classified and gives these more weight in the next iteration. This adjusts the classifier as these inputs are now more important to classify correctly. The final classifier is then constructed based on the weighted sum of the individual classifiers, according to their respective accuracy, which then is a classifier that is trained to minimize the error in its predictions.

Other classification algorithms use a different approach to finding the class to which new input belongs. Support Vector Machines (SVMs) are powerful in finding the *support vectors* that separate the classes with the largest distance (or *margin*) between the nearest class points on either side and the vectors. In linear cases (2D space) this is a line, in 3D a plane, and a hyperplane in higher dimensional feature space. However, in cases where the segregation of classes is not possible using linear functions, SVMs can use different *kernels* to transform the data into a feature space into a higher dimension and thus find a possible solution for linearly separating the classes. This is what makes SVMs so powerful, along with its regularization parameters C and gamma. The C parameter is used to adjust whether the decision boundary should be smoother, i.e., simpler, or more focused on exactly classifying all training samples correctly. The gamma parameter regulates how much "reach" each training sample has, i.e. to what degree samples far from the decision boundary impacts the location of the decision boundary. In the k-nearest neighbors (kNN) algorithm, a new input is classified based on the class of the k nearest samples. This method is computationally expensive and is sensitive to the chosen k data points to use for classification. Another algorithm, the k-means, attempts to cluster the samples in a training set with regard to a centroid point that is iteratively moved until it finds the position with the least mean squared distance to the samples around it. A new sample is then classified based on the class of the centroid it is closest to. Here, k refers to the number of centroids that are present in the data.

3.2.1 Previous Work Related to Classification of Physical Activity

Classification of movement patterns has been performed in a large variety of settings and contexts. Within human movement science, activity monitoring and recognition is the most prolific area of research using machine learning methods. This has grown out of the possibility to recognize the activity performed in free-living settings using low-cost sensors (accelerometers or IMU's) that are attached to the body, or by using smartphones [101]. Different methods for activity recognition using video data have also been developed [102]. Typically, classification methods have been developed to distinguish between a wide variety of activities, from physical activity types (running, walking, sitting, biking, see e.g. [103, 104, 105]), to more fine-grained activities performed for example in the kitchen (cutting food, baking, pan-frying, etc.[106]). There is also a large body of research on monitoring the *amount of* activity a person is performing in the course of a day or a week using pedometers or other accelerometer based systems [107, 108]. In an industrial context, activity recognition is being used in for example ergonomic risk assessment [109], and in surveillance settings behavior can be recognized using video data [110].

Another area within physical activity classification is recognition of exercises, and more specifically, classification of the performance of movement patterns within an exercise (e.g. [111, 112]). This has been conducted in elite sports athletes looking to fine-tune movement techniques [113], but also in rehabilitation of orthopedic patients [114]. Here, activity classification is useful to inform the person about their movement pattern, by providing feedback on how it can be improved to keep exercise effectiveness high. In an exergame setting, assessing movement patterns is an essential part of the game system. The assessment is used to evaluate whether the user completed a task, and what feedback to provide - i.e., whether to reward the user with points or other in-game rewards. This assessment has typically been done using coarse rules and thresholds, although in recent years more and more machine learning methods have been used to assess performance compared to a template movement. In Zhao et al. [115] and Gal et al. [116], movements during exergaming are assessed using pre-determined rules that specify the intended movement pattern. Others have developed comprehensive systems for eliciting, modeling, and assessing human movement, such as Ofli et al. [117], Lam et al. [118], and Tao et al. [119], where movement patterns are compared to template movements as well. Even though these methods work to a satisfactory degree in healthy adults, it is not known how well they perform in older adults or in patient populations.

In contrast to this earlier work, we propose to train machine learning methods by directly using joint center position movement patterns. In this manner, the algorithms are trained on a data set containing weight-shifting movement repetitions that have been classified into either being correctly or incorrectly performed, with no further information about *why* the repetitions were classified like this. This way, there are no hard constraints coded into the classification system, only example movement patterns representing correctly and incorrectly performed weight shifts. This allows for more flexibility in the system since the models are trained on people of different ages, body sizes and -shapes.

3.3 Paper III - Force Estimation

Exercising balance during in-home exergaming is challenging because existing exergame systems do not measure the force you exert on each foot while exercising and thus are in lieu of trustworthy data to provide feedback on. And, as we know, feedback is essential for efficient exercise, as improving and maintaining correct movement patterns depends on receiving information about your movement performance. To make exergames for balance training feasible and available, it is vital to enable access to force data without introducing new equipment. One manner of doing that could be to use machine learning models such as recurrent neural networks that are created to estimate force from already available time series data, such as joint center positions. Therefore, this is also a task that is subsequent to pose estimation, although not necessarily to activity classification.

Sequentially ordered data inherently has a time-dependency. This can be for example data from speech, data from weather sensors, stock trade data, or video image data. This time-dependency entails that the information in previous data points can be used to inform what the next data points can, or might, be. Furthermore, not just the previous data point, but additional data points further back in the data stream might be important for predicting the next data point. Neural networks, as described earlier, organize the input data so that one input node corresponds to one data point in the data set. This manner of organizing data makes it extremely difficult to capture dependencies between the current and past data points [59]. Researchers thus developed *recurrent* neural network (RNN) architectures, where the data output from a node is passed back to the same node and used to inform the next iteration of output from that layer in the network. This makes it possible to learn long-term dependencies without needing exponentially more training data. One of the core features is *parameter sharing* between the recurrent hidden nodes, along with the feature of returning the output of a node to that same node for each iteration of training [56].

Even though a typical task in time-series modelling is to forecast or predict the next data point, e.g. the next word in a sentence or the next day's stock value, it can also be used in other tasks where the input data is a time series. In the context of this thesis, Long Short-Term Memory (LSTMs)[76] recurrent neural networks were favored in the estimation task of force during weight-shifting, because of the inherent time-dependency between the previous and current joint positions observed and force level measured.

LSTM recurrent neural networks also introduce a new concept within each (recurrent) node. Here, there are *gates* inside the node that controls the flow of information through activation functions that the predictions are passed through after each gate has made a prediction based on previous experience. Figure 3.6 shows the different gate types that are used in LSTMs: forget gates, selection gates, and ignoring gates. After input is provided to an LSTM node, it is passed to each gate simultaneously, along with the new input data. In Figure 3.6, the left-most path is a prediction, which is a normal vote for the weight that the node should provide to each output node, run through a sigmoid activation layer. The ignore gate (middle-left path) is often called an attention mechanism. This picks up on which information can be ignored in a signal and blocks that information from the prediction. This information is then passed through an activation function, usually tanh, which squeezes the values from the prediction to lie between -1 and 1. This information is then added to the next prediction by element-wise multiplication, effectively blocking, attenuating, or letting signals in the next prediction go through. The middle-right path is the forget gate. After the first output from the node is made, the forget gate has stored information about the previous input. This gate then stores what is most likely **not** a good prediction in the next iteration. Some LSTM models leave this gate out, but we chose to keep it in since it has been shown that removing information that is unnecessary from the prediction improves performance [120]. The last path (right-most in the figure), the selection gate, works in the opposite way of the forget gate. It has learned the predictions of what are most likely good predictions, and by using a tanh activation it enhances or attenuates signals, also through element-wise multiplication of the prediction.

Some hyperparameters are important to be conscious of when training an LSTM network. First of all, the number of layers, and nodes in each layer, will be decisive in finding the right balance between an underfitting and an overfitting network (described above). Too few layers and nodes can result in an underfitting model, and too many can cause overfitting. In our context, the complexity of the data set, and output data warranted more than one LSTM layer, and a relatively large number of nodes [120]. We chose to use three LSTM layers with 512 nodes each. To avoid overfitting, we added dropout (20 %) to the model. This layer reduces the sensitivity of the model to specific nodes, as it makes the model skip a randomly selected percentage of nodes. Furthermore, the learning rate of the model is one of the most important hyperparameters, as it defines how large steps the model will take to search for a smaller loss, or cost. Too large steps can result in a model that misses the optimal loss area, and conversely, a too small learning rate can cause a very slow learning rate. We set a learning rate of 0.0001 [120]. The number of iterations, called epochs, that the model trains on the data set are also important. To few epochs can create an underfitting model, as the model has not been trained enough, and too many epochs can result in an overfitting model. We set an initial 200 epochs for training, but with an early stopping method where the training was stopped if the loss was reduced by less than 0.0003 for three consecutive epochs. To account for the complex training of recurrent neural networks, we used the adaptive ADAM optimizer was employed [121]. The activation function used in this study was the sigmoid, as it produces output that lies in the 0-1 interval which is similar to the output data we are estimating. The output layer consists of six nodes, as there are six dimensions we are estimating force data for (x,y, and z axis for left and right foot).

3.3.1 Previous Work Related to Force Estimation

Force estimation from kinematic data has been investigated in other settings than balance training. In gait and running analysis, kinetic data such as ground reaction force (GRF) is important for assessing the causes of (sub-optimal) movements and gaining insights into the mechanics involved in movement deficiencies and injuries. There are two distinct approaches to the estimation of GRF, namely using biomechanical modeling such as inverse dynamics [122], and the use of machine learning approaches, which has been rising in recent years. Estimating GRF or center of mass (CoM) displacement over time using data from full-body IMU setups for inverse dynamics methods, or other biomechanical models has been shown to yield good results during gait [123, 124, 125, 126]. However, these models are computationally expensive and might not be feasible to use in low-resource devices or in-home settings. GRF and moments during gait have also been estimated with high accuracy using machine learning models with additional input from biomechanical models, calculated from full-body IMU data [127, 128]. Some also use IMUs to successfully estimate joint loading [129], or ground reaction force during activities of daily living [123].

The drawback of using full-body IMU systems is the practical requirements of use. Not only is attaching all the sensors time consuming and requiring expert help to do correctly, but also maintaining and managing such sensor systems adds to them not being feasible to use in elderly care homes or rehabilitation hospital settings. Kinematic data from simpler systems such as depth-based cameras or AI-based pose estimation systems using standard digital video are much more feasible for use in these settings. Previous research on using data from a biomechanical model from a camera-based 3D motion capture system also shows promising results. For example Mundt et al. [130] showed that both a recurrent LSTM neural network and a feed forward (FF) neural network are able to estimate both tri-axial joint moments in the lower body and ground reaction force with high accuracy. Oh et al. [131] and Choi et al. [132] also used features from a biomechanical model calculated from 3DMoCap data to estimate tri-axial GRF and moments during gait, with high accuracy in their results.

Using recent developments regarding powerful machine learning models, we propose a solution that could circumvent the computationally and practically costly approaches that require full-body marker placements and biomechanical models to estimate GRF. The combination of ML models that capture time-dependency and are able to model complex relationships between features and target values, such as LSTM, and joint center positional data, could provide a method for estimating GRF without needing the computational layer of biomechanical modeling. If these joint center positions were extracted from an ML-based image-analysis system such as DeepLabCut, GRF could be estimated by using a standard video camera only. However, to our knowledge, the direct use of joint center positional data to inform a machine learning model has not been investigated in terms of performance and accuracy in estimation, which is what we aim to do in this thesis.

3.4 Model evaluation procedures

As mentioned above, the ability of the models to generalize to other data than the specific data set it has been trained on is essential. This means that the model should be able to represent the data-target relationship in such a way that it can accurately predict, estimate, or classify input from the same context it has not seen before. If the model is to be used in real-world applications, this aspect is vital. Measuring a model's performance in a standardized manner is therefore essential to assess how well a model performs in the given task. There are several evaluation methods available, and here we detail the most important ones and the ones we have used in the work this thesis is based on.

The terms *overfitting* and *underfitting* are used to describe a model's ability to generalize to new data, as depicted in Figure 3.7. Both underfitting and overfitting models have low accuracy in the classification of a new data point and have a high error in estimation situations. In underfitting, the model is not specific enough in following the data points' changes over time or in small variations in the features, and in overfitting, the model is too specific and fails to generalize to new data, despite following the training data accurately. The middle panel in Figure 3.7 is the preferred performance of the model on the training data. Here, the model is not too specialized on the training set but specialized enough to make generalizations to new data inputs.

To test for overfitting or underfitting, it is common to use what is called a holdout technique. Here, part of the data is not being shown to the model while it is trained, and the model is then evaluated ("tested") for accuracy on this held-out data set. This is then repeated until all the data has been used both as training and as test data, and the performance of the model is represented as the average accuracy over these iterations. This technique is called *cross-validation (CV)*. There are several different ways to construct this test/train split in the data, i.e., different cross-validation methods. The most appropriate CV method to use depends on the task at hand, the structure of the data-set, and the domain, amongst other factors. For example, in the current setting of exergaming for balance training in older adults, it was natural to keep out all data from one participant for each iteration. This provided an average performance based on the models' ability to generalize to the movement pattern of a person it had not seen before. This method is called *leave-one-group-out* CV, or LOGOCV. Other methods include k-fold CV, where k sets of the data (e.g., data from three participants) is held out from training, and holdout CV where a randomly selected subset of for example 20% of the data, across participants, is held out. In time series data for forecasting, it is also common to use a CV method where the data at the last time point is held out, while the rest of the data is used for training. This is then performed iteratively until all data points of the time series have been used as holdout data.

There are several different ways to represent how well a model performs in a given classification task [133]. Accuracy, the ratio between correctly classified samples over the total number of samples, is just one of many metrics. As accuracy is sensitive to imbalanced data sets, where there is a major disproportion between the number of samples in the different classes, other metrics such as F1-score have been developed that provide a better picture of model performance. The ratio between correctly and incorrectly classified samples can be represented in different ways. These metrics might be more appropriate to use, depending on the important factor to consider is when evaluating the model. In figure 3.8 the different metrics are illustrated using what is called a *confusion matrix*. The confusion matrix is a commonly used evaluation method in other domains as well where statistical evaluation is important, such as in epidemiology. As figure 3.8 shows, *recall* is the ratio between positive samples correctly classified as belonging to the positive class. *Precision*, however, is the ratio between positive samples that were correctly classified over the total number of positively predicted samples. These are important metrics with important distinctions, as they inform about the ratio between the number of samples that were predicted to be in the positive class and the samples that actually belonged to the positive class. *Specificity*, on the other hand, represents the number of samples correctly classified as belonging to the negative class, over the total number of negative samples. The two remaining quadrants, false positives and false negatives, represent what is more commonly known as Type I and Type II errors, respectively. In some contexts, for example classification of animal species from images, these errors might not be serious, but in medical domains, a false negative (or Type II) error can have grave consequences for a person seeking medical attention, as a condition that is present in the patient is not detected and thus not treated appropriately. Thus, the choice of metric is an important step in the evaluation of model performance, and domain knowledge must inform decisions on what is more important to avoid or focus on in evaluation.

In estimation or forecasting tasks, model performance is commonly evaluated using a distance metric from the predicted value to the actual value in a data set. There are several different metrics in this area as well, and the choice of method again depends on the purpose and domain of the model in question. The common component in these methods is that they represent the (average) error in the estimations performed by the model. Mean square error (MSE) and root mean square error (RMSE) are two related metrics, as the latter is the square root of the former. RSME is considered more useful as the squared average (e.g, squared seconds) is more complicated to comprehend than its square root (seconds), as it brings the metric back to the original unit of measure of the data. These metrics are calculated as shown in Equations 3.1 and 3.2. Here, $\hat{y}$ is the predicted value of y (i.e., the output from a model), and $\overline{y}$ is the mean value of y.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2 \tag{3.1}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \tag{3.2}$$

In time series prediction, it is also common to evaluate the performance in terms of correlation, using the coefficient of determination (R^2). This metric is used for determining how much of the variability in the regression model can be explained by the features in the data set [56], and is calculated using Equation 3.3.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \overline{y})^2} \tag{3.3}$$

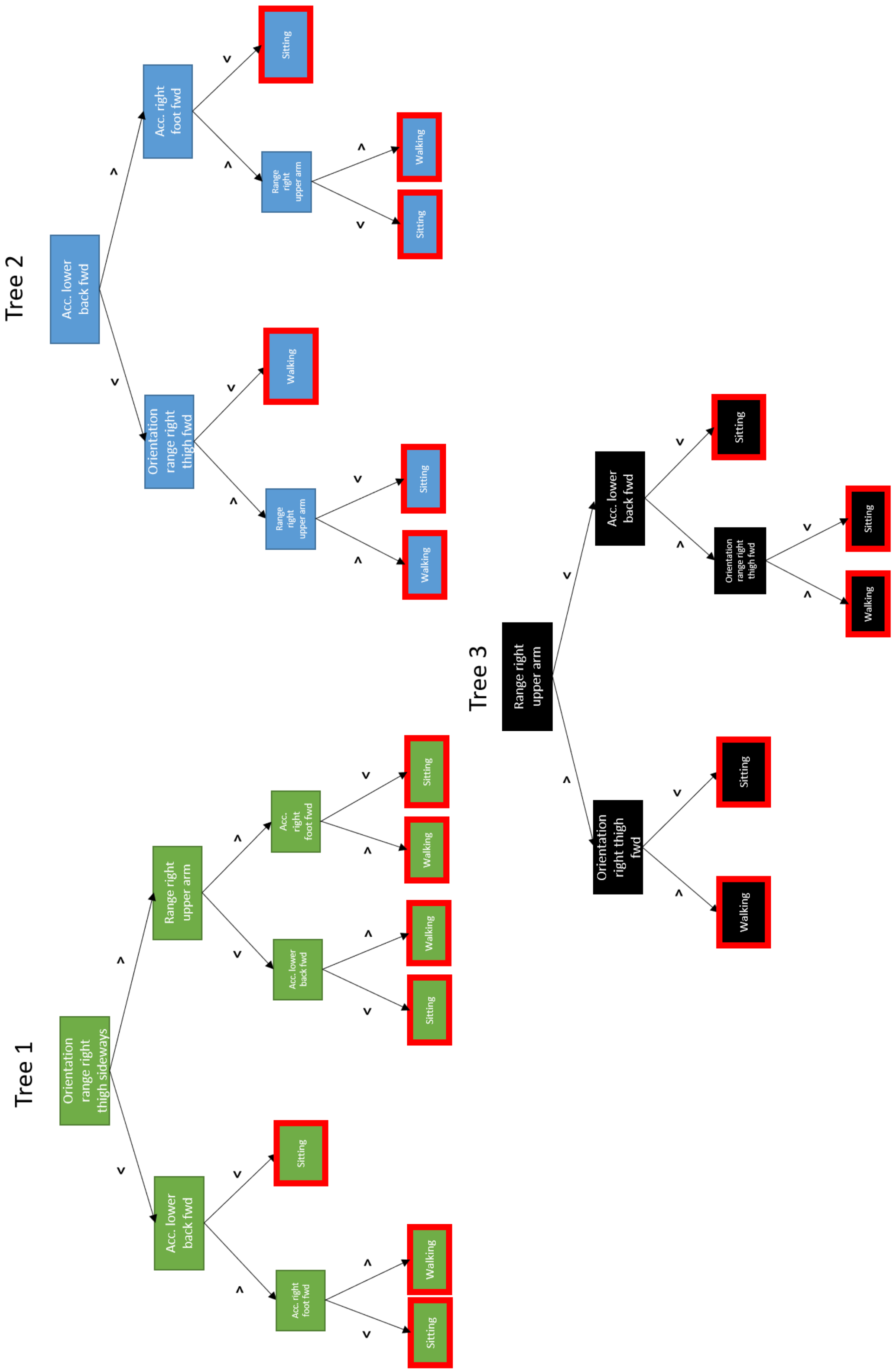


Figure 3.5: Example of three decision trees with randomly split subsets of features.

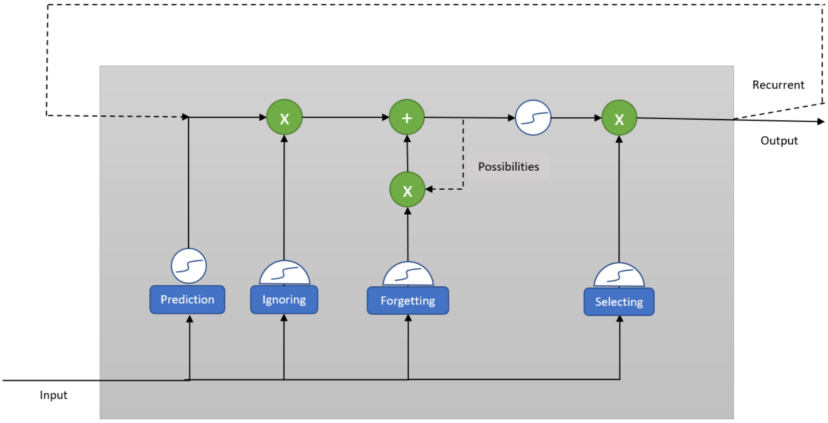


Figure 3.6: A schematic of an LSTM with an input, a prediction, an ignore gate, a forget gate, and a selection gate. X-symbols mark element-wise multiplication, and $+$-symbols mark element-wise addition. Tanh activation functions are denoted with a half-circle, sigmoid with a complete circle. The dotted line is the recurrent information that is passed back to the gate from the output.

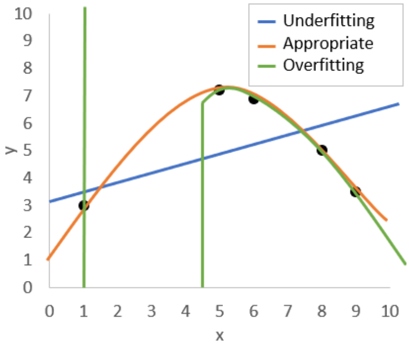
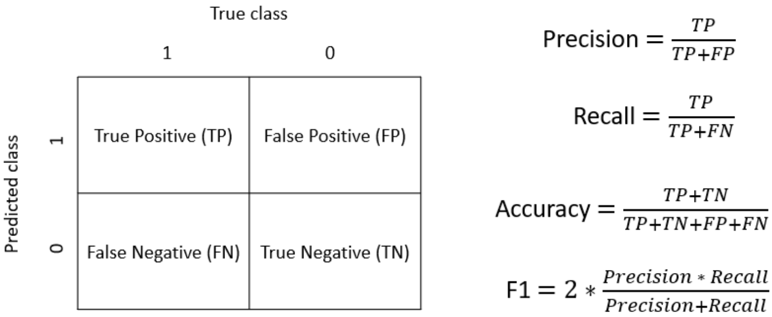


Figure 3.7: Underfitting model (blue), appropriate model (orange), and overfitting model (green). *Adapted from Goodfellow et al 2016.*



Predicted class	True class: 1	True class: 0
1	True Positive (TP)	False Positive (FP)
0	False Negative (FN)	True Negative (TN)

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{F1} = 2 * \frac{Precision * Recall}{Precision+Recall}$$

Figure 3.8: Description of precision and recall, accuracy and F1-score.

Chapter 4

Methods

4.1 Design and Ethical Considerations

The studies this thesis consists of are based on data collected at the Motion Capture and Visualization Laboratory at the University of Science and Technology in Trondheim, Norway, in the spring of 2019. The thesis is within the medical technology field, an interdisciplinary field combining medicine and health sciences with technology and computer science.

4.1.1 Design of data collection

With the research questions in mind, the data collection was designed to provide the necessary data in the most efficient manner possible. Through careful planning, we found the optimal way to acquire all the data needed for all three studies without requiring the participants to spend more than 1.5 hours at the laboratory. Because there is scarce research on using machine learning in exergaming, the movement pattern was designed to be simple and not contaminated from noisy data, yet relevant for balance training. This potentially provides us with more clear answers than more complex movement patterns could have yielded as performance by the algorithms is likely to be more difficult to interpret if the movement patterns are very noisy or complex.

This choice of design is also reflected in the data protocol we decided to employ. By having low-noise, high accuracy data from a 3DMoCap system, simultaneously captured with in-home appropriate modalities such as video data and depth camera data, we could also compare results using these different data types, which resulted in interesting analyses as presented in Chapter 5.

4.1.2 Ethical considerations

The project and data collection were assessed by the Regional Ethics Committee prior to project start. In addition, the National Data Security committee approved the project with regard to data security, handling of personal information, and plans for anonymizing the data.

In accordance with the Declaration of Helsinki, before data collection started, participants were given oral and written information about the project, the data collection procedures, possible benefits and inconveniences, and had the opportunity to ask questions whenever they might arise. All participants signed a written informed consent form.

As the project aimed at improving technology meant for use by older adults, the recruited participants were in this demographic. Having an elderly participant group perform balance training movements could pose a safety risk, because of the changes in the neuro-muscular system described in Chapter 1. To minimize the risk of injury or adverse events during data collection, extra care was taken to ensure the safety of the participants, both in the design of the exergame and in planning the movements to be performed. The exergame was designed in close collaboration with a computer engineering student, who developed the game prototype (see [134]). The game was initially designed for stroke patients, so it lent itself nicely to use for elderly users as the principles for designing games for these two demographics are virtually the same (low-key music, low contrast colors, slow game speed) (ref). The design made it possible to follow the game play for all participants, as it was flexible enough to let participants find the movement speed and strategy they felt safe and comfortable using. No sudden or rapid movements were required, and the extent they needed to lean medio-laterally to get the in-game reward was manageable for all participants. Furthermore, throughout the data collection, there were two project representatives present, one of which was responsible for the safety of the participant during playing.

4.2 Participants

The recruited participants were healthy elderly (age >65 years) with no history of balance issues or loss of motor function that could render their participation potentially hazardous. Participants were given information about the project from a researcher in the project visiting their local exercise groups for seniors. Here, participants were given an opportunity to ask questions and sign up for participation at their preferred time and date. This recruitment process was used in both the pilot study and in the main study.

4.3 The exergame

For the purpose of eliciting movements from the participants, a custom exergame was designed and developed. This game consisted of two versions, where each version was designed to elicit a specific movement pattern from the participants. Common for the two versions was the interface and the goal of the game, as presented to participants. A screenshot can be seen in Figure 4.1. In the top right corner, the number of coins collected can be seen, and in the bottom left corner the speed of the cart can be controlled. The coins can be seen on either side of the rail. There were approximately 100 coins in total, and there were never more than two coins successively on each side at a time. For each coin, the player could receive between 1-3 points, depending on their movement pattern. Above the rail, in the center of the screen, a white star is superimposed on a multi-colored bar. The star shows the position of the virtual marker (shown in Figure 4.3(a) and 4.3(b)) in the x-direction (sideways).

Its position on the colored bar as the cart hits a coin determines the number of points the player receives for that coin. This is where the two game versions differ: in the first game version, the bar was grey with a dividing line in the middle, as seen in Figure 4.2(a). In the second game version, the bar is divided into three colors, as seen in Figure 4.2(b). The two color bars provided points based on the position of the star. In the first version, with the gray bar, the maximum score possible was awarded if the player managed to have the star as far laterally as possible when a coin was hit. By performing this movement pattern, the player completed a correct weight shift. The upper body was leaned over the foot, making the majority of their body weight loaded on the foot they were leaning towards, as shown in Figure 4.3(a). Conversely, the colored bar in the second version of the game gave three points if the star was in the green area in the middle of the bar. To achieve this, the player had to lean their lower body sideways to make the cart hit the coin, but keep their upper body leaned in the opposite direction to keep the star within the green area, shown in Figure 4.3(b).

4.4 Equipment

For 3D motion capture data, a setup with four different data modalities was used to capture simultaneous data. Four OQUS cameras (MX 400, Qualisys AB, Göteborg, Sweden) was used. Cameras were capturing at 90Hz. 36 reflective markers were placed according to the Plug-in-Gait Full-Body marker placement guide (PiG-FB, [135], head and hands excluded), by an experienced human movement scientist and an experienced assistant. Two digital cameras (GoPro Hero 3+ Black, 1400 × 720 px, GoPro Inc.) were used, one positioned sagittally, at the right-hand side of the participants, and one frontally, 200 cm behind the participants. These

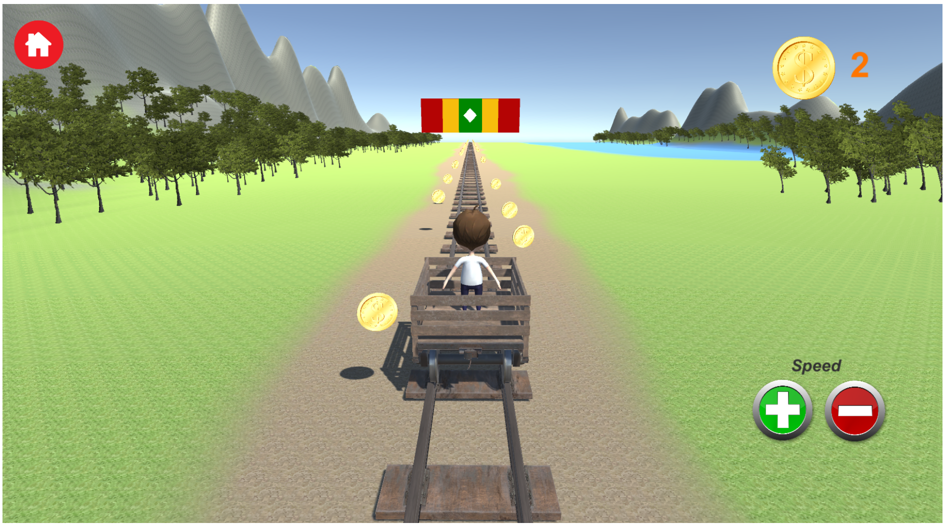
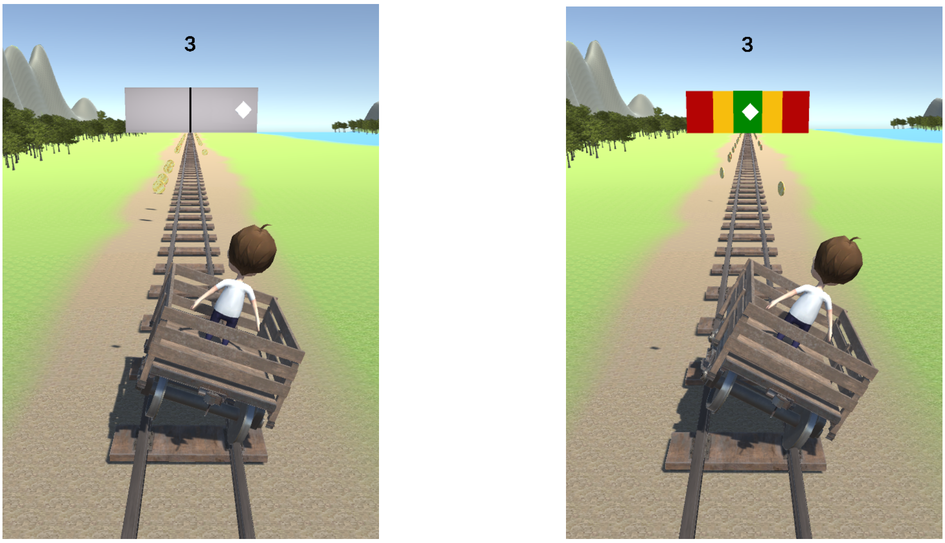


Figure 4.1: Interface of the game.

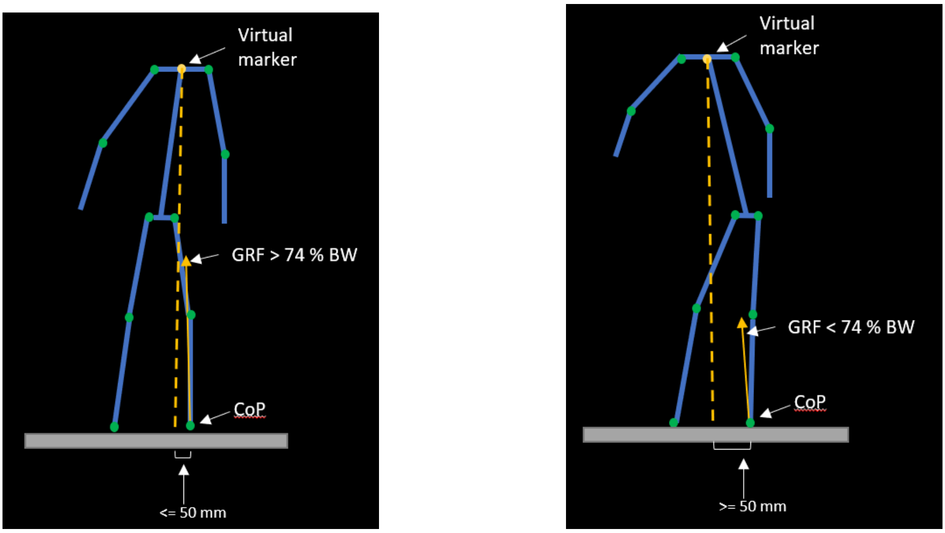


(a) Part 1. Two-split grey bar, shown at the end of the track, with the star to the right of the dividing line, rewarding 3 points.

(b) Part 2. Three-split color bar, shown at the end of the track, with the star in the middle 33%, rewarding 3 points.

Figure 4.2: The two versions of the exergame.

cameras were synchronized with a remote control. Two force plates (1000 Hz, 60×40×5 cm, Kistler Nordic AB) were positioned in the center of the playing



(a) Typical body posture when being rewarded 3 points in part 1 of the game. Here, the player is leaning their upper body over their weight-bearing foot, resulting in the distance between the virtual marker and the CoP of the weight-bearing foot being <50 mm, and the GRF Z-component being >74% of body weight. BW = body weight. GRF = Ground reaction force. CoP = Center Pressure.

(b) Typical body posture when being rewarded 3 points in part 2 of the game. Here, the player is not leaning their upper body over their weight-bearing foot, resulting in the distance between the virtual marker and the CoP of the weight-bearing foot being >50 mm, and the GRF z-component being <74% of body weight. GRF = Ground reaction force. CoP = Center of Pressure.

Figure 4.3: Typical body postures when playing the two different exergame versions

area, one for each foot of the participants, with a 60×30×5 cm platform extending on each lateral side of the force plates. As input to control the game, a Kinect (V2, 25 Hz, Microsoft Inc.) was placed according to recommendations in front of the player. A schematic of the experimental setup can be seen in Figure 4.4.

4.5 Data collection

4.5.1 Pilot study protocol

Data for the pilot study was collected in November 2017. 11 participants were recruited, six females and five males, with an average age of 69.3 years (1SD 4.0). Participants were recruited in the same manner as in the main study. The pilot study aimed to investigate the accuracy ML models could classify incorrectly and correctly performed movement patterns of weight-shifting. The movements performed in the pilot study were not elicited from a game but were instructed movement patterns. Two movement patterns were performed: one without a clear weight-shift, i.e., an incorrect movement, and one movement pattern with a clear weight shift, i.e. a correctly performed movement. Three ML models that are commonly used in activity classification were employed: kNN, RF, and SVM. The results were promising, with an average accuracy of 0.989, 0.949, 0.958 for RF, kNN, and SVM, respectively. The highest was RF on all joint centers and SVM on shoulder joint centers (both 0.996). The lowest was k-NN on ankle joint centers (0.879). Results showed that it is indeed feasible to classify movement patterns, prompting us to continue with a data collection where movements were naturally elicited from an exergame.

4.5.2 Main study protocol

Upon arrival at the laboratory, participants were given oral instructions and information about the activity to be performed. After changing into comfortable shoes and clothes as instructed, height and weight were recorded. The reflective markers were attached and the data collection started. Participants performed three trials of each version of the game, totaling six trials for each participant. Between all trials, the player was offered a break where they could sit on a chair to relax if needed.

4.6 Data processing and analysis

3D joint centers from the 3DMoCap system were used in all three studies. Using the standardized PiG-FB biomechanical model [135], joint centers were extracted for each time frame of the data using Vicon Nexus (v. 2.10, Vicon Motion Systems Ltd, UK). Joint center locations are represented as a three-dimensional point vector in Euclidean space relative to the origin of the Qualisys lab coordinate system. A schematic of the extracted joint centers can be seen in Figure 4.5, with details of joint center definitions in [135].

2D joint centers from the DLC framework, based on GoPro video data, were used in study I and study III. To obtain joint center positions from the digital video data, an experienced human movement scientist manually labeled the joint center locations as seen in Figure 4.5 in three images from two videos of each participant, totaling 194 images. This is in line with the recommendations for training data for the DLC framework [87]. The joint center location data and the video images were then used to train the DLC model, and predictions of joint center locations in unseen videos were produced. The train/test split was set to 95/5, and the model was trained for 220000 iterations with loss plateauing at 0.0012. In four participants (4,8,9, and 10) the frontal camera view of the ankles was obstructed, and thus not labeled. Ankle joint position predictions from these participants were therefore excluded from the analysis.

For all further analyses, Python (v. 3.7-3.11) was used, with the machine learning packages Sci-Kit Learn [136], Keras [137], DLC[1] , and TSFresh[2].

4.6.1 Paper I - Pose Estimation

This paper focused on comparing the variability of segment lengths from the 3DMoCap system, the Kinect, and the DLC image analysis system.

Joint center positions from the Kinect camera were extracted using Kinect Studio (ref) in conjunction with Kinect2Toolbox (ref). Segment lengths from all three systems were extracted as the distance in Euclidean space between joint centers, e.g., the lower arm segment was the distance between the wrist and the elbow joint center, as seen in Figure 4.5. These lengths were then calculated for each time frame for each of the three data types. Variability in segment lengths was represented as mean standard deviation (SD) and mean coefficient of variability (CoV). To evaluate the statistical significance of the difference in segment length variability, a Shapiro–Wilks test for normality was first conducted. This gave a $p < 0.05$ for all segment lengths. Therefore, the non-parametric Friedman test was conducted to assess statistical differences in segment length variability between the three systems. Lastly, a Wilcoxon's Signed Rank test post hoc analysis was conducted on the statistically significant differences from the Friedman test in order to extract which between-system differences were statistically significant. A Bonferroni correction was used, resulting in $\alpha = 0.017$.

4.6.2 Paper II - Movement Classification

Paper II was an assessment of the performance of different ML algorithms in classifying correctly and incorrectly performed weight-shifting movements. Each repetition of a movement was extracted from a whole trial sequence of movements. One repetition was defined as the movement between the most lateral position in the y-direction of the right shoulder marker and its most contralateral position. To determine movement correctness, the midpoint between shoulder joints and the GRF was used, as shown in Figures 4.3(a) and 4.3(b). A repetition resulting in <74 % BW, as well as a midpoint horizontal position of >50 mm from the ankle joint in the y-axis, was deemed as incorrectly performed (i.e. was not considered a complete weight shift). Each repetition was then represented by a set of statistical features calculated using the TSFresh package [138] for Python. The original features were also run through a Principal Component Analysis (PCA) to reduce the dimensionality of the feature space. Both the statistical features and the PCA features were used as input data to the ML models separately. Each model was trained on the entire dataset, using all joints, and also on just one set of joint positions at a

[1] https://github.com/DeepLabCut/DeepLabCut

[2] https://github.com/blue-yonder/tsfresh

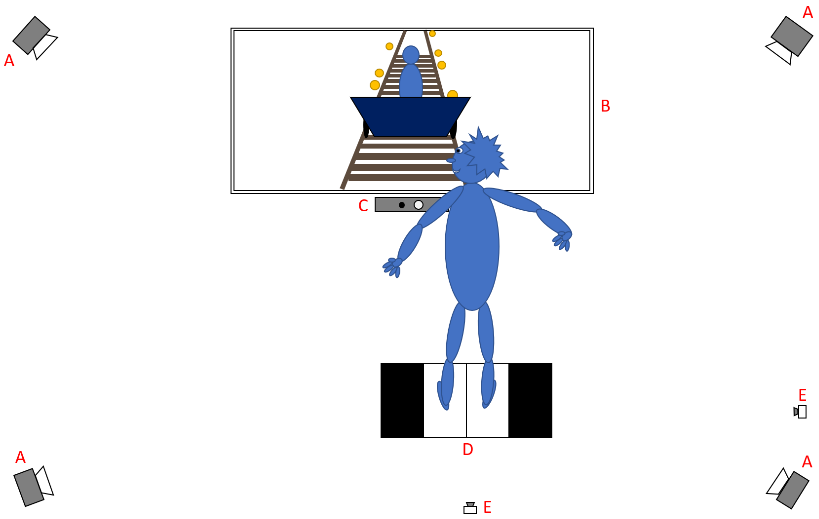


Figure 4.4: Experimental setup illustration. A = Qualisys cameras, B = 82" TV, C = Kinect camera, D = Force plates, E = GoPro cameras. Not to scale.

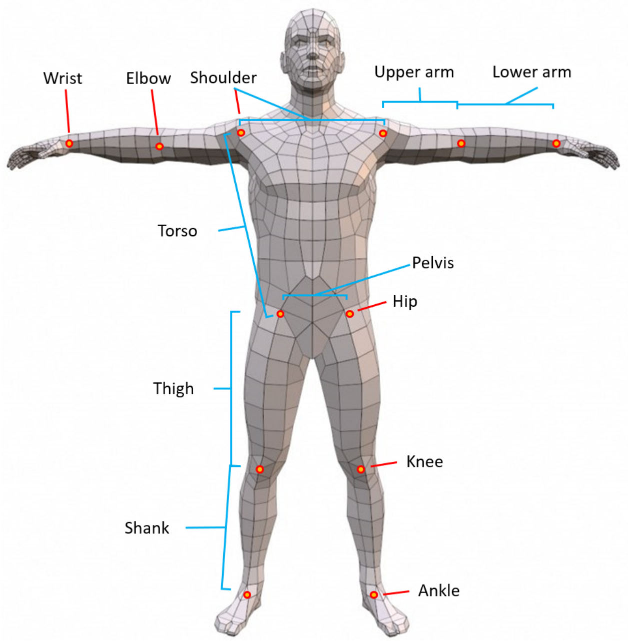


Figure 4.5: Schematic of the joint center positions (red) and segments (blue).

time, for example shoulder joints only or hip joints only.

The models employed in this study were kNN, random forest, support vector machine, and a neural network classifier (multi-layer perceptron, MLP). A leave-one-group-out cross-validation was performed, where a group consisted of one participant's data, resulting in 11-fold cross-validation. Recall and F1-score were the metrics extracted for the evaluation of model performances.

4.6.3 Paper III - Force Estimation

In Paper III, we investigated the ability of an LSTM network in the estimation of GRF using kinematic data. Here, we used DLC and 3DMoCap joint center positions scaled to the 0-1 range as training data in separate model training and testing iterations. The x,y, and z force components from the force plates, normalized to the participants' body weight, were used as target variables. A stacked, 3-layer LSTM network, an XGBoost model [139], and a standard multiple linear regression model (LinReg) were employed. A 11-fold cross-validation was performed similarly to the procedure in Paper II. Evaluation metrics were mean and standard deviation of R^2 and Root Mean Square Error (RMSE) over the cross-validation iterations.

Chapter 5

Results

This chapter presents an overview of the results from the three studies that were conducted, with a focus on the results' contribution to answering each of the research questions presented in Chapter 1. Details from the results can be found in the respective papers in Part II. Figure 5.1 presents an overview of results from Paper I, Figure 5.2 shows results from Paper II, and Figures 5.3 and 5.6 show the results from Paper III.

5.1 Results overview

Overall, the findings from this thesis demonstrate that in all three aspects focused on, ML can indeed improve usability and availability. The motion tracking technology can be simplified using deep learning and image analysis on standard digital video data while retaining the all-important tracking accuracy needed to provide appropriate feedback. We can also use ML-based algorithms to analyze movement patterns in a manner that identifies correctly performed weight-shifting movements regardless of e.g. body size. Lastly, our findings show that it is possible to measure weight-shifting in an accurate manner without needing specialized force sensing equipment. This can enable feedback on weight-shifting performance just by using a standard digital camera.

5.1.1 Research Question I

How does a state-of-the-art deep learning 2D image analysis system perform with regard to segment length variability compared to a 3DMoCap system and Kinect system?

The overall aim here was to compare the accuracy of a deep learning-based pose estimation system (DeepLabCut, DLC) to a state-of-the-art marker-based system and a widely used marker-less system. To this end, we compared the intra-segment variability of the three systems, as presented in Paper 1. Segment variability, i.e., the change in segment lengths over time, is in and of itself also an interesting measure of performance in any motion tracking system, as stability in motion tracking is essential to ensure that the player's movement pattern is consistently represented in the game system. These results would directly affect how to conduct the analysis for answering RQ III, where DL-based joint center data could potentially be utilized. As Figure 5.1 shows, the average DLC and Kinect system joint center location variability were overall higher than that of the 3DMoCap system. However, not all differences were statistically significant, indicating that in some segments the variability of the DLC and Kinect was not systematically higher than in the 3DMoCap. The segments where variability was not statistically different between 3DMoCap and DLC were left upper arm, left torso, and left and right shanks. Compared to Kinect segment variability, the only statistically significant differences were in the left lower arm and left thigh. Thus, in all segments, the DLC system performed with comparable segment length variability to either the ToF or the 3DMoCap system. In all three systems, the shoulder segment variability was the lowest, and the lowest overall SD was in the shoulders (2.8 mm) using 3DMoCap.

These findings show that a ResNet-based deep learning image analysis system, DeepLabCut, is a viable option for accurately extracting joint center locations for use in in-home exergame settings that require low body segment length variability.

5.1.2 Research Question II

To what extent can machine learning classification models identify correctly performed movement patterns during weight-shifting exergaming?

RQ II was the first to be explored, as the original aim of the thesis was to identify correctly and incorrectly performed movements during exergaming. To answer this inquiry, different movement patterns from the participants were elicited during exergaming: incorrectly and correctly performed medio-lateral weight shifts. The task for the machine learning models was first of all to identify the correctly performed movements, as these are essential for rewarding the player appropriately during gaming. The analysis also focused on classification accuracy using only joint subsets versus using all joint center data combined, and on whether different cross validation methods and feature representations affected classification performance.

Results for this analysis showed that the MLP and RFC models outperformed the kNN and SVM models, as seen in Figure 5.2. The MLP achieved a mean recall of 93.4% when trained on any joint subset, which was the highest performance

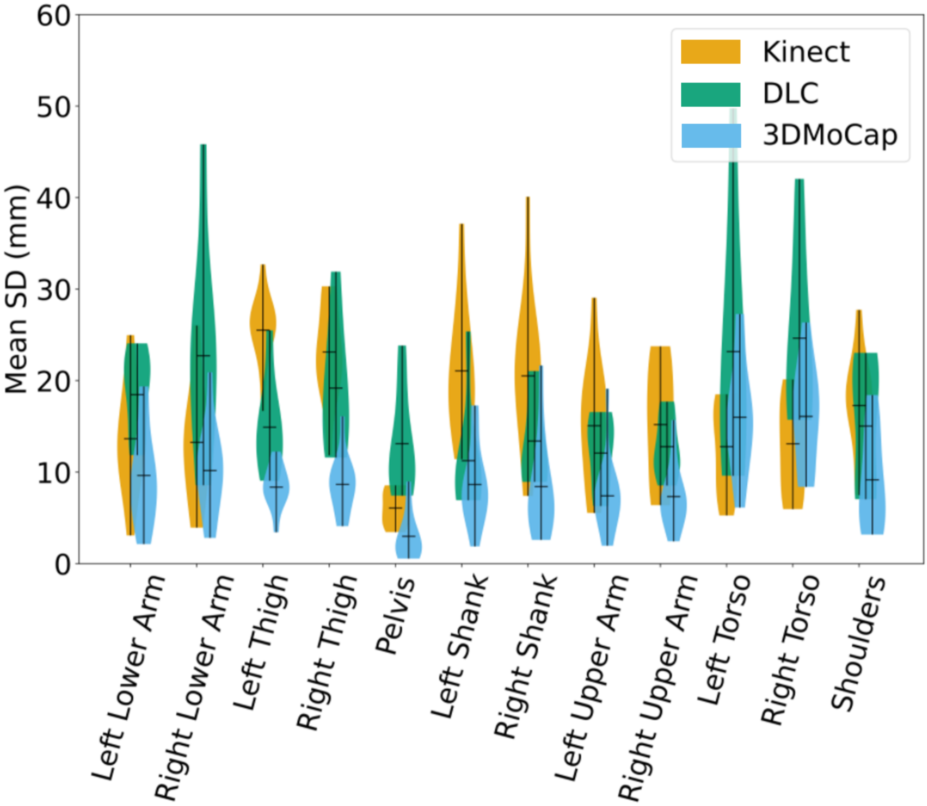


Figure 5.1: Results from Paper 1 - Pose Estimation. Mean standard deviation (SD, in mm) of each body segment length by the different motion capture systems.

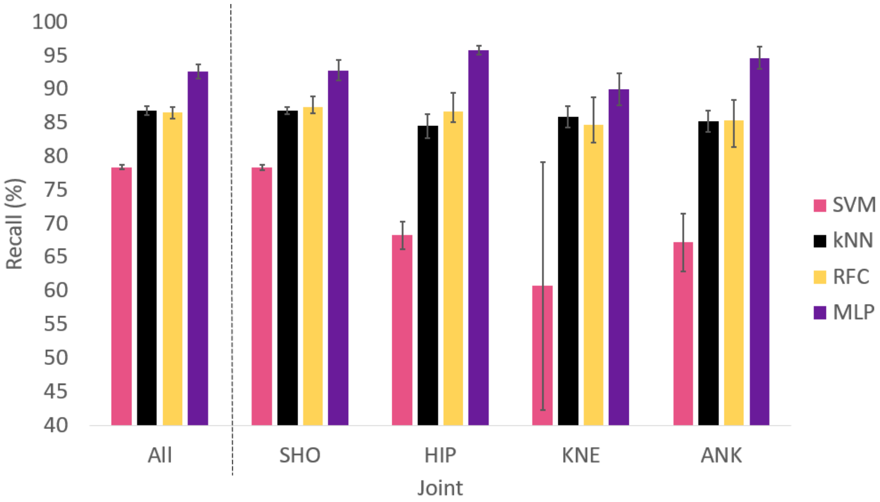


Figure 5.2: Results from Paper 2 - Movement Classification. Mean recall of classification performance for all four models, using all data and using joint subsets. SHO = shoulders, HIP = hips, KNE = knees, ANK = ankles. RFC = random forest classifier, SVM = support vector machine, kNN = k-Nearest Neighbors, MLP = multi-layer perceptron.

of all models. The RFC model achieved a 7% lower mean recall when trained on any joint subset, while the kNN and SVM models achieved 85.6 % and 68.7% mean recall, respectively. The highest performance on a single joint subset was by the MLP using hip joints (96.5% recall), but the most consistent joint subset for all models was the shoulder joint, which had the highest mean recall of 86.3%. Between the two CV methods, 10-fold (CV10) and leave-one-group out (LOGO), there was no clear systematic difference in performance, which was also the case using different feature representations (statistical features and PCA features). The MLP model was also the highest performing model in classification of incorrect repetitions; accuracy here was 70% compared to 59%, 58%, and 43% in the RFC, kNN, and SVM models, respectively.

This shows that an MLP model can classify correctly performed medio-lateral weight-shifts in >9 out of 10 repetitions, without using pre-determined rules or thresholds.

5.1.3 Research Question III

What accuracy can a recurrent neural network model achieve on estimation of force data, using kinematic data from 2D and 3D motion capture systems?

To answer this research question, joint center data from 3DMoCap and DV cameras were used as input to the estimation algorithms. The force (% bodyweight) data from the force plates was used as target data. The 3DMoCap data provides information in three dimensions (x, y, z), while the DV data only provides two (x and y, the depth dimension is not present here). Therefore, estimation of GRF in the depth dimension might be challenging when using DV data. Estimation was performed by an LSTM and XGBoost model, with a multivariate linear regression model as a baseline. Feature selection procedures revealed that when using 3DMoCap data there were eight joints that contributed with >82 % of the information in the GRF estimations, and these were kept in the analysis while the rest were discarded. For DV data there were also eight joints that contributed with the majority of the information, with >78%. Figures 5.3 and 5.4 detail the contributions of the different joint centers when using 3DMoCap and DV data, respectively. Using the selected joint centers, the three models were re-trained and estimations were again performed.

When using 3DMoCap data, the LSTM model performed at an excellent level with an average of >69%, 89%, and 98% explained variance (R^2) for Fx, Fy, and Fz, respectively, and RSME of <8%, 6% and 4% BW, respectively. The XGBoost and LinReg models achieved a slightly lower R^2, but also these achieved very good results in the Fz component with >85% R^2 and RMSE of <12% BW in both models.

The performance of all models decreased slightly when using DV data, although the LSTM results were still excellent. RMSE increased to 10%, 8%, and 8% in the LSTM model for Fx, Fy, and Fz components, respectively. R^2 decreased to 56% in Fx, 77% in Fy, and 92% in Fz for the LSTM model. In the XGBoost and LinReg models, R^2 in Fz decreased to 56% and 61%, 20%, and 19% RMSE, respectively.

These results indicate that ground reaction force can be reliably estimated from joint center kinematic data using a recurrent neural network (LSTM) during mediolateral weight shifts.

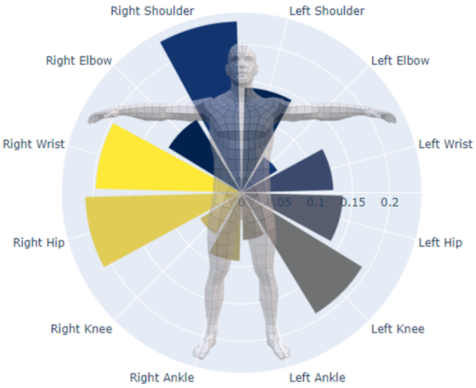


Figure 5.3: Results from Paper 3 - Force Estimation. Feature importance, or contributions, of different joint centers using MoCap data.

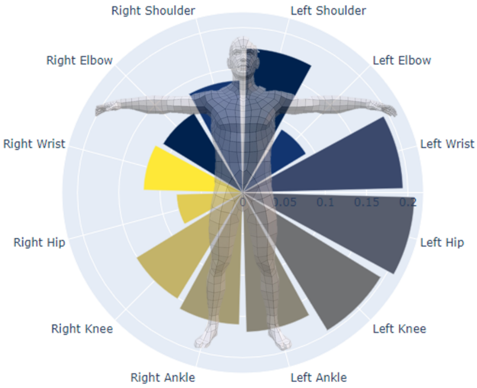


Figure 5.4: Results from Paper 3 - Force Estimation. Feature importance, or contributions, of different joint centers using DV data.

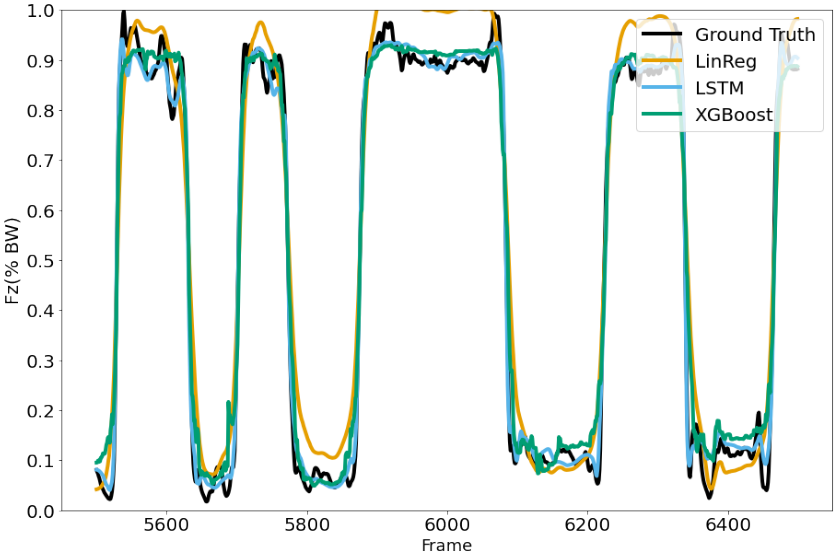


Figure 5.5: Results from Paper 3 - Force Estimation. Fz component estimation by the three models using 3DMoCap data.

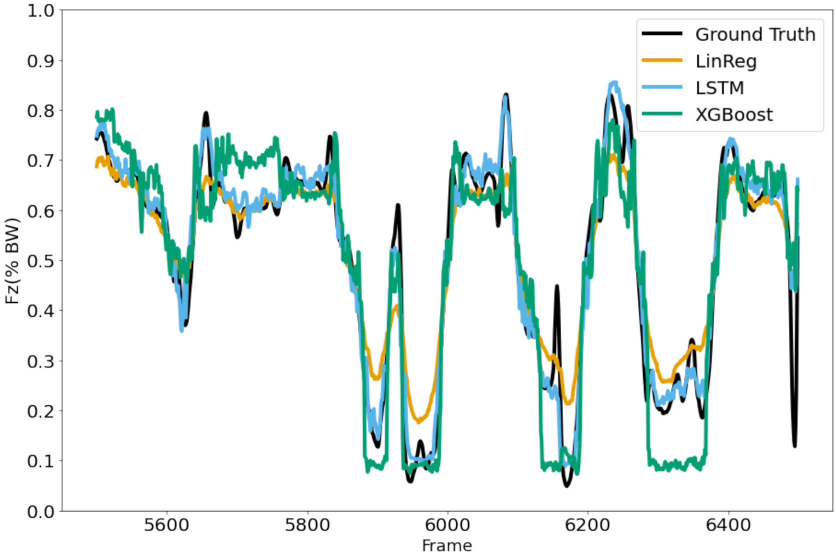


Figure 5.6: Results from Paper 3 - Force Estimation. Fz component estimation by the three models using DV data.

Chapter 6

Discussion

6.1 Overall discussion

The overall aim of this thesis was to investigate how existing solutions in exergame systems for weight-shifting exercise can be improved using state-of-the-art machine learning approaches. In the context of this thesis, *improvement* is specifically tied to the RQ of three core aspects of an exergaming system. For RQ1, improvement entailed providing motion capture data in a more accessible, yet accurate, manner using deep learning image analysis. For RQ2, improvement consisted of assessing movement patterns based on a self-learning approach of what a correct weight-shifting movement pattern is, rather than rules and templates. For RQ3, we sought to improve the usability of in-home exergames for balance training by enabling the generation of feedback on a data modality not previously available without specialized equipment.

As can be discerned from the results summarized in Chapter 5, using machine learning approaches suitable for these three endeavors can indeed improve exergame systems for balance exercise. Although similar research has been conducted previously in other areas, this does not guarantee that the machine learning models will perform equally well in the context of balance training exergames, as the No Free Lunch Theorem suggests [55]. However, the results of the studies in this thesis are encouraging, as we were able to demonstrate that our approaches reach similar levels of accuracy as adjacent research in their respective studies.

Even though our results are promising, it is important to keep in mind that realizing in-home use of exergames as a tool for physical exercise still has a long way to go. As with implementing any technology in the health domain, adoption of exergames is tedious work that requires attention and effort from both researchers, developers, the healthcare system, and the users themselves. One of the most essential aspects is to demonstrate to users that exergames are, in fact, useful for serious business and not suited for entertainment only. This is why improving technological tools is important. Bad experiences such as difficulties in setting up a system to make it work, or not working properly once it has been set up, can quickly lead to it being put away and not being recommended for use. For those designing and creating such systems, it is therefore imperative to be conscious of the choices being made for the game [140, 141, 118] and the supporting technology.

The relationship between study II and study III shows that there are several ways to use AI to provide feedback during exergaming. Performance of the entire weight-shifting repetition can be evaluated for being correct or not, or the weight-shifting performance can be provided as a real-time stream of information. This warrants careful consideration of what the most useful and meaningful feedback is for a particular person when taking exergames into use, especially as older adults are a heterogeneous group where individual levels of physical and cognitive capacity will influence what is most suitable. Furthermore, it might be useful to provide both types of feedback in the same setting. Seeing your own weight-shifting pattern in real-time, and then being rewarded rightfully based on that performance can be very helpful for improving a person's understanding of what a correct movement is - and guiding them towards making it. This can make people more aware of their own movement pattern as it can provide information during all phases of the movement and then reinforce or correct that movement pattern using an appropriate reward system after each repetition is performed.

Enabling and facilitating the use of exergames in in-home settings is a crucial focal point in the work towards healthy and active aging. Other endeavors aiming at developing methods of providing information about movement patterns, and methods of assessing movement patterns, during balance-training exergaming contribute as well towards understanding the most useful and appropriate setup for exergames for older adults. Because this setting is complex, all innovative approaches can and should add valuable tools to a larger toolbox that needs to be diverse, flexible, and adaptive to be able to accommodate the various needs, preferences, and body types of different users.

6.2 Key Findings and Contributions

The findings from the three studies in this thesis represent promising pieces of the puzzle being laid towards making exergames for balance exercise more accessible and easy to use, both in and outside clinical settings. This is the core contribution of this thesis, providing evidence that it is possible to use state-of-the-art analysis methods for technological solutions that make exergames more available and useful for in-home use.

In detail, the key findings were as follows: 1) A ResNet-based deep learning image analysis system, DeepLabCut, is a viable option for use in in-home exergame settings that depend upon low body segment length variability. 2) An MLP and an RFC model can classify correctly performed medio-lateral weight-shifts in >9 out of 10 repetitions, without comparing to pre-determined rules or templates. 3) Ground reaction force can be estimated reliably from joint center kinematic data from both 3D and 2D video using a recurrent neural network (LSTM) during medio-lateral weight shifts.

Furthermore, the three studies each contribute with their own insights into using machine learning methods in different facets of exergaming. Study I contributes to the critical first step of capturing player movements accurately with a single digital video camera. Using a state-of-the-art, pre-trained image analysis tool and a relatively small training data set ($\tilde{2}00$ images), player movement patterns can be extracted automatically. Existing solutions necessitate specialized equipment for motion capture to be available to the exergame. Even though cameras using depth-sensing technology are specifically designed to capture movements of the player, they are overall not more accurate than our proposed solution, as the results from Study I show. This is a key step in making exergames accessible for users that are not technologically advanced, as it enables the use of exergames without having to obtain and use specialized equipment such as depth sensing cameras or IMU sensors. It is generally difficult to compare our findings in Study I to that of previous pose estimation research that was performed in other domains and contexts. Our work is specifically oriented towards the application of a pose estimation framework within active, healthy aging, and is thus concerned with the practical side of the performance of a framework - i.e., the segment length variability. In contrast, earlier work on pose estimation algorithms was largely concerned with the evaluation of new or adjusted pose estimation algorithms per se, and therefore evaluated performance from a more technical point of view. One example that illustrates why our study is difficult to compare to the existing literature is the use of percentage correct key points (PCK) within a certain distance from the actual target key point (i.e., joint center) location. Using a threshold of the predicted location being within 50% or even 10% of a segment length can lead to estimated joint center locations many centimeters from the actual, real-world joint center. In contrast, our approach is novel in that it investigates the segment length variability calculated in each system, which is a more useful measure as it shows how well a motion capture system performs over time.

Study II is an investigation into which classification models are most suited for identifying correctly performed medio-lateral weight-shifting movements. Here, we document the superior ability of a neural network and random forest model over other models in this task and concluded that the random forest model should be preferred because of the faster prediction time and more transparent decision making process. Our approach here differs from previous, template-based assessment methods in that it does not use pre-determined goals for the movement pattern that is more or less based on domain knowledge of the movement performed. Our self-learning method transfers this task to the machine learning models instead, as the training data consists of features from joint center locations during one repetition of medio-lateral weight-shifting, and the corresponding label (correct or incorrect). The random forest classifier and the MLP classifier then are able to distinguish a correctly performed repetition from an incorrectly performed repetition in as much as >9 out of 10 times. This study also contributes with crucial information needed for exergame developers when choices are to be made when implementing machine learning models for classification of movement pattern correctness.

In study III, a novel and highly valuable way of extracting real-time force data while performing medio-lateral weight-shifts have been developed. In in-home settings, this information was previously unavailable due to the lack of suitable equipment for accurately measuring ground reaction force (i.e., how much weight is being put on each foot). Our approach shows that it is possible to use both 3D and 2D data to obtain GRF information, enabling more informative and useful feedback to the player about their weight-shifting performance. This study also underscores that the quality of motion capture data is important. Data with more noise makes it harder for ML methods to discern the true relationship between the input data and the target data. Furthermore, our approach uses joint center positions directly as input to the LSTM model, which avoids the computationally expensive use of biomechanical models for feature engineering as used in previous research on force estimation. Hence, all three studies can affect the choices made in the further development of exergames. We have illustrated that by using state-of-the-art machine learning models, it is possible to make movement analysis tasks simpler by skipping computational layers where biomechanical models, rules, and templates are involved. For independent, in-home exergaming, these are crucial improvements as they can make exergame systems more accessible, less resource demanding, and easier to use. These are features that enable older adults to take agency over their own level of physical function and implement exercise regimes from the comfort of their home, thereby contributing to preventing falls and other detrimental effects of poor physical function.

In addition to the scientific findings from the papers, the work conducted during these four years has produced know-how and methodological insights into implementing machine learning into human movement science. Furthermore, a comprehensive data set has been curated, which could be of use to other researchers in the field. This will be published in an open-source repository.

6.3 Methodological Considerations

The results and main conclusions in this thesis should be viewed in light of some considerations around the methodological approaches employed. Firstly, the use of a high-end motion capture laboratory has both strengths and drawbacks. Using gold-standard motion capture equipment (3DMoCap) provides high-quality data about the movement patterns performed in the studies. This contributes to the correct interpretation of results, as model performance was not significantly influenced by poor data quality. Furthermore, using a laboratory setting allows us to control the surroundings of the experiments, but this has low ecological similarity to the setting where exergaming might be performed in a real-world setting.

The data collection provided information from twelve healthy older adults. This is not a large data set, which could affect the generalizability of the data as the participants only represent a small part of the heterogeneous demography of the older adult population. Furthermore, all participants exercised regularly and were in relatively good health, which might also have influenced the variability of the movement patterns in the data-set. Investigation of the performance of the current methods and models using data from persons with frailty or limited physical function is an interesting future avenue to pursue further, as their movement patterns can be expected to be slower and more variable than those in the current data set.

One drawback that limits the generalizability of our results in this thesis is the missing implementation and testing in a real-world setting. Even though the results show encouraging results, this is still offline testing only. In a real-time implementation, there are other considerations in addition to the classification or estimation performance, such as latency in predictions and thus delays in feedback, which is an important side of machine learning implementation in general [142, 143].

In terms of these methods being transferable to a real-world setting, one major issue can be curating a data set for training models before implementing them in an exergame system. The data collection itself can be time and resource demanding, and post-processing of motion capture data to prepare them for machine learning models is also a time-consuming job. Nevertheless, this can be mediated by using frameworks that allow for using pre-trained data sets in combination with a specialized training layer, with video data from the context-specific setting in which the exergame is going to be implemented.

6.4 Implications

This thesis shows that methods for improving exergame solutions are available, in that there might not be a need for specialized equipment to be able to track, assess and measure important data for balance training. This implies that there is a possibility to realize the potential of balance exergames that research has documented, as the possible barriers of use relating to technological solutions can be avoided. Health care professionals should look beyond the solutions created for the casual gaming consumer market for more appropriate and specialized exergames, as exergames can be used to automatically supervise exercise and provide feedback to the player. The possibility to tailor the exergame tasks to the person playing is also an important feature, making it possible to target specific functions to train.

Game developers should also look towards developing exergame systems that do not require additional equipment to be purchased and maintained by the player or institution. Developing exergames going forward should focus on ease of use and availability of games from a technical perspective, by taking advantage of the possibilities within machine learning.

6.5 Future Work

To continue on the path towards implementation of balance-training exergames, and exergames in general, the next steps should involve including machine learning models when developing exergames. Only by testing these solutions in a real-world setting will we be able to know how well the different pieces work. This is both regarding the technical side of it, i.e., whether the machine learning models are fast enough in real-time predictions and estimations, and regarding whether the users actually find the systems usable and useful. It would also be interesting to see whether it is possible to provide more fine-grained feedback on movement pattern correctness using classification models like the ones used in study II. The most challenging side of this is that it requires a large data set with different movement pattern errors being naturally elicited, which would be challenging and time-consuming to procure. Furthermore, the effectiveness of exergames using different types of motion capture technologies and algorithms should be a focus of research going forward, to build a stronger knowledge foundation that can enable informed decisions when designing and implementing exergames. Especially in-home, independent use should be explored, as this is where the major potential of exergames lies.

Chapter 7

Conclusions

Considering the potential of exergames to be a facilitator of motivating and efficient in-home exercise for the aging population, this thesis aimed at assessing new methods of capturing, assessing, and enabling use of information in the motion capture aspect of balance training exergames. This can enable the use of more available and easy-to-use technologies, such as smartphone cameras or web-cameras, in an in-home setting. To this end, state-of-the art machine learning models were employed in image analysis, classification of movement patterns, and estimation of force data from kinematic information. The results provide comprehensive documentation that different machine learning models perform at a high level in these tasks, which is encouraging with regard to future development and implementation of balance training exergames.

In conclusion, there are easy-to-use technological improvements available through machine learning frameworks that should be considered for implementation when an exergame system is created. These help to ensure that the quality of the systems are at the level that users need and expect when using them in serious settings, which in this setting means low levels of jittering and erroneous motion tracking, appropriate reward systems that facilitate efficiency and motivation, and accurate feedback on weight-shifting performance in real-time without specialized equipment.

Bibliography

[1] Jennifer Taylor, Sarah Walsh, Wing Kwok, Marina B Pinheiro, Juliana Souza De Oliveira, Leanne Hassett, Adrian Bauman, Fiona Bull, Anne Tiedemann, and Catherine Sherrington. A scoping review of physical activity interventions for older adults. pages 1–15, 2021.

[2] Laurence Z. Rubenstein. Falls in older people: Epidemiology, risk factors and strategies for prevention. *Age and Ageing*, 35(SUPPL.2):37–41, 2006.

[3] Nancye May Peel. Epidemiology of falls in older age. *Canadian Journal on Aging*, 30(1):7–19, 3 2011.

[4] L D Gillespie, M C Robertson, W J Gillespie, C Sherrington, S Gates, L M Clemson, and S E Lamb. Interventions for preventing falls in older people living in the community. *Cochrane Database Syst Rev*, 9(9):Cd007146, 2012.

[5] Sebastian Deterding. Gamification: designing for motivation. *Interactions*, 19(4):14–17, 7 2012.

[6] Nina Skjæret, Ather Nawaz, Kristine Ystmark, Yngve Dahl, Jorunn L. Helbostad, Dag Svanæs, and Beatrix Vereijken. Designing for movement quality in exergames: Lessons learned from observing senior citizens playing stepping games. *Gerontology*, 61(2):186–194, 2015.

[7] Elizabeth J. Lyons. Cultivating Engagement and Enjoyment in Exergames Using Feedback, Challenge, and Rewards. In *Games for Health Journal*, volume 4, pages 12–18. Mary Ann Liebert Inc., 2 2015.

[8] Katie Jane Brickwood, Greig Watson, Jane O'brien, and Andrew D Williams. Consumer-based wearable activity trackers increase physical activity participation: Systematic review and meta-analysis, 4 2019.

[9] Nina Skjæret, Ather Nawaz, Tobias Morat, Daniel Schoene, Jorunn Lægdheim, and Beatrix Vereijken. Exercise and rehabilitation delivered through exergames in older adults : An integrative review of technologies, safety and efficacy. *International Journal of Medical Informatics*, 85(1):1–16, 2016.

[10] Joep Janssen, Olaf Verschuren, Willem Jan Renger, Jose Ermers, Marjolijn Ketelaar, and Raymond Van Ee. Gamification in physical therapy: More than using games. *Pediatric Physical Therapy*, 29(1):95–99, 2017.

[11] Jan David Smeddinck, Marc Herrlich, and Rainer Malaka. Exergames for Physiotherapy and Rehabilitation: A Medium-term Situated Study of Motivational Aspects and Impact on Functional Reach. *Proceedings of the ACM CHI'15 Conference on Human Factors in Computing Systems*, 1:4143–4146, 2015.

[12] Ather Nawaz, Nina Skjaeret, Jorunn Laegdheim Helbostad, Beatrix Vereijken, Elisabeth Boulton, and Dag Svanaes. Usability and acceptability of balance exergames in older adults: A scoping review. *Health informatics journal*, 22(4):911–931, 12 2016.

[13] Damla Kiziltas and Ufuk Celikcan. Knee Up:an Exercise Game for Standing Knee Raises by Motion Capture with RGB-D Sensor. In *Smart Tools and Applications in Graphics*, 2018.

[14] Halim Tannous, Dan Istrate, Aziz Benlarbi-Delai, Julien Sarrazin, Didier Gamet, Marie Christine Ho Ba Tho, and Tien Tuan Dao. A new multi-sensor fusion scheme to improve the accuracy of knee flexion kinematics for functional rehabilitation movements. *Sensors (Switzerland)*, 16(11), 2016.

[15] KE Laver, B Lange, S George, JE Deutsch, G Saposnik, and M Crotty. Virtual reality for stroke rehabilitation. *Cochrane Database of Systematic Reviews*, (11):CD008349, 2017.

[16] Lufang Zheng, Guichen Li, Xinxin Wang, Huiru Yin, Yong Jia, Minmin Leng, Hongyan Li, and Li Chen. Effect of exergames on physical outcomes in frail elderly: a systematic review. *Aging Clinical and Experimental Research 2019 32:11*, 32(11):2187–2200, 9 2019.

[17] Maziah Mat Rosly, Hadi Mat Rosly, Glen M. Davis OAM, Ruby Husain, and Nazirah Hasnan. Exergaming for individuals with neurological disability: a systematic review. *https://doi.org/10.3109/09638288.2016.1161086*, 39(8):727–735, 4 2016.

[18] Mike van Diest, Claudine CJC Lamoth, Jan Stegenga, Gijsbertus J Verkerke, and Klaas Postema. Exergaming for balance training of elderly: state of the art and future developments. *Journal of NeuroEngineering and Rehabilitation*, 10(1):101, 2013.

[19] Seline Wüest, Nunzio Alberto Borghese, Michele Pirovano, Renato Mainetti, Rolf van de Langenberg, and Eling D. de Bruin. Usability and Effects of an Exergame-Based Balance Training Program. *Games for Health Journal*, 3(2):106–114, 2014.

[20] Emma K. Stanmore, Alexandra Mavroeidi, Lex D. De Jong, Dawn A. Skelton, Chris J. Sutton, Valerio Benedetto, Luke A. Munford, Wytske Meekes, Vicky Bell, and Chris Todd. The effectiveness and cost-effectiveness of strength and balance Exergames to reduce falls risk for people aged 55 years and older in UK assisted living facilities: A multi-centre, cluster randomised controlled trial. *BMC Medicine*, 17(1):1–14, 2 2019.

[21] RPS Van Peppen, G Kwakkel, S Wood-Dauphinee, HJM Hendriks, PhJ Var der Wees, and J Dekker. The impact of physical therapy on functional outcomes after stroke : what ' s the evidence ? *Clinical Rehabilitation*, 18:833–862, 2004.

[22] Catherine E. Lang, Jillian R. MacDonald, Darcy S. Reisman, Lara Boyd, Teresa Jacobson Kimberley, Sheila M. Schindler-Ivens, T. George Hornby, Sandy A. Ross, and Patricia L. Scheets. Observation of Amounts of Movement Practice Provided During Stroke Rehabilitation. *Archives of Physical Medicine and Rehabilitation*, 90(10):1692–1698, 2009.

[23] Christine C. Chen and Rita K. Bode. Factors influencing therapists' decision-making in the acceptance of new technology devices in stroke rehabilitation. *American Journal of Physical Medicine and Rehabilitation*, 90(5):415–425, 2011.

[24] Ai Vi Nguyen, Yau Lok Austin Ong, Cindy Xin Luo, Thiviya Thuraisingam, Michael Rubino, Mindy F. Levin, Franceen Kaizer, and Philippe S. Archambault. Virtual reality exergaming as adjunctive therapy in a sub-acute stroke rehabilitation setting: facilitators and barriers. *Disability and Rehabilitation: Assistive Technology*, 14(4):317–324, 2019.

[25] Stepan Obdrzalek, Gregorij Kurillo, Ferda Ofli, Ruzena Bajcsy, Edmund Seto, Holly Jimison, and Michael Pavel. Accuracy and robustness of Kinect pose estimation in the context of coaching of elderly population. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 1188–1193, 2012.

[26] Catherine Sherrington, Nicola J. Fairhall, Geraldine K. Wallbank, Anne Tiedemann, Zoe A. Michaleff, Kirsten Howard, Lindy Clemson, Sally Hopewell, and Sarah E. Lamb. Exercise for preventing falls in older people living in the community. *Cochrane Database of Systematic Reviews*, 2019(1), 2019.

[27] John R. Beard, A. T. Jotheeswaran, Matteo Cesari, and Islene Araujo De Carvalho. The structure and predictive value of intrinsic capacity in a longitudinal study of ageing. *BMJ Open*, 9(11), 11 2019.

[28] Mark W. Rogers, Marjorie E. Johnson, Kathy M. Martinez, Marie Laure Mille, and Lois D. Hedman. Step training improves the speed of voluntary step initiation in aging. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, 58(1):46–51, 2003.

[29] Michelle M. Lusardi, Geraldine L. Pellecchia, and Marjorie Schulman. Functional Performance in Community Living Older Adults. *Journal of Geriatric Physical Therapy*, 26(3):14–22, 2003.

[30] Marjorie Woollacott and Anne Shumway-Cook. Attention and the control of posture and gait: A review of an emerging area of research. *Gait and Posture*, 16(1):1–14, 2002.

[31] Chih Hsuan Chou, Chueh Lung Hwang, and Ying Tai Wu. Effect of exercise on physical function, daily living activities, and quality of life in the frail older adults: A meta-analysis. *Archives of Physical Medicine and Rehabilitation*, 93(2):237–244, 2012.

[32] Michael D. Denkinger, Albert Lukas, Thorsten Nikolaus, and Klaus Hauer. Factors associated with fear of falling and associated activity restriction in community-dwelling older adults: A systematic review. *American Journal of Geriatric Psychiatry*, 23(1):72–86, 2015.

[33] United Nations Department of Economic and Social Affairs Population Division. World Population Prospects 2019: Highlights (ST/ESA/SER.A/423). Technical report, United Nations Department of Economic and Social Affairs Population Division, New York, 2019.

[34] Qun Fang, Parisa Ghanouni, Sarah E. Anderson, Hilary Touchett, Rebekah Shirley, Fang Fang, and Chao Fang. Effects of Exergaming on Balance of Healthy Older Adults: A Systematic Review and Meta-analysis of Randomized Controlled Trials. *Games for Health Journal*, 9(1):11–23, 2020.

[35] Martin G. Jorgensen, Uffe Laessoe, Carsten Hendriksen, Ole Bruno Faurholt Nielsen, and Per Aagaard. Efficacy of nintendo wii training on mechanical leg muscle function and postural balance in community-dwelling older adults: A randomized controlled trial. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, 68(7):845–852, 2013.

[36] Yoshiro Okubo, Daniel Schoene, and Stephen R Lord. Step training improves reaction time, gait and balance and reduces falls in older people: a systematic review and meta-analysis. *British Journal of Sports Medicine*, 2016.

[37] Jeffrey M. Hausdorff, Dean A. Rios, and Helen K. Edelberg. Gait variability and fall risk in community-living older adults: A 1-year prospective study. *Archives of Physical Medicine and Rehabilitation*, 82(8):1050–1056, 2001.

[38] John R Beard and David E Bloom. Towards a comprehensive public health response to population ageing. *The Lancet*, 385(9968):658–661, 2 2015.

[39] Jorunn L. Helbostad, Beatrix Vereijken, Clemens Becker, Christop Todd, Kristin Taraldsen, Mirjam Pijnappels, Kamiar Aminian, and Sabato Mellone. Mobile health applications to promote active and healthy ageing. *Sensors (Switzerland)*, 17(3):1–13, 2017.

[40] Marie Chan, Daniel Estève, Jean Yves Fourniols, Christophe Escriba, and Eric Campo. Smart wearable systems: Current status and future challenges. *Artificial Intelligence in Medicine*, 56(3):137–156, 2012.

[41] S. Singhal and S. Carlton. The era of exponential improvement in healthcare? *McKinsey & Company Review*, May, 2019.

[42] Dereli E.E. and Yaliman A. Comparison of the effects of a physiotherapist-supervised exercise programme and a self-supervised exercise programme on quality of life in patients with Parkinson's disease. *Clinical rehabilitation*, 24(4):352–362, 2010.

[43] C. M. Woodard and M. J. Berry. Enhancing adherence to prescribed exercise: Structured behavioral interventions in clinical exercise programs. *Journal of Cardiopulmonary Rehabilitation*, 21(4):201–209, 2001.

[44] Agnes Zijlstra, Martina Mancini, Lorenzo Chiari, and Wiebren Zijlstra. Biofeedback for training balance and mobility tasks in older populations: A systematic review. *Journal of NeuroEngineering and Rehabilitation*, 7(1):1–15, 2010.

[45] Aseel Berglund, Erik Berglund, Fabio Siliberto, and Erik Prytz. Effects of reactive and strategic game mechanics in motion-based games. *2017 IEEE 5th International Conference on Serious Games and Applications for Health, SeGAH 2017*, 2017.

[46] Sruti Subramanian, Yngve Dahl, Nina Skjaret Maroni, Beatrix Vereijken, and Dag Svanas. Twelve Ways to Reach for a Star: Player Movement Strategies in a Whole-Body Exergame. *2019 IEEE 7th International Conference on Serious Games and Applications for Health, SeGAH 2019*, (August), 2019.

[47] Qicheng Ding, Ian H. Stevenson, Ninghua Wang, Wei Li, Yao Sun, Qining Wang, Konrad Kording, and Kunlin Wei. Motion games improve balance control in stroke survivors: A preliminary study based on the principle of constraint-induced movement therapy. *Displays*, 34(2):125–131, 4 2013.

[48] R. A. Geiger, J. B. Allen, J. O'Keefe, and R. R. Hicks. Balance and mobility following stroke: Effects of physical therapy interventions with and without biofeedback/forceplate training. *Physical Therapy*, 81(4):995–1005, 2001.

[49] Vassilia Hatzitaki, Ioannis G. Amiridis, Thomas Nikodelis, and Styliani Spiliopoulou. Direction-Induced Effects of Visually Guided Weight-Shifting Training on Standing Balance in the Elderly. *Gerontology*, 55(2):145–152, 3 2009.

[50] Deepesh Kumar, Nirvik Sinha, Anirban Dutta, and Uttama Lahiri. Virtual reality-based balance training system augmented with operant conditioning paradigm. *BioMedical Engineering OnLine 2019 18:1*, 18(1):1–23, 8 2019.

[51] Julia M. Leach, Martina Mancini, Robert J. Peterka, Tamara L. Hayes, and Fay B. Horak. Validating and calibrating the Nintendo Wii balance board to derive reliable center of pressure measures. *Sensors (Switzerland)*, 14(10):18244–18267, 9 2014.

[52] Ross A. Clark, Adam L. Bryant, Yonghao Pua, Paul McCrory, Kim Bennell, and Michael Hunt. Validity and reliability of the Nintendo Wii Balance Board for assessment of standing balance. *Gait and Posture*, 31(3):307–310, 3 2010.

[53] Harrison L. Bartlett, Lena H. Ting, and Jeffrey T. Bingham. Accuracy of force and center of pressure measures of the Wii Balance Board. *Gait and Posture*, 39(1):224–228, 1 2014.

[54] Kelly J. Bower, Ross A. Clark, Jennifer L. McGinley, Clarissa L. Martin, and Kimberly J. Miller. Clinical feasibility of the Nintendo Wii™ for balance training post-stroke: A phase II randomized controlled trial in an inpatient setting. *Clinical Rehabilitation*, 28(9):912–923, 2014.

[55] Tom M. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[56] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 103 of *Springer Texts in Statistics*. Springer New York, New York, NY, 2013.

[57] Xindong Wu, Xingquan Zhu, Gong Qing Wu, and Wei Ding. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 26(1):97–107, 1 2014.

[58] Stephen Marshland. *Machine Learning - An Algorithmic Perspective*. Taylor {\&} Francis Group, 2nd edition, 2015.

[59] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[60] Jackie Ayoub, Feng Zhou, Shan Bao, and X. Jessie Yang. From Manual Driving to Automated Driving. In *AutomotiveUI'19*, pages 70–90, 2019.

[61] Manohar Mishra, Janmenjoy Nayak, Bighnaraj Naik, and Ajith Abraham. Deep learning in electrical utility industry: A comprehensive review of a decade of research. *Engineering Applications of Artificial Intelligence*, 96(August):104000, 2020.

[62] Sayyeda Saadia Razvi, Shaw Feng, Anantha Narayanan, Yung Tsun Tina Lee, and Paul Witherell. A review of machine learning applications in additive manufacturing. *Proceedings of the ASME Design Engineering Technical Conference*, 1:1–10, 2019.

[63] Tahereh Pourhabibi, Kok Leong Ong, Booi H. Kam, and Yee Ling Boo. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, 133(April):113303, 2020.

[64] Jian Wang, Hengde Zhu, Shui Hua Wang, and Yu Dong Zhang. A Review of Deep Learning on Medical Image Analysis. *Mobile Networks and Applications*, 26(1):351–380, 2021.

[65] Nathan LaPierre, Chelsea J.T. Ju, Guangyu Zhou, and Wei Wang. MetaPheno: A critical evaluation of deep learning and machine learning in metagenome-based disease prediction. *Methods*, 166(February):74–82, 2019.

[66] Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428, 2018.

[67] Yucheng Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, 192(January):102897, 2020.

[68] Rafael Caldas, Tariq Fadel, Fernando Buarque, and Bernd Markert. Adaptive predictive systems applied to gait analysis: A systematic review. *Gait and Posture*, 77(January):75–82, 2020.

[69] Delaram Jarchi, James Pope, Tracey K.M. Lee, Larisa Tamjidi, Amirhosein Mirzaei, and Saeid Sanei. A Review on Accelerometry-Based Gait Analysis and Emerging Clinical Applications. *IEEE Reviews in Biomedical Engineering*, 11:177–194, 2018.

[70] B. Dumphart, D. Slijepčević, F. Unglaube, A. Kranzl, A. Baca, M. Zeppelzauer, and B. Horsak. An automated deep learning-based gait event detection algorithm for various pathologies. *Gait & Posture*, 90:50–51, 2021.

[71] Damith Senanayake, Saman Halgamuge, and David C. Ackland. Real-time conversion of inertial measurement unit data to ankle joint angles using deep neural networks. *Journal of Biomechanics*, 125(June):110552, 2021.

[72] William S. Burton, Casey A. Myers, and Paul J. Rullkoetter. Machine learning for rapid estimation of lower extremity muscle and joint loading during activities of daily living. *Journal of Biomechanics*, 123:110439, 2021.

[73] L. Minh Dang, Kyungbok Min, Hanxiang Wang, Md Jalil Piran, Cheol Hee Lee, and Hyeonjoon Moon. Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition*, 108, 2020.

[74] Espen A. F. Ihlen, Ragnhild Støen, Lynn Boswell, Raye-Ann de Regnier, Toril Fjørtoft, Deborah Gaebler-Spira, Cathrine Labori, Marianne C. Loennecken, Michael E. Msall, Unn I. Möinichen, Colleen Peyton, Michael D. Schreiber, Inger E. Silberg, Nils T. Songstad, Randi T. Vågen, Gunn K. Øberg, and Lars Adde. Machine Learning of Infant Spontaneous Movements for the Early Prediction of Cerebral Palsy: A Multi-Site Cohort Study. *Journal of Clinical Medicine*, 9(1):5, 2019.

[75] Emily E. Cust, Alice J. Sweeting, Kevin Ball, and Sam Robertson. Machine and deep learning for sport-specific movement recognition: a systematic review of model development and performance. *Journal of Sports Sciences*, 37(5):568–600, 2019.

[76] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[77] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep Sparse Rectifier Neural Networks. In *Proceedings of the 14th International Con- ference on Artificial Intelligence and Statistics (AISTATS)*, pages 315–323, 2011.

[78] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. Technical report.

[79] J Deng, W Dong, R Socher, L.-J. Li, K Li, and L Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[80] Tsung-Yi Lin, Michael Maire, Serge J Belongie, Lubomir D Bourdev, Ross B Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft {COCO:} Common Objects in Context. *CoRR*, 2014.

[81] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. Technical report.

[82] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. Technical report.

[83] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:770–778, 2016.

[84] J Deng, W Dong, R Socher, L.-J. Li, K Li, and L Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[85] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 249–256, 2010.

[86] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model. In *European Conference on Computer Vision*, pages 34–50, 5 2016.

[87] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. DeepLabCut: Markerless Pose Estimation of User-Defined Body Parts with Deep Learning. *Nature Neuroscience*, 21(September), 2018.

[88] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. Towards Understanding Action Recognition. In *International Conference on computer vision*, pages 3192–3199, 2013.

[89] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. HumanEva : Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. *International Journal of Computer Vision 2009 87:1*, 87(1):4–27, 8 2009.

[90] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 2014.

[91] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. TOTAL CAPTURE: POSE ESTIMATION FUSING VIDEO AND IMU DATA Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors.

[92] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 1 2021.

[93] Daniel Groos, Heri Ramampiaro, and Espen A.F. Ihlen. EfficientPose: Scalable single-person pose estimation. *Applied Intelligence*, 51:2518–2533, 2021.

[94] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral Human Pose Regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–545, 2018.

[95] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting Temporal Context for 3D Human Pose Estimation in the Wild. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3395–3404, 2019.

[96] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seide, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. XNect. *ACM Transactions on Graphics (TOG)*, 39(4):17, 7 2020.

[97] B. Bonnechère, B. Jansen, P. Salvia, H. Bouzahouene, L. Omelina, F. Moiseev, V. Sholukha, J. Cornelis, M. Rooze, and S. Van Sint Jan. Validity and reliability of the Kinect within functional assessment activities: Comparison with standard stereophotogrammetry. *Gait and Posture*, 39(1):593–598, 2014.

[98] Andrea Bravi, André Longtin, and Andrew J.E. Seely. Review and classification of variability analysis techniques with clinical applications. *BioMedical Engineering Online*, 10(90), 10 2011.

[99] Thomas G. Dietterich. Ensemble Methods in Machine Learning. *Multiple Classifier Systems*, 1857:1–15, 2000.

[100] Leo Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001.

[101] Robert-Andrei Voicu, Ciprian Dobre, Lidia Bajenaru, and Radu-Ioan Ciobanu. Human Physical Activity Recognition Using Smartphone Sensors. *Sensors (Basel, Switzerland)*, 19(3), 2 2019.

[102] Allah Bux, Plamen Angelov, and Zulfiqar Habib. Vision based human activity recognition: A review. *Advances in Intelligent Systems and Computing*, 513:341–371, 2017.

[103] Shian-Ru Ke, Hoang Le Uyen Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi. A Review on Video-Based Human Activity Recognition. *Computers*, 2:88–131, 6 2013.

[104] Michalis Vrigkas, Christophoros Nikou, and Ioannis A. Kakadiaris. A Review of Human Activity Recognition Methods. *Frontiers in Robotics and AI*, 0(NOV):28, 2015.

[105] Antonio A. Aguileta, Ramon F. Brena, Oscar Mayora, Erik Molino-Minero-Re, and Luis A. Trejo. Multi-Sensor Fusion for Activity Recognition—A Survey. *Sensors*, 19(17), 9 2019.

[106] Colin Lea, Michael D Flynn Ren, Austin Reiter, and Gregory D Hager. Temporal Convolutional Networks for Action Segmentation and Detection. In *CVPR*, pages 156–165, 2016.

[107] O.M. Giggins, I. Clay, and L. Walsh. Physical Activity Monitoring in Patients with Neurological Disorders: A Review of Novel Body-Worn Devices. *Digital biomarkers*, 1(1), 6 2017.

[108] Lisa McGarrigle and Chris Todd. Promotion of physical activity in older people using mHealth and eHealth technologies: Rapid review of reviews. *Disability and Rehabilitation*, 22(12), 2020.

[109] Behnoosh Parsa, Athma Narayanan, and Behzad Dariush. Spatio-temporal pyramid graph convolutions for human action recognition and postural assessment. *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, pages 1069–1079, 2020.

[110] Shugang Zhang, Zhiqiang Wei, Jie Nie, Lei Huang, Shuang Wang, and Zhen Li. A Review on Human Activity Recognition Using Vision-Based Method. *Journal of Healthcare Engineering*, 2017, 2017.

[111] Portia E Taylor, Gustavo J M Almeida, Jessica K Hodgins, and Takeo Kanade. Multi-label classification for the analysis of human motion quality. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012(August):2214–8, 2012.

[112] Michel Antunes, Renato Baptista, Girum Demisse, Djamila Aouada, and Björn Ottersten. Visual and human-interpretable feedback for assisting physical activity. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9914 LNCS:115–129, 2016.

[113] Emily E Cust, Alice J Sweeting, Kevin Ball, and Sam Robertson. Machine and deep learning for sport-specific movement recognition: a systematic review of model development and performance. *Journal of Sports Sciences*, 37(5):568–600, 3 2018.

[114] Portia E Taylor, Gustavo J M Almeida, Takeo Kanade, and Jessica K Hodgins. Classifying Human Motion Quality for Knee Osteoarthritis Using Accelerometers. *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 339–343, 2010.

[115] Wenbing Zhao, Ann M. Reinthal, Deborah D Espy, and Xiong Luo. Rule-Based Human Motion Tracking for Rehabilitation Exercises : Realtime Assessment , Feedback , and Guidance. *IEEE Access*, 5:21382–21394, 2017.

[116] Norbert Gal, Diana Andrei, Dan Ion Nemeş, Emanuela Nadasan, and Vasile Stoicu-Tivadar. A Kinect based intelligent e-rehabilitation system in physi-

cal therapy. In *Studies in Health Technology and Informatics*, volume 210, pages 489–493. IOS Press, 2015.

[117] Ferda Ofli, Gregorij Kurillo, Štěpán Obdržálek, Ruzena Bajcsy, Holly Brugge Jimison, and Misha Pavel. Design and evaluation of an interactive exercise coaching system for older adults: Lessons learned. *IEEE Journal of Biomedical and Health Informatics*, 20(1):201–212, 1 2016.

[118] Agnes W.K. Lam, Danniel Varona-Marin, Yeti Li, Mitchell Fergenbaum, and Dana Kulić. Automated Rehabilitation System: Movement Measurement and Feedback for Patients and Physiotherapists in the Rehabilitation Clinic. *Human-Computer Interaction*, 31(3-4):294–334, 2016.

[119] Lili Tao, Adeline Paiement, Dima Damen, Majid Mirmehdi, Sion Hannuna, Massimo Camplani, Tilo Burghardt, and Ian Craddock. A comparative study of pose representation and dynamics modelling for online motion quality assessment. *Computer Vision and Image Understanding*, 148:136–152, 2016.

[120] Klaus Greff, Rupesh K Srivastava, Jan Koutnik, Bas R Steunebrink, and Jurgen Schmidhuber. LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232, 2017.

[121] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15, 2015.

[122] David A. Winter. *Biomechanics and Motor Control of Human Movement: Fourth Edition*. JOHN WILEY & SONS, INC, 2009.

[123] R. Fluit, M. S. Andersen, S. Kolk, N. Verdonschot, and H. F.J.M. Koopman. Prediction of ground reaction forces and moments during various activities of daily living. *Journal of Biomechanics*, 47(10):2321–2329, 2014.

[124] Angelos Karatsidis, Giovanni Bellusci, H. Schepers, Mark de Zee, Michael Andersen, and Peter Veltink. Estimation of Ground Reaction Forces and Moments During Gait Using Only Inertial Motion Capture. *Sensors*, 17(12):75, 12 2016.

[125] Lei Ren, Richard K. Jones, and David Howard. Whole body inverse dynamics over a complete gait cycle based only on measured kinematics. *Journal of Biomechanics*, 41(12):2750–2759, 8 2008.

[126] M. Hayashibe, A. González, and P. Fraisse. Personalized Modeling for Home-Based Postural Balance Rehabilitation. In *Human Modeling for Bio-Inspired Robotics: Mechanical Engineering in Assistive Technologies*, pages 111–137. Elsevier Inc., 1 2017.

[127] Gustavo Leporace, Luiz Alberto Batista, Leonardo Metsavaht, and Jurandir Nadal. Residual analysis of ground reaction forces simulation during gait using neural networks with different configurations. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2015-Novem:2812–2815, 2015.

[128] Sean T. Osis, Blayne A. Hettinga, and Reed Ferber. Predicting ground contact events for a continuum of gait types: An application of targeted machine learning using principal component analysis. *Gait and Posture*, 46:86–90, 2016.

[129] Ahmed J. Aljaaf, Abir J. Hussain, Paul Fergus, Andrzej Przybyla, and Gabor J. Barton. Evaluation of machine learning methods to predict knee loading from the movement of body segments. *Proceedings of the International Joint Conference on Neural Networks*, 2016-Octob:5168–5173, 2016.

[130] Marion Mundt, Arnd Koeppe, Sina David, Franz Bamer, Wolfgang Potthast, and Bernd Markert. Prediction of ground reaction force and joint moments based on optical motion capture data during gait. *Medical Engineering and Physics*, 86:29–34, 12 2020.

[131] Seung Eel Oh, Ahnryul Choi, and Joung Hwan Mun. Prediction of ground reaction forces during gait based on kinematics and a neural network model. *Journal of Biomechanics*, 46(14):2372–2380, 9 2013.

[132] Ahnryul Choi, Jae Moon Lee, and Joung Hwan Mun. Ground reaction forces predicted by using artificial neural network during asymmetric movements. *International Journal of Precision Engineering and Manufacturing*, 14(3):475–483, 2013.

[133] Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 17(1):168–192, 2018.

[134] Camilla Velvin. Exergame for rehabilitering av balansen til slagpasienter. Technical report, Norwegian University of Science and Technology, 2019.

[135] Vicon Motion Systems Limited. Plug-in Gait Reference Guide. Technical report, 2016.

[136] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2012.

[137] Francois Chollet. Keras, 2015.

[138] Maximilian Christ, Andreas W. Kempa-Liehr, and Michael Feindt. Distributed and parallel time series feature extraction for industrial big data applications. 2016.

[139] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[140] Phillipp Anders, Espen Ingvald Bengtson, Karoline Blix Grønvik, Nina Skjæret-Maroni, and Beatrix Vereijken. Balance Training in Older Adults Using Exergames: Game Speed and Cognitive Elements Affect How Seniors Play. *Frontiers in Sports and Active Living*, 0:54, 5 2020.

[141] Sruti Subramanian, Nina Skjæret-Maroni, and Yngve Dahl. Systematic Review of Design Guidelines for Full-Body Interactive Games. *Interacting with Computers*, 32(4):367–406, 7 2020.

[142] Kjetil Raaen and Tor-Morten Grønli. Latency Thresholds for Usability in Games: A Survey. In *NIK*, 2014.

[143] Jan-Philipp Stauffert, Florian Niebling, and Marc Erich Latoschik. Latency and Cybersickness: Impact, Causes, and Measures. A Review. *Frontiers in Virtual Reality*, 1(November):1–10, 2020.

Part II

Publications

Chapter 8

Paper I

Comparison of a Deep Learning-Based Pose Esimation System to Marker-Based and Kinect Systems in Exergaming for Balance Training

Authors: Elise Klæbo Vonstad, Xiaomeng Su, Beatrix Vereijken, Kerstin Bach, Jan Harald Nilsen.

Published in Sensors 2020, 20, 6940; DOI: 10.3390/s20236940 sensors

MDPI

Article

Comparison of a Deep Learning-Based Pose Estimation System to Marker-Based and Kinect Systems in Exergaming for Balance Training

Elise Klæbo Vonstad [1,*], Xiaomeng Su [1], Beatrix Vereijken [2], Kerstin Bach [1] and Jan Harald Nilsen [1]

[1] Department of Computer Science, Norwegian University of Science and Technology, 7034 Trondheim, Norway; xiaomeng.su@ntnu.no (X.S.); kerstin.bach@ntnu.no (K.B.); jan.h.nilsen@ntnu.no (J.H.N.)

[2] Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, 7030 Trondheim, Norway; beatrix.vereijken@ntnu.no

* Correspondence: elise.k.vonstad@ntnu.no

Received: 26 October 2020; Accepted: 1 December 2020; Published: 4 December 2020

check for updates

Abstract: Using standard digital cameras in combination with deep learning (DL) for pose estimation is promising for the in-home and independent use of exercise games (exergames). We need to investigate to what extent such DL-based systems can provide satisfying accuracy on exergame relevant measures. Our study assesses temporal variation (i.e., variability) in body segment lengths, while using a Deep Learning image processing tool (DeepLabCut, DLC) on two-dimensional (2D) video. This variability is then compared with a gold-standard, marker-based three-dimensional Motion Capturing system (3DMoCap, Qualisys AB), and a 3D RGB-depth camera system (Kinect V2, Microsoft Inc). Simultaneous data were collected from all three systems, while participants (N = 12) played a custom balance training exergame. The pose estimation DLC-model is pre-trained on a large-scale dataset (ImageNet) and optimized with context-specific pose annotated images. Wilcoxon's signed-rank test was performed in order to assess the statistical significance of the differences in variability between systems. The results showed that the DLC method performs comparably to the Kinect and, in some segments, even to the 3DMoCap gold standard system with regard to variability. These results are promising for making exergames more accessible and easier to use, thereby increasing their availability for in-home exercise.

Keywords: motion capture; image analysis; markerless motion capture; exergaming; segment lengths; kinect; deep learning; human movement

1. Introduction

The proportion of older adults that are in need of guided physical exercise is expected to increase due to the coming demographic change. There is a need to develop technological tools that can aid and relieve clinicians in this effort [1,2]. In recent years, exercise gaming (exergaming) has emerged as a viable alternative, or addition, to traditional exercise. Exergames are designed to make the player move in a specific manner to train a specific function, such as stepping, sideways leaning, or moving their arms over their head. Typically, the person playing controls the game by moving their body: the game system captures their movements and uses this as input to control the game [3]. Exergames have also been shown to be more motivating and fun than traditional exercise [4,5], and they could potentially be used to provide quality, high-volume exercise guidance without having a physical therapist present to supervise [6,7].

Finding a motion capture tool that is suitable for this context is one of the areas in exergaming that is most challenging. If an older adult is going to use an exergame system, it needs to be user friendly, i.e., easy to understand and use efficiently, while providing input to the game that reliably represents the persons' movements. The latter is a prerequisite for the exergame system being useful in a serious setting: accurate and reliable capture of a person's movement is needed in order to ensure that the game is rewarding the player for correctly performed exercise movements and suggesting improvements to less-correctly performed movement patterns. This can facilitate good quality in performed movement patterns and increased motivation for exercise by providing appropriate rewards [8].

In this study, we investigate performance of a DL-based motion capture system by assessing the systems' temporal variation (i.e., variability) in estimating body segment lengths as compared to the gold standard 3DMoCap system. In the remaining of this section, we describe the three relevant systems and the rationale of segment length variation as an exergame relevant measure.

In kinematic analysis in settings, such as in exergaming, the movement patterns of body segments are used as input to the game. Therefore, segment lengths are important to keep relatively constant to avoid erroneous representation of the player in the game. Segment lengths are also vital in scaling biomechanical models to the person being measured [9], and segment definitions affect the kinematic analysis of movement [10,11]. The most accurate tools, i.e., the gold standard for measuring human movement, are marker-based 3D motion capture systems (3DMoCap). These systems are expensive, in terms of cost, time, and knowledge required to use them, and they are, as such, infeasible to use in a person's home.

One of the most popular tools for motion capture in exergaming is the Kinect (Microsoft Inc, Redmond WA, USA), a multiple-camera device while using RGB and depth (RGB-D) information in combination with machine learning-based (ML) analysis to detect human body parts and estimate three-dimensional (3D) joint positions of people, detected within the camera field of view [12]. In some contexts, Kinect cameras are useful, as they provide joint center position data directly with no need for additional processing of the depth or image data, as seen in, e.g., [13]. Kinect-based games vary in the gestures and movements that they elicit from the player. However, as there are many games that are designed for older adults, movements that challenge balance and posture control are common, as seen in [3]. Even though the joint center positions are not as accurately defined as in a 3DMoCap system [14], Kinect cameras have been shown to provide data that are sufficiently valid and reliable for some exergaming purposes [15,16]. However, we must be conscious of temporal variability in distal joints, such as wrists, elbows, knees, and ankles [17–19] when using Kinect-based systems for exergaming purposes. Additionally, even though Kinect cameras are more accessible than 3DMoCap systems, this still is an extra device that needs to be acquired and correctly set up before having access to exergames. This could be a potential barrier of use and it could be circumvented by utilizing standard digital cameras available in most homes today, such as smartphones, web cameras, and tablets.

Using standard digital video in motion capture has received increased attention in recent years, due to advances in DL-based image processing techniques. Frameworks, such as OpenPose [20], DeeperCut [21], and EfficientPose [22], provide kinematic information by extracting joint positions from video. DeepLabCut (DLC, [23]) is another interesting DL-based system that could potentially be used for pose estimation in humans. This has previously been used to reliably track points of interest on animals and insects in standard video and it could potentially be used to acquire human motion data during exergaming. One interesting feature of this framework is that it was shown to require a relatively low number of training samples in order to accurately predict joint center locations in unseen videos [23]. This is achieved by using transfer learning, which specializes the network to the specific context at hand. As noted in [24], using context-specific training data can optimize the pose estimation model, which might improve performance. This can potentially provide exergame users with a tool that performs with high accuracy in specialized settings, where the reliability of the exergame is vital in ensuring proper feedback and guidance. The framework is available through an open source, easy to use toolbox (github.com/deeplabcut/deeplabcut). Athough the above mentioned DL methods for pose estimation and joint tracking are evaluated for accuracy by comparing the estimated joint locations to the human pre-labelled training data, the tracking accuracy is rarely compared to a gold standard 3DMoCap system. This might be sufficient for their respective contexts, but, if we are to use such systems in settings such as exergaming in rehabilitation, where accuracy and stability in motion tracking are vital, we need to know how these systems perform when compared to the gold standard. Despite the potential of using standard digital video for motion tracking for exergaming, comparisons of a DL system, such as DLC to gold standard motion capture systems, has not yet been extensively investigated with regard to exergame relevant measures, such as variability of for example segment lengths. To our knowledge, the current study is the first to compare DLC to a 3DMoCap system and a Kinect camera system in terms of variability in segment lengths.

It is crucial for the further development and implementation of exergames to develop markerless motion capture systems that are easy to use and reliable to ensure proper feedback and guidance to the users during exergaming. In order to contribute to this goal, this study aims at comparing a transfer learning-based pose estimation model (DLC) to the Kinect system and a gold-standard 3DMoCap system.

The remainder of this paper is structured as follows: the methods and materials are found in Section 2. Section 3 details our results, and the discussion is found in Section 4.

2. Materials and Methods

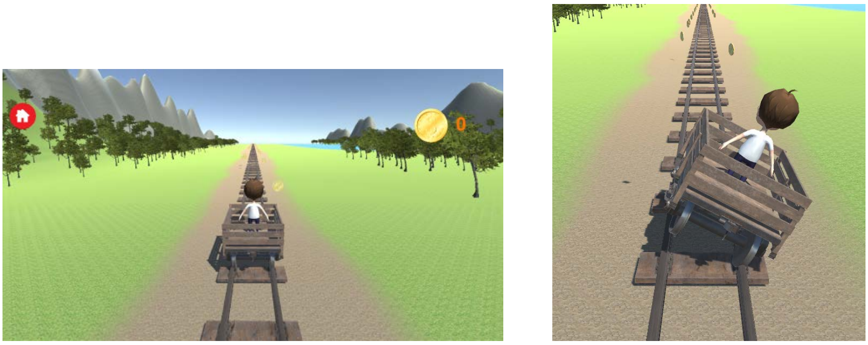
2.1. Participants

We recruited healthy older adults from a local exercise group for seniors in Trondheim Norway. The inclusion criteria were age >65 years and the absence of physical or cognitive impairments or conditions that affected balance or gait ability. There were 12 participants in total (10F); the average age was 70.4 years (SD 11.4, range 54–92). The average height and weight were 172.3 (±11.4) cm and 70.4 (±12.1) kg, respectively. The exclusion criteria were physical or cognitive injuries/impairments that affected their balance or gait ability, and age <50 or age >80 years. The participants were given oral and written information regarding the study and gave their written consent, and the Norwegian Center for Research Data approved the study (reference number 736906). The participants attended one session each, and all completed the data collection without incident. The participants wore t-shorts and shorts of different types, colors, and fabrics, and some wore shorts and/or a sports bra.

2.2. Protocol

2.2.1. The Exergame

As medio-lateral weight-shifting movements are paramount in prevention of loss of balance function in older adults [25], and they are often used in Kinect-based games [26], a custom weight-shifting exergame was developed for the purpose of this study. The exergame was designed in order to elicit medio-lateral weight shifts from the participants: On the screen, an avatar in a rail cart was controlled by the participant, as seen in Figure 1A. If the participant leaned to one side the rail cart also tilted to the same side, allowing it to hit coins along the rail track as the cart moved down the track (Figure 1B).



(**A**) Start of game (**B**) Cart lateral movement

Figure 1. Screenshots from the game. (**A**) shows the start of the game, and (**B**) shows the cart leaning sideways with the movements of the player to hit coins along the track.

2.2.2. Equipment

As input to control the game, a Kinect (v2, 30 Hz, Microsoft Inc., Redmond, WA, USA) camera system was used in order to track player movements. Participants' movements were simultaneously measured while using 36 reflective markers placed according to the Plug-in-Gait Full-Body marker placement guide (PiG-FB, [27]) and a 3DMoCap system (90 Hz, Qualisys AB, Gothenburg, Sweden) consisting of four MX400 cameras, and a normal digital camera (30 Hz, 1400 × 720 px, GoPro Hero Black 3+, GoPro Inc., San Mateo, CA, USA) positioned 200 cm behind the center of the starting position of the participants. The participants' height, weight, and age was also recorded. Figure 2 depicts a schematic of the experimental setup. The participants were standing on a 160 × 60 × 5 cm game platform while playing, where they had one force plate (Kistler Group, Winterthur, Switzerland) under each foot. Measurements from this equipment were not used in the current study.

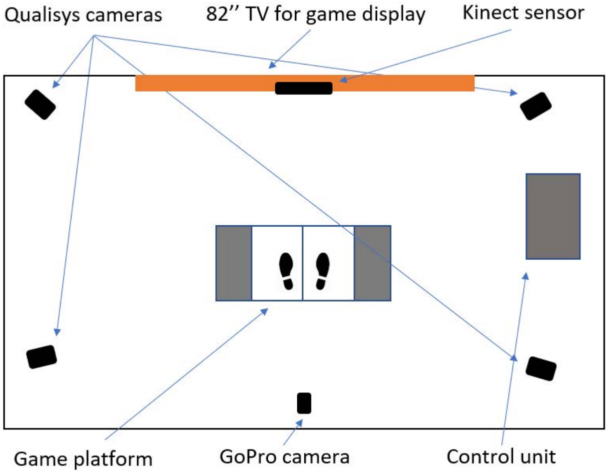


Figure 2. Experimental setup.

2.3. Processing and Analysis

Processing was conducted while using Python (v 3.7), and statistical analyses were performed while using SPSS Statistics (v. 26, IBM Corp, Armonk, New York, NY, USA).

2.3.1. Dataset

The current dataset consists of data from participants leaning from side to side to control the exergame. Each participant played six trials of the game, with each trial lasting for around five minutes. There is 3DMoCap data from 11 participants, DLC data from 12 participants, and Kinect data from 12 participants. In some participants (4, 8, 9, 10), the field of view of the GoPro camera resulted in ankles not being visible in the video. The ankle joints from these participants were excluded from analysis by the DLC system. 3DMoCap data from one participant (2) was corrupted and it was not included in the analysis.

2.3.2. Preprocesssing of Kinect and 3DMoCap Data

The standard PiG-FB biomechanical model was used in order to extract joint center locations from 3DMoCap data. Joint center positions (distance to lab-coordinate system origin, mm) from shoulders, elbows, wrists, hips, knees, and ankles were extracted (Figure 3B). Normally, segment lengths (i.e., distance between joint center locations) are calibrated by the PiG-FB biomechanical model at the beginning of a trial and kept constant throughout the data capture. In this study, however, variability of segment lengths during the trial are of particular interest and will be used instead of the constant segment length calculated from the PiG-FB biomechanical model. For Kinect camera data, joint positions of shoulder, elbow, wrist, hip, knee, and ankle joints in the X and Y-axis (relative to the coordinate system origin within the camera itself) of the camera were extracted from the Kinect skeletal model (Figure 3C) using Kinect Studio (v. 2.0.14, Microsoft Inc) and the Kinect2Toolbox [28]. Because the data from Kinect were originally reported in meters, they was converted to millimeters to be comparable to DLC and 3DMoCap data.

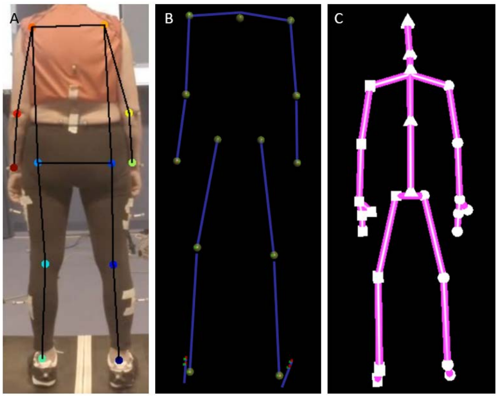


Figure 3. Joint centers as defined by the three motion capture systems (not to scale). (**A**) = DeepLabCut, (**B**) = 3DMoCap, (**C**) = Kinect.

2.3.3. Preprocessing of DeepLabCut Data

The DLC framework is based on a feature detector method from one of the state-of-the-art human pose estimation frameworks, DeeperCut [21]. This employs a variant of ResNet that was pre-trained on the ImageNet [29] database. Semantic segmentation of images containing body parts is performed on all frames of the video data, and deconvolutional layers are used in order to up-sample the images after convolution and, thus, ensure sufficient spatial resolution for body part detection. Instead of the classification output layer of the CNN, score maps for the predictions of a body part in an image is produced. These spatial probability densities are then fine-tuned for each body part while using labelled images.

Digital video data were analyzed while using the DLC implementation software DeepLabCut (DLC, github.com/DeepLabCut/DeepLabCut). Joint center locations of shoulders, elbows, wrists, hips, knees, and ankles (Figure 3A) were manually applied to three images from two videos from each participant by an experienced human movement scientist, totaling 194 labelled images. This is in line with recommendations in [23]. These joint center locations in images were then used as training data for the neural network (ResNet101). The train/test split was set to 95/5. The DLC was trained for 220,000 iterations, with loss plateauing at 0.0012 at a p-cutoff of 0.01. After training, all of the videos were analyzed by the DLC, and the predicted joint center pixel locations were extracted. The DLC was then evaluated on the 5 % left out data to assess whether overfitting had occurred. To convert video pixel data to mm, the distance between shoulder joint centers were extracted from 3DMoCap data and then used as a reference to calculate pixel size. One pixel was found to be approx. 3.5 mm: the DLC pixel data was converted to mm by multiplying the pixel information with 3.5.

2.3.4. Calculation of Segment Lengths and Variability

Segment lengths of shoulders, left and right upper arms, left and right lower arms, left and right torso sides, pelvis, left and right thighs, and left and right shanks were found by calculating the Euclidean distance between the joint centers for each joint set for each participant (Figure 4). Data points that were outside of mean ±3 standard deviations (SD) were considered to be outliers and removed from the dataset.

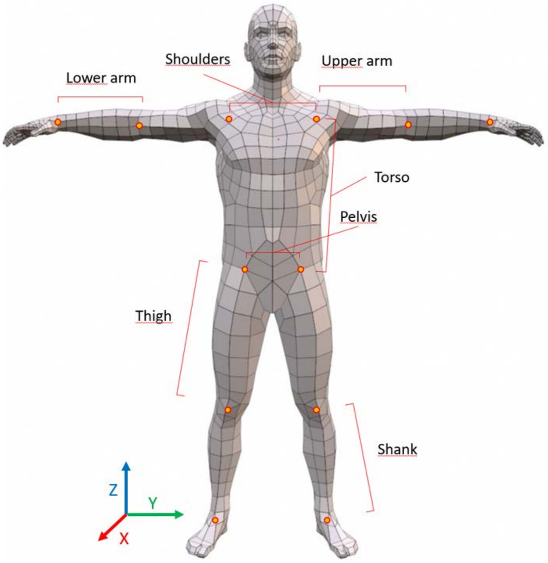


Figure 4. Joint locations, axes directions, and segment length definitions extracted from all three motion capture systems.

2.3.5. Statistical Analysis

The 3DMoCap, Kinect, and DLC segment length variability were assessed by analyzing the Euclidean distance between joints in each time frame for each of the camera systems. To represent variability in these distances, standard deviation and coefficient of variation (coeffVar; dispersion of data around the mean) was employed. The Shapiro–Wilks test for normality gave a $p < 0.05$ for all segment lengths, which revealed that the data were not normally distributed. Therefore, the non-parametric Friedman test was conducted to assess statistical differences in segment length variability between the three systems. Subsequently, a post hoc analysis was conducted on the statistically significant differences from the Friedman test in order to extract which between-system differences were statistically significant. The post hoc analysis was conducted by using Wilcoxon's signed-rank test with a Bonferroni correction, resulting in α = 0.017.

3. Results

Table 1 shows the results from the mean segment lengths from each motion capture system. Table 2 shows the mean SD of segment lengths from each motion capture system. The results from the Friedman analyses of statistically significant difference can be found in Table 3. In Figure 5, a comparison of the variability of the segment length over 1000 frames of the shoulder and shank segments can be found. Figure 6 shows the results from the post hoc-test, as well as the median and IQR values as box plots.

Table 1. Mean segment lengths (mm (1SD)). L = left, R = Right. *N* = data from number of participants. 3DMoCap = 3D motion capture system, DLC = DeepLabCut

Segment	Side	3DMoCap		DLC		Kinect	
		N		*N*		*N*	
Shoulders		11	328.8 (23.5)	12	308.8 (25.5)	12	330.9 (20.2)
Upper arm	L	11	269.5 (19.1)	12	351.0 (20.5)	12	269.8 (15.9)
	R	11	279.5 (22.6)	12	357.8 (23.2)	12	267.1 (13.2)
Lower arm	L	11	228.5 (20.4)	12	228.7 (16.6)	12	235.8 (13.3)
	R	11	225.1 (12.8)	12	231.9 (16.1)	12	235.3 (14.8)
Torso	L	11	444.7 (27.8)	12	568.5 (33.7)	12	503.8 (27.4)
	R	11	439.9 (25.9)	12	566.1 (38.3)	12	497.9 (27.6)
Pelvis		11	148.6 (5.6)	12	280.6 (28.5)	12	154.8 (9.5)
Thigh	L	11	409.0 (33.2)	12	405.9 (21.9)	12	373.8 (26.1)
	R	11	410.6 (33.0)	12	411.9 (27.1)	12	372.5 (29.4)
Shank	L	11	404.9 (23.4)	8	415.2 (34.0)	12	378.8 (29.0)
	R	11	402.8 (22.4)	8	414.4 (33.5)	12	374.3 (27.3)

Table 2. Mean standard deviation (mm, (coefficient of variation)) of segment lengths. L = left, R = Right. *N* = data from number of participants. 3DMoCap = 3D motion capture system, DLC = DeepLabCut. Light green = lowest mean SD within system; light red = highest SD within system. Bright green = overall lowest mean SD; bright red = overall highest mean SD.

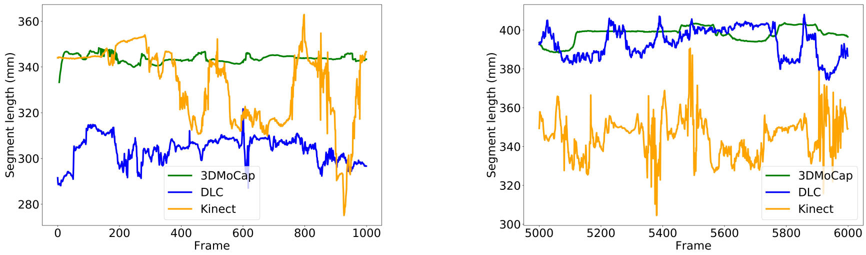
Segment	Side	3DMoCap		DLC		Kinect	
		N		*N*		*N*	
Shoulders		11	9.1 (0.02)	12	16.6 (0.04)	12	17.3 (0.05)
Upper arm	L	11	7.4 (0.03)	12	11.7 (0.04)	12	15.1 (0.05)
	R	11	7.3 (0.02)	12	13.0 (0.04)	12	15.2 (0.06)
Lower arm	L	11	9.6 (0.04)	12	14.4 (0.08)	12	13.7 (0.06)
	R	11	10.2 (0.04)	12	20.4 (0.08)	12	13.3 (0.05)
Torso	L	11	15.9 (0.03)	12	22.5 (0.04)	12	12.8 (0.02)
	R	11	15.7 (0.09)	12	22.5 (0.03)	12	13.1 (0.02)
Pelvis		11	2.8 (0.01)	12	7.3 (0.04)	12	6.1 (0.03)
Thigh	L	11	8.3 (0.02)	12	16.4 (0.03)	12	25.5 (0.06)
	R	11	8.7 (0.02)	12	20.5 (0.04)	12	23.1 (0.06)
Shank	L	11	8.6 (0.02)	8	14.5 (0.02)	12	21.1 (0.05)
	R	11	8.6 (0.02)	8	13.6 (0.02)	12	20.5 (0.05)

Table 3. Chi-square (X^2), *p*-value, and mean ranks from the Friedman test of statistical difference between mean segment length standard deviation. Df = degrees of freedom. L = left, R = Right. 3DMoCap = 3D motion capture system, DLC = DeepLabCut.

Segment	Side			Mean Rank		
		X^2(df)	*p*	3DMoCap	DLC	Kinect
Upper arm	L	3.8 (2)	0.148	1.55	2.09	2.36
	R	8.7 (2)	0.023	1.27	2.36	2.36
Lower arm	L	11.6 (2)	0.003	1.27	2.73	2.0
	R	7.81 (2)	0.020	1.45	2.64	1.91
Shoulders		11.1 (2)	0.004	1.18	2.45	2.36
Torso	L	5.6 (2)	0.060	1.91	2.55	1.55
	R	5.5 (2)	0.103	2.00	2.45	1.55
Pelvis		20.2 (2)	0.000	1.09	3.00	1.91
Thigh	L	16.5 (2)	0.000	1.18	1.91	2.91
	R	16.9 (2)	0.000	1.0	2.36	2.64
Shank	L	4.6 (2)	0.102	1.43	2.00	2.57
	R	8.9 (2)	0.012	1.14	2.14	2.71

3.1. Mean Lengths

Representative examples shown in Figure 5 depict temporal segment length variabilities in the three different motion capture systems. These are examples of the variability seen in segment lengths of the shoulder (Panel 5a) and right shank (Panel 5b) during 1000 frames, which corresponds to 33.3 s. The mean length of the shoulder segment that is seen in Figure 5A is approx. 344 mm, as measured by the 3DMoCap system with a range of approx. 15 mm. The DLC has a shorter segment (304 mm) with slightly higher variability (range 35 mm) when compared to the 3DMoCap system. The Kinect system starts with about the same segment length as the 3DMoCap system (333 mm), but shows increasing variability throughout the trial with a range of 88 mm. Differences in the average segment lengths are most likely due to different definitions of where the joint centers are located, as seen in Figure 3. The right shank segment (Figure 5B) shows similar results: the 3DMoCap system and the DLC measure similarly in mean shank length, at 393 mm and 398 mm, respectively, but the DLC has a larger range of 33 mm as compared to 15 in the 3DMoCap system. The Kinect system has a lower shank length, averaging at about 345 mm, and the variability is higher throughout the trial, with a range of 86 mm.



(**A**) Shoulders

(**B**) Right shank

Figure 5. Comparison of variations in temporal segment lengths of the shoulder (**A**) and right shank (**B**). Image recording frequency 30 Hz.

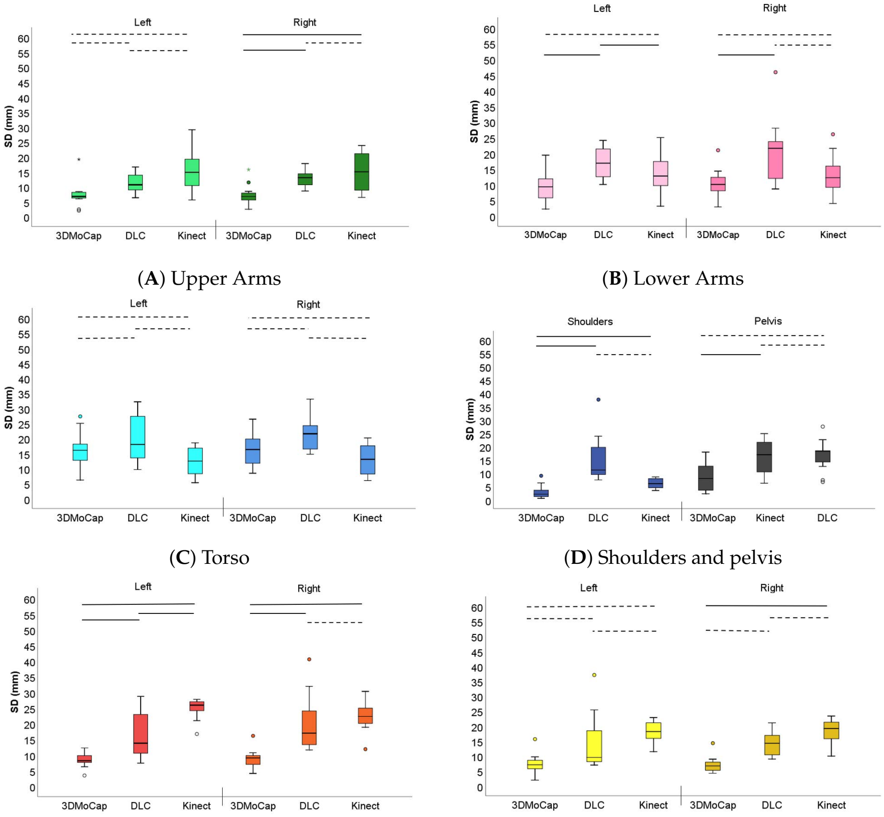
The mean segment lengths calculated from all data for all participants vary between the three motion capture systems, as shown in Table 1. This reflects the different biomechanical models mentioned earlier, although some segments are defined similarly and have similar lengths. Furthermore, the pelvis lengths show that, even though the Kinect system typically defines the hip joints more superior, or towards the head, as compared to the 3DMoCap system [14], the size of the segment is similar between the two systems. An example of this is the shoulder segment, where the difference between the Kinect and the 3DMoCap system is <2.5 mm, and approximately 20 mm between the DLC and the 3DMoCap. The Kinect system also seems to underestimate lower body segment lengths compared to 3DMoCap, while the DLC shows similar segment lengths here.

3.2. Segment Length Variability

Table 2 shows the variability of all systems for all segment lengths. This shows that the overall highest mean SD was 25.5 mm (Kinect, left thigh) and the overall lowest mean SD was 2.8 mm (3DMoCap, pelvis). The average of all mean SDs was 9.4 mm (SD 3.6) in the 3DMoCap system, 16.1 (SD 4.5) in the DLC system, and 16.4 mm (SD 5.1) in the Kinect system. The table also shows that the coeffVar was generally low, further indicating low variability in all three systems.

3.2.1. Upper and Lower Arm

Figure 6A, B, shows arm segment variability, where panel A shows the upper arm and panel B shows the lower arm.



(**A**) Upper Arms

(**B**) Lower Arms

(**C**) Torso

(**D**) Shoulders and pelvis

(**E**) Thighs

(**F**) Shanks

Figure 6. Box plots of variation of standard deviations of upper arms (**A**), lower arms (**B**), torso (**C**), shoulders and pelvis (**D**), thighs (**E**), and shanks (**F**) for the left and right side of the body. 3DMoCap = 3D motion capture system, DLC = DeepLabCut. Dotted lines signify $p > 0.017$, solid lines $p < 0.017$ from Wilcoxons Signed Rank test.

In the left upper arm (Panel A), the difference in SD between DLC (median 10.6 mm, IQR 8.2 to 14.0), and Kinect (median 14.8 mm, 7.5 to 19.2) is not statistically significant. The Kinect and the DLC system both had a higher SD than the 3DMoCap system (median 6.7 mm, IQR 5.9 to 8.2), but this was not statistically significant. In the right upper arm, the only difference in mean SD that is not statistically significant is between DLC (median 13.0 mm, IQR 10.3 to 14.8) and Kinect (median 15.0 mm, IQR 7.5 to 21.8), and both show statistically significant higher SD as compared to the 3DMoCap (median 6.7 mm, IQR 4.9 to 8.5).

In the left lower arm (Panel B), Kinect (median 12.7 mm, IQR 8.5 to 18.2) variability was not statistically different from the 3DMoCap (median 9.3 mm, IQR 4.8 to 11.9). However, the DLC (median 16.8 mm, IQR 12.0 to 21.7) showed statistically significant higher variability than the 3DMoCap and the Kinect. In the right lower arm, the DLC (median 21.5 mm, IQR 11.1 to 24.0) had statistically significant higher variability than 3DMoCap (median 10.0 mm, IQR 8.0 to 12.3), but not to the Kinect (median 12.2 mm, IQR 8.3 to 16.3).

3.2.2. Torso and Shoulders

Panel C shows the median SD difference between the systems in the torso (left and right side). Here, the difference in mean SD was non-significant between any of the systems on neither the left (3DMoCap median 16.0 mm (IQR 11.4 to 18.3), Kinect median 12.4 mm (IQR 7.6 to 17.1) DLC median 18.0 mm (IQR 13.4 to 28.2)) nor the right side (3DMoCap median 16.3 mm (IQR 9.6 to 20.4), Kinect median 13.0 mm (7.2 to 18.6) DLC median 21.5 mm (IQR 16.4 to 26.6)). In the shoulder segment (Panel D) the difference between DLC (median 17.5 mm, IQR 8.8 to 23.0) and Kinect (median mm 18.7, IQR 13.2 to 18.9) is also non-significant. The median SD for the 3DMoCap shoulder segment was 8.7 mm (IQR 3.8 to 14.6).

3.2.3. Pelvis

The pelvis segment SD (Panel D) difference was not statistically significant between the DLC (median 11.1 mm (IQR 9.4 to 21.0)) and Kinect (median 6.0 mm (IQR 4.2 to 8.0). The difference between the Kinect and 3DMoCap (median 2.1 mm, IQR 1.1 to 3.8) was not statistically significant.

3.2.4. Thigh and Shanks

In panel E, the results for mean SD difference between the systems in the thigh segments are presented. These show that differences between systems are statistically significant on both the left and right thighs, except for DLC as compared to Kinect on the right side. The median SD of the thigh segment in 3DMoCap was 8.1 mm (IQR 7.6 to 9.9) and 9.1 mm (IQR 6.4 to 9.9) on the left and right side, respectively, while the median SD of DLC was 13.7 mm (IQR 10.5 to 23.1) and 17.0 mm (IQR 12.5 to 28.4), for the left and right side, respectively. The thigh median SD was the highest in the Kinect, with 25.9 mm (IQR 23.7 to 27.1) and 22.3 mm (IQR 18.9 to 15.8) on the left and right side, respectively.

On the left side, there are non-significant differences between 3DMoCap (median 7.0 mm, IQR 6.0 to 9.7) and DLC (median 9.5 mm, IQR 7.0 to 25.4), between DLC and Kinect (median 18.1 mm, IQR 15.1 to 21.9) and between 3DMoCap and Kinect, as seen in the results of shank segment variability (Panel F). On the right side, the difference between DLC (median 14.2 mm, IQR 9.5 to 19.1) and Kinect (median 19.1 mm, IQR 14.3 to 21.3) is not statistically significant, but the Kinect has a significantly higher mean SD than 3DMoCap (median 6.6 mm, IQR 5.0 to 8.9). The difference between DLC and 3DMoCap is not statistically significant.

4. Discussion

In this paper, we compared the DLC image analysis system to a Kinect system and a gold standard 3D motion capture system by studying the variability in segment lengths. Overall, the DLC method and Kinect system showed slightly higher variability than the 3DMoCap system, but this was not statistically significant for all of the comparisons. The highest mean SD found (25.5 in left thigh from

Kinect, Table 2) shows that the systems generally perform with acceptable variability, even in the worst performing segment.

Our analyses show that the DLC variability difference is not statistically significant from the gold standard 3DMoCap system in several segments, namely left upper arm, left and right torso, and left and right shank. The Kinect variability is not statistically significantly different from the gold standard 3DMoCap system in the left upper arm, left and right lower arm, left and right torso, pelvis, and left shank. The difference in variability between the DLC and the Kinect is not statistically significant in the left and right upper arms, the left lower arm, left and right torso, shoulders, pelvis, right thigh, and left and right shanks.

4.1. Implications

The low variability of the DLC in predicting shank joint centers is promising for applying it in balance training settings while using weight-shifting movements, where stability in foot tracking is essential. This also applies to stepping exercises, which also constitute an important part of recommended exercises for older adults [30].

In the upper arms, the DLC and the Kinect systems both show slightly higher variability than the 3DMoCap system in the right arm. The Kinect showed highest variability, but this difference was not statistically significant when compared to the 3DMoCap. Reaching and leaning are often used in exercise for postural control and balance in older adults [31], and these results show that it might be feasible to use DLC in exergames that aim at eliciting such movements from players. The results from the shoulder segment also support this; even though the variability was highest in the DLC system, it was comparable to the Kinect.

Even though the hip joint centers often are difficult to track for marker-less models [32], our results show that the pelvis is tracked with a stability that is comparable to the 3DMoCap by both the Kinect and the DLC systems. This is also the case in the torso segments. They are able to reliably track the pelvis and torso segments is important in many exergames, as these often provide the base segments for assessing balance movements [33].

The thigh segment variability differences were statistically significant between all systems in both the left and right segment, except for the DLC as compared to Kinect in the right thigh. Here, the Kinect showed the highest median variability of all segments, while the DLC showed high IQR. The use of DLC in balance training for older adults is still feasible, as the variability is only slightly higher than in the 3DMoCap system. However, this needs to be taken into consideration by developers and clinicians who aim at introducing DL-based motion tracking into exergames for balance training, as there is some uncertainty of stability in the lengths of these segments.

4.2. Related Work

Our results are in line with previous studies on the validity of the Kinect system [14,15,17,18,34,35]. The Kinect generally performs with some difference to the gold standard 3DMoCap system, but within acceptable ranges in the contexts studied. We can compare our results to these studies, as pose estimation from (monocular) image data using other DL-based methods is an adjacent field of research. There are several methods utilizing ResNets in order to predict joint positions, as shown in Chen et al. [36]. These typically achieve Mean Per Joint Position Error (MPJPE, mm) of 48–108, which is comparable to our results of mean SD. Other DL methods achieve varying MPJPE [36], which ranges from 130 to 40, where the latter was achieved by Sun et al. [37] while using a volumetric representation of 3D pose by heat maps and joint regression. Furthermore, Arnab et al. [38] and Mehta et al. [39] developed models that were able to reach MPJPE of 54.3 mm and 63.6 mm, respectively, when testing on the Human3.6M dataset [40]. Despite the importance of knowing the temporal variability that a human pose estimation system has, there is limited research on the variability of segment lengths. It is reported in e.g., [18], where thigh, shank, upper, and lower arm variability was found to be higher in the Kinect when compared to the 3DMoCap system. Moreover, a similar study to ours was conducted by Nakano et al [41], where the results of using a variant of the OpenPose model showed a mean absolute error (MAE) of joint positions ranging from <5 mm to >40 mm when compared to 3DMoCap. Note that this study used five synchronized cameras for 3D OpenPose joint tracking and did not directly report variability. To our knowledge, this study is the first to evaluate the DLC system on variability in human pose estimation as compared to Kinect and 3DMoCap.

4.3. Further Considerations and Future Directions

Being aware of the potential positive and negative sides of different motion capture systems is vital when assessing whether a system is suitable for use in a given situation. Different measurement techniques can inherently influence data quality. Even though 3DMoCap is considered to be the gold standard for accuracy, it is prone to marker placement errors, soft issue artifacts, and marker occlusion [42,43]. Kinect based systems, or specifically the Kinect joint location algorithm, can have poor joint tracking in situations where a body part is not visible to the camera, unusual poses, or interaction with objects [18]. Previous research has extensively documented these weaknesses in both 3DMoCap and Kinect systems. However, because the DL-based systems in motion capture settings are still in their infancy, they have not yet undergone the scrutiny that widespread use gives. Therefore, the possible limitations and sources of error in DL methods, such as DLC, can be found in the technical details of how the method works. For instance, some CNN-based methods, such as DLC, are pre-trained on an enormous dataset (ImageNet, [29]) and then trained on data from specific contexts (such as in the current study) in the last stage of training. This means that the DLC might already be biased before it is trained on our context-specific image data, which could impact our results in unknown ways. Generally, DL models require such large datasets as ImageNet (>14.2 million images) in order to have sufficient training data for a given classification/estimation problem. However, the DLC system avoids this issue and it only requires around 200 images to predict joint centers in context-specific situations. This is possible due to pre-training, i.e., transfer learning, and the use of spatial probability densities for locating body parts in images (see [21,23] for details). This requirement of around 200 images is much more feasible to achieve for context-specific motion tracking, and it can make DL-based analysis within reach for users who do not have access to the normally required large-scale datasets for their applications.

Because DL models are restricted to learning from the training data that they see, the training data will significantly influence the outcome of the predictions that they make. In practice, this implies that, for a DLC system to be usable for any type of person, we need a dataset that represents all types of person in order to avoid errors that result from a non-representative dataset. This can be both a practical and an ethical issue in these systems, while being avoided in systems that (semi-)directly measure the points of interest, such as Kinect and 3DMoCap systems, or by using pre-trained networks. It is vital to build the knowledge required to introduce new technology into settings where users are particularly vulnerable, such as patient settings, where efficient time usage is critical for progress in regaining physical function. Errors or unintended bias in such systems can have consequences that are detrimental to a persons' health or quality of life, which is why it is essential to highlight and discuss these issues in the context of motion capture for the rehabilitation or prevention of loss of physical function. Even though these issues are important to keep in mind when using DL systems, there is great potential in using such methods for motion capture in settings that require ease of use and where other motion capture systems are not feasible to use. An interesting future direction would be benchmarking DLC on large, variable datasets, and research is underway in order to investigate the performance of DLC in 3D settings with a larger variety in movement directions and types.

4.4. Limitations

There are some limitations to this study that are important to be aware of. There were only 12 participants, which might limit the generalizability of these results to the elderly population. Furthermore, the movements performed while playing the balancing exergame were mostly limited to the frontal plane of the participants. Other more complex movements might make the prediction of segment lengths more challenging. This is an important area for future research. The Kinect camera and the digital video camera were set up in their optimal configuration for capturing motion data from participants: the Kinect camera was in the anterior view of the frontal plane of participants and the GoPro in the posterior view of the frontal plane. However, the 3DMoCap system only consisted of four cameras, which led to missing information in some of the marker trajectories because of the occlusion of markers in one or more cameras, possibly contributing to the 3DMoCap system not achieving the performance possible in such systems.

4.5. Conclusions

Overall, these results are encouraging. The aim of introducing novel technological solutions in exergaming is to improve the cost efficiency and ease of use, thereby making exergaming more accessible for older adults or patients in the home or at a rehabilitation clinic. This is dependent on technological solutions that provide sufficient information regarding the person's movement while exergaming. Using readily available digital cameras to track movement during exergaming could provide such a solution, thereby making it possible to use exergames without needing technical assistance. The results of the current study are the first to show that a DL-based motion capture system using transfer learning can achieve measurement stability in segment lengths that were comparable to a popular motion capture camera in exergaming for balance training, and, in some segments, even comparable to the gold-standard in motion capture.

Although 3DMoCap provides the best possible accuracy, in the trade-off between ease of use and accessibility versus accuracy, the former is given priority because of the requirements of the home/older adult context. In other situations, the difference between a DLC and a 3DMoCap might be considered too large, e.g., in a clinical gait analysis setting. The segment length variability that was found in this study would impact joint angle measurements, which results in the clinical assessment of the gait function of the patient potentially being altered. Such contexts will continue to require high accuracy in order to limit the risk of incorrect data, and marker-based 3DMoCap systems are therefore still the best option here—for the time being. Continuing research for improving marker-less systems is an important direction to take because of the potential benefits in resource use, usability, and flexibility. The results of the current study warrant further investigation into using DLC or similar systems in more complex movement patterns and other camera positions, and also implementing it in real-time in exergame settings.

Author Contributions: Conceptualization, E.K.V. and J.H.N.; methodology, E.K.V. and J.H.N.; software, E.K.V.; validation, J.H.N., K.B. and B.V.; formal analysis, E.K.V.; investigation, E.K.V.; resources, E.K.V.; data curation, E.K.V.; writing—original draft preparation, E.K.V.; writing—review and editing, X.S., J.H.N., K.B. and B.V.; visualization, E.K.V.; supervision, J.H.N.; project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: Authors would like to thank the participants for kindly spending their valuable time and effort to partake in the data collection. Furthermore, authors would like to thank Martin Aursand Berge for his contribution to the data collection as a research assistant.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DL	Deep Learning
ML	Machine Learning
RBG-D	Red Blue Green - Depth
3DMoCap	3D Motion Capture
IMU	Inertial Measurement Unit

ResNet	Residual Neural Network
DLC	DeepLabCut
CNN	Convolutional Neural Network
SD	Standard Deviation
CoeffVar	Coefficient of Variation

References

1. Beard, J.R.; Bloom, D.E. Towards a comprehensive public health response to population ageing. *Lancet* **2015**, *385*, 658–661, doi:10.1016/S0140-6736(14)61461-6.
2. Proffitt, R.; Lange, B. Considerations in the Efficacy and Effectiveness of Virtual Reality Interventions for Stroke Rehabilitation: Moving the Field Forward. *Phys. Ther.* **2015**, *95*, 441–448, doi:10.2522/ptj.20130571.
3. Skjæret, N.; Nawaz, A.; Morat, T.; Schoene, D.; Lægdheim, J.; Vereijken, B. Exercise and rehabilitation delivered through exergames in older adults : An integrative review of technologies, safety and efficacy. *Int. J. Med. Inform.* **2016**, *85*, 1–16, doi:10.1016/j.ijmedinf.2015.10.008.function.
4. Wuest, S.; Borghese, N.A.; Pirovano, M.; Mainetti, R.; van de Langenberg, R.; de Bruin, E.D. Usability and Effects of an Exergame-Based Balance Training Program. *Games Health J.* **2014**, *3*, 106–114, doi:10.1089/g4h.2013.0093.
5. Chen, M.H.; Huang, L.L.; Wang, C.H. Developing a Digital Game for Stroke Patients' Upper Extremity Rehabilitation—Design, Usability and Effectiveness Assessment. In Proceedings of the 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, Las Vegas, NV, USA, 26–30 July, 2015; Volume 3, pp. 6–12, doi:10.1016/j.promfg.2015.07.101.
6. Rand, D.; Givon, N.; Weingarden, H.; Nota, A.; Zeilig, G. Eliciting upper extremity purposeful movements using video games: A comparison with traditional therapy for stroke rehabilitation. *Neurorehabilit. Neural Repair* **2014**, *28*, 733–739, doi:10.1177/1545968314521008.
7. Lang, C.E.; Macdonald, J.R.; Reisman, D.S.; Boyd, L.; Kimberley, T.J.; Schindler-Ivens, S.M.; Hornby, T.G. Observation of Amounts of Movement Practice Provided During Stroke Rehabilitation. *Arch. Phys. Med. Rehabil.* **2009**, *90*, 1692–1698, doi:10.1016/j.apmr.2009.04.005.
8. de Rooij, I.J.M.; van de Port, I.G.L.; Meijer, J.W.G. Effect of Virtual Reality Training on Balance and Gait Ability in Patients with Stroke: Systematic Review and Meta-Analysis. *Phys. Ther.* **2016**, *96*, 1905–1918, doi:10.2522/ptj.20160054.
9. Crabolu, M.; Pani, D.; Raffo, L.; Conti, M.; Cereatti, A. Functional estimation of bony segment lengths using magneto-inertial sensing: Application to the humerus. *PLoS ONE* **2018**, *13*, e203861, doi:10.1371/journal.pone.0203861.
10. Kainz, H.; Modenese, L.; Lloyd, D.; Maine, S.; Walsh, H.; Carty, C. Joint kinematic calculation based on clinical direct kinematic versus inverse kinematic gait models. *J. Biomech.* **2016**, *49*, 1658–1669, doi:10.1016/j.jbiomech.2016.03.052.
11. Tak, I.; Wiertz, W.P.; Barendrecht, M.; Langhout, R. Validity of a New 3-D Motion Analysis Tool for the Assessment of Knee, Hip and Spine Joint Angles during the Single Leg Squat. *Sensors* **2020**, *20*, 4539, doi:10.3390/s20164539.
12. Shotton, J.; Fitzgibbon, A.; Blake, A.; Kipman, A.; Finocchio, M.; Moore, B.; Sharp, T. Real-Time Human Pose Recognition in Parts from a Single Depth Image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011.
13. Gaglio, S.; Re, G.L.; Morana, M. Human Activity Recognition Process Using 3-D Posture Data. *IEEE Trans. Hum.-Mach. Syst.* **2015**, *45*, 586–597, doi:10.1109/THMS.2014.2377111.
14. Obdrzalek, S.; Kurillo, G.; Ofli, F.; Bajcsy, R.; Seto, E.; Jimison, H.; Pavel, M. Accuracy and robustness of Kinect pose estimation in the context of coaching of elderly population. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, San Diego, CA, USA, 28 August–1 September 2012; pp. 1188–1193, doi:10.1109/EMBC.2012.6346149.
15. van Diest, M.; Stegenga, J.; Wörtche, H.J.; Postema, K.; Verkerke, G.J.; Lamoth, C.J. Suitability of Kinect for measuring whole body movement patterns during exergaming. *J. Biomech.* **2014**, *47*, 2925–2932, doi:10.1016/j.jbiomech.2014.07.017.

16. Ma, M.; Proffitt, R.; Skubic, M. Validation of a Kinect V2 based rehabilitation game. *PLoS ONE* **2018**, *13*, e202338, doi:10.1371/journal.pone.0202338.
17. Otte, K.; Kayser, B.; Mansow-Model, S.; Verrel, J.; Paul, F.; Brandt, A.U.; Schmitz-Hübsch, T. Accuracy and Reliability of the Kinect Version 2 for Clinical Measurement of Motor Function. *PLoS ONE* **2016**, *11*, e166532, doi:10.1371/journal.pone.0166532.
18. Bonnechère, B.; Jansen, B.; Salvia, P.; Bouzahouene, H.; Omelina, L.; Moiseev, F.; Sholukha, V.; Cornelis, J.; Rooze, M.; Van Sint Jan, S. Validity and reliability of the Kinect within functional assessment activities: Comparison with standard stereophotogrammetry. *Gait Posture* **2014**, *39*, 593–598, doi:10.1016/j.gaitpost.2013.09.018.
19. Shu, J.; Hamano, F.; Angus, J. Application of extended Kalman filter for improving the accuracy and smoothness of Kinect skeleton-joint estimates. *J. Eng. Math.* **2014**, *88*, 161–175, doi:10.1007/s10665-014-9689-2.
20. Cao, Z.; Hidalgo Martinez, G.; Simon, T.; Wei, S.E.; Sheikh, Y.A. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, doi:10.1109/tpami.2019.2929257.
21. Insafutdinov, E.; Pishchulin, L.; Andres, B.; Andriluka, M.; Schiele, B. DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model. In *European Conference on Computer Vision Computer Vision—ECCV 2016*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2016; Volume 9910.
22. Groos, D.; Ramampiaro, H.; Ihlen, E.A. EfficientPose: Scalable single-person pose estimation. *Appl. Intell.* **2020**, doi:10.1007/s10489-020-01918-7.
23. Mathis, A.; Mamidanna, P.; Cury, K.M.; Abe, T.; Murthy, V.N.; Mathis, M.W.; Bethge, M. DeepLabCut: Markerless Pose Estimation of User-Defined Body Parts with Deep Learning. *Nat. Neurosci.* **2018**, *21*, 1281–1289, doi:10.1038/s41593-018-0209-y.
24. Mathis, A.; Schneider, S.; Lauer, J.; Mathis, M.W. A Primer on Motion Capture with Deep Learning: Principles, Pitfalls, and Perspectives. *Neuron* **2020**, *108*, 44–65, doi:10.1016/j.neuron.2020.09.017.
25. Howe, T.E.; Rochester, L.; Neil, F.; Skelton, D.A.; Ballinger, C. Exercise for improving balance in older people. *Cochrane Database Syst. Rev.* **2011**,*11*, doi:10.1002/14651858.cd004963.pub3.
26. Da Gama, A.; Fallavollita, P.; Teichrieb, V.; Navab, N. Motor Rehabilitation Using Kinect: A Systematic Review. *Games Health J.* **2015**, *4*, 123–135, doi:10.1089/g4h.2014.0047.
27. Vicon Motion Systems Limited. Plug-In Gait Reference Guide. Available online: https://usermanual.wiki/Document/Plugin20Gait20Reference20Guide.754359891/amp (accessed on 13 May 2020).
28. Zhu, Y.; Zhao, Y.; Zhu, S.C. Understanding Tools: Task-Oriented Object Modeling, Learning and Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
29. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–26 June 2009.
30. McCrum, C.; Gerards, M.H.; Karamanidis, K.; Zijlstra, W.; Meijer, K. A systematic review of gait perturbation paradigms for improving reactive stepping responses and falls risk among healthy older adults. *Eur. Rev. Aging Phys. Act.* **2017**, *14*, 3, doi:10.1186/s11556-017-0173-7.
31. Sherrington, C.; Fairhall, N.J.; Wallbank, G.K.; Tiedemann, A.; Michaleff, Z.A.; Howard, K.; Clemson, L.; Hopewell, S.; Lamb, S.E. Exercise for preventing falls in older people living in the community. *Cochrane Database Syst. Rev.* **2019, ***1*, doi:10.1002/14651858.CD012424.pub2.
32. Sarafianos, N.; Boteanu, B.; Ionescu, B.; Kakadiaris, I.A. 3D Human pose estimation: A review of the literature and analysis of covariates. *Comput. Vis. Image Underst.* **2016**, *152*, 1–20, doi:10.1016/j.cviu.2016.09.002.
33. van Diest, M.; Lamoth, C.C.; Stegenga, J.; Verkerke, G.J.; Postema, K. Exergaming for balance training of elderly: State of the art and future developments. *J. NeuroEng. Rehabil.* **2013**, *10*, 101, doi:10.1186/1743-0003-10-101.
34. Capecci, M.; Ceravolo, M.G.; Ferracuti, F.F.; Iarlori, S.; Longhi, S.; Romeo, L.; Russi, S.N.; Verdini, F. Accuracy evaluation of the Kinect v2 sensor during dynamic movements in a rehabilitation scenario. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society—EMBS, Orlando, FL, USA, 16–20 August 2016; pp. 5409–5412, doi:10.1109/EMBC.2016.7591950.
35. Xu, X.; McGorry, R.W. The validity of the first and second generation Microsoft Kinect for identifying joint center locations during static postures. *Appl. Ergon.* **2015**, *49*, 47–54, doi:10.1016/j.apergo.2015.01.005.

36. Chen, Y.; Tian, Y.; He, M. Monocular human pose estimation: A survey of deep learning-based methods. *Comput. Vis. Image Underst.* **2020**, *192*, 102897, doi:10.1016/j.cviu.2019.102897.
37. Sun, X.; Xiao, B.; Wei, F.; Liang, S.; Wei, Y. Integral Human Pose Regression. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 529–545.
38. Arnab, A.; Doersch, C.; Zisserman, A. Exploiting temporal context for 3D human pose estimation in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 3395–3404.
39. Mehta, D.; Sotnychenko, O.; Mueller, F.; Xu, W.; Elgharib, M.; Fua, P.; Seidel, H.P.; Rhodin, H.; Pons-Moll, G.; Theobalt, C. XNect. *ACM Trans. Graph.* **2020**, *39*, 1–24, doi:10.1145/3386569.3392410.
40. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1325–1339, doi:10.1109/TPAMI.2013.248.
41. Nakano, N.; Sakura, T.; Ueda, K.; Omura, L.; Kimura, A.; Iino, Y.; Fukashiro, S.; Yoshioka, S. Evaluation of 3D Markerless Motion Capture Accuracy Using OpenPose With Multiple Video Cameras. *Front. Sport. Act. Living* **2020**, *2*, 50, doi:10.3389/fspor.2020.00050.
42. Della Croce, U.; Leardini, A.; Chiari, L.; Cappozzo, A. Human movement analysis using stereophotogrammetry Part 4: Assessment of anatomical landmark misplacement and its effects on joint kinematics. *Gait Posture* **2005**, *21*, 226–237, doi:10.1016/j.gaitpost.2004.05.003.
43. Mündermann, L.; Corazza, S.; Andriacchi, T. The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications. *J. Neuroeng. Rehabil.* **2006**, *3*, 1–11.

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapter 9

Paper II

Assessment of Machine Learning Models for Classification of Movement Patterns During a Weight-Shifting Exergame

Authors: Elise Klæbo Vonstad, Beatrix Vereijken, Kerstin Bach, Xiaomeng Su, Jan Harald Nilsen.

Published in IEEE Transactions on Human-Machine Systems. Volume 51, Issue 3, 2021. DOI: 10.1109/THMS.2021.3059716

Assessment of Machine Learning Models for Classification of Movement Patterns During a Weight-Shifting Exergame

Elise Klæbo Vonstad, Beatrix Vereijken, Kerstin Bach, Xiaomeng Su, and Jan Harald Nilsen

***Abstract*—In exercise gaming (exergaming), reward systems are typically based on rules/templates from joint movement patterns. These rules or templates need broad ranges in definitions of correct movement patterns to accommodate varying body shapes and sizes. This can lead to inaccurate rewards and, thus, inefficient exercise, which can be detrimental to progress. If exergames are to be used in serious settings like rehabilitation, accurate rewards for correctly performed movements are crucial. This article aims to investigate the level of accuracy machine learning/deep learning models can achieve in classification of correct repetitions naturally elicited from a weight-shifting exergame. Twelve healthy elderly (10F, age 70.4 SD 11.4) are recruited. Movements are captured using a marker-based 3-D motion-capture system. Random forest (RF), support vector machine, k-nearest neighbors, and multilayer perceptron (MLP) are the employed models, trained and tested on whole body movement patterns and on subsets of joints. MLP and RF reached the highest recall and F1-score, respectively, when using combined data from joint subsets. MLP recall range are 91% to 94%, and RF F1-score range 79% to 80%. MLP and RF also reached the highest recall and F1-score in each joint subset, respectively. Here, MLP ranged from 93% to 97% recall, while RF ranged from 73% to 80% F1-score. Recall results, show that >9 out of 10 repetitions are classified correctly, indicating that MLP/RF can be used to identify correctly performed repetitions of a weight-shifting exercise when using full-body data and when using joint subset data.**

***Index Terms*—Classification, exergaming, machine learning, movement patterns, movement quality, reward systems, weight-shifting.**

I. Introduction

WITH the overall rise in gamification in recent years, serious games have been employed in a wide variety of fields, including education [1], professional training [2], [3], cognitive training [4], and physical exercise (e.g., [5]). Gamification refers to the introduction of elements from gaming, such as goals, reward systems, and challenges, into ordinary tasks to make them more fun and thereby increase motivation and adherence [6]. An essential element when designing serious games is how to determine whether the player's answer or action is correct and thus should be rewarded in the games. Typically, serious games predefine correct answers or actions, and track the performance of players directly using controllers, keyboards, or smartphones, allowing for relatively straight-forward checks of correctness. In serious games for exercise ("exergaming"), the player is interacting with the game using bodily movements that are captured by cameras or other devices [5]. Movements are subsequently assessed against predefined decision rules or thresholds, as seen in e.g., [7], [8], and rewards are given if these body parts performed as predefined, regardless of the correctness of the movements of other parts of the body.

As commercial exergames aim at being entertaining and easy to use, broad ranges and definitions of what is considered "correct" by the game are necessary to accommodate different body shapes and sizes. Because of these broad definitions, players often figure out quickly what the minimum required behavior is for receiving rewards [9]. When the game rewards the player even when performing the movements in this manner, players can easily cheat, or worse, not even know whether they were performing the movements correctly or incorrectly. For entertainment purposes, this may well be irrelevant. However, in the context of regaining or maintaining physical function, performing the correct movements is essential for effectiveness and progress [10]. Effective exercise depends on performing the necessary movements correctly, thus supervised exercise programs typically report better results than nonsupervised exercise programs [11].

For older adults, exergaming is regarded as a promising tool to deliver guided exercise without the presence of therapists or clinicians. Furthermore, exercise delivered through exergames has been shown to be more fun and motivating than traditional exercise [5], [12]. This could help increase adherence and motivation for exercising in older adults, which is a prerequisite for mediating the strain the ongoing demographic change will place on our health care systems [13]. Older adults often have different requirements for movements during exercise compared to healthy people, as they might have physical constraints due to ageing [14]. ColorRules and settings in exergames therefore need to be adapted to individual constraints and goals, but still allow for proper form and tempo to progress in training [15]. If

Manuscript received May 29, 2020; revised November 12, 2020 and January 19, 2021; accepted January 28, 2021. Date of publication; date of current version. This article was recommended by Associate Editor M. S. Neubert. *(Corresponding author: Elise Klæbo Vonstad.)*

Elise Klæbo Vonstad, Kerstin Bach, Xiaomeng Su, and Jan Harald Nilsen are with the Department of Computer Science, Norwegian University of Science and Technology, 7034 Trondheim, Norway (e-mail: elise.k.vonstad@ntnu.no; kerstin.bach@ntnu.no; xiaomeng.su@ntnu.no; jan.h.nilsen@ntnu.no).

Beatrix Vereijken is with the Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, 7034 Trondheim, Norway (e-mail: beatrix.vereijken@ntnu.no).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2021.3059716.

Digital Object Identifier 10.1109/THMS.2021.3059716

exergames are to be effective in serious exercise settings such as rehabilitation, we need game systems that accurately identify and reward correctly performed movements to ensure efficiency and progress [16], [17].

One alternative to using broadly defined rules and thresholds to determine the correctness of a movement is to study the occurrences of movement patterns with good or poor quality and build models that embody the features of each of these. These models can then be used to assess movement quality, potentially with high accuracy, as the model is trained to recognize features of a correctly performed movement pattern without being fed predefined rules or thresholds. Recent developments in machine learning (ML)/deep learning (DL) have made it possible to efficiently analyze large amounts of data, which is promising for using high-volume data from whole-body movement patterns. Such models have been used successfully to recognize different everyday activities like walking, sitting, and lying down [18], [19], and movement patterns during traditional exercise (e.g., [20]). However, to the best of our knowledge, it has yet to be applied to assessment of movement pattern quality during exergaming.

A. *Pilot Study*

To study the suitability and potential of applying ML for our objective, we conducted a pilot study first to investigate whether ML models can distinguish between similar full-body movement patterns where some are performed correctly and others incorrectly [21]. In this pilot study, participants (N = 11, 6 F, mean age 69.3 years, SD 4.0) performed repetitions of weight-shifting movements where half of the movements were performed with clear incomplete weight shifts (i.e., incorrectly performed repetitions), and the other half with clear complete weight-shifts (i.e., correctly performed repetitions). Participants were instructed on how to perform the movements to ensure that the right movement patterns for incorrect and correct repetitions were recorded. A marker-based 3-D motion capture system (3DMoCap) was used to track participants' movements, and statistical features were calculated for each repetition. Three different ML models [Random forest (RF), support vector machine (SVM), and K-nearest neighbor (K-NN)] were trained and evaluated for classification performance using leave-one-group-out (LOGO) cross-validation. All three models achieved good performance (>90% accuracy, [18]). These results encouraged us to investigate whether ML models can accurately classify movements that are *naturally* elicited (i.e., not instructed) from a balancing exergame. As naturally elicited movements are more varied, both within and across participants, classification can be more challenging.

B. *Aim of This Article*

The present aticle investigates what level of F1-score and recall four different ML/DL models can achieve in classifying correctly performed whole-body and joint-subset movement patterns naturally performed during a balance exergame.

C. *Article Organization*

This article is organized as follows. Related work is outlined in Section II. The experimental set-up and data analysis procedures are described in Section III. Section IV presents results comparing four different ML models in the classification of movement correctness. Discussion of the results and limitations of the study are presented in Section V. Conclusion and future work are presented in Section VI.

II. Related Work

In general, exergaming for older adults is considered a promising tool for facilitating unsupervised exercise at home or in an elderly care center (e.g., [4], [22], [23]). Research has shown that exergames are effective in delivering exercise for several physical and mental functions, such as balance and postural control [24], gait [25], upper body movements [12], cognitive function [26], problem solving [27], and memory [28]. Exergames are also found to be more motivating and fun than traditional exercise [9], [29], which is an essential feature that could facilitate adherence and motivation for exercise [12]. In addition, the technologies that exergames are based on make it possible to tailor games to individual needs and goals [30], which is a major advantage that could make exergaming even more effective than traditional exercise. Furthermore, to ensure that exergames are appropriate for older adults, extensive research has been conducted into the design and usability requirements for this population, resulting in guidelines and design principles that apply to exergames for older adults [16], [31].

In recent years, there has been a proliferation of work implementing the (semi)automatic classification and recognition of actions and activities based on multimodal data recorded from human movement [18]. Although research on movement classification, as shown in [18], is an adjacent field of research, these models only focus on identifying *what* movement has been performed, not the *quality* of the movement (e.g. how well the movement was performed). We are particularly interested in assessing the quality of movement and will therefore focus primarily on related work that sheds light on evaluating movement quality.

High-quality research has been conducted with the aim of identifying errors in movement patterns compared to predefined movement templates [32]–[35], and rules/thresholds [7], [8], [36]. Movement performance compared to the predefined goal is used to provide feedback on how to improve movement patterns. Comparison of movements to thresholds and/or rules is also done in comprehensive work on modelling and evaluation of human movement, as seen in [15], [14], and [37].

Using template movements and decision rules can be appropriate for players that do not have physical constraints or do not need individual adaptation of movement patterns during exercise and are aiming to perform the exercises similarly to a healthy person. As mentioned, participants need to have goals that are adjusted to their needs and constraints, so comparing their movements to a healthy person or a template movement can be detrimental to motivation or might push them to perform the exercise outside their safe limits.

One earlier study also aimed to classify movement quality in a more naturally elicited, less instructed, fashion [38]. Here, exercise repetitions near exhaustion were used as examples of incorrectly performed movements and classified as correct or incorrect using ML models. This study was conducted on healthy children, using a smartphone (i.e, an inertial measurement unit) to capture movements.

In conclusion, we find that there is a wide variety of settings and contexts where automatic identification of movement errors during exercise is receiving attention, including technique analysis in general fitness and elite sports, as well as exercising for elderly at home or in rehabilitation centers. However, research into classification of movement quality specifically during exergaming is scarce, especially regarding identification of correctly and incorrectly performed movements.

Further, a large body of the related work demonstrated that errors in movement patterns can be identified during exercise by comparing performed movements to rules and template movements or expert scoring. Conversely, our study aims to build ML/DL models that can classify correctly performed movements that are naturally elicited, without comparing to a template movement or a set of rules or thresholds. Then, we assess the accuracy with which these models can identify correctly performed movements in unseen samples of the movement patterns.

III. Experimental Setup and Analysis: Assessing Movement Patterns Using ML

1) Participants: Participants were healthy older adults recruited from local exercise groups in the municipality. All participants gave their written, informed consent. There were 12 participants in total (10F); average age was 70.4 (SD 11.4) years (range 54–92). Average height and weight were 172.3 (SD 11.4) cm and 70.4 (SD 12.1) kg, respectively. Exclusion criteria were physical or cognitive injuries/impairments that affected their balance and gait ability, and age <50 or age >80 years. The project was approved by the Norwegian Regional Ethics Committee and the Norwegian Centre for Research Data (REK case number: 2017/2078-1).

2) Experimental Protocol: The experiment was conducted at the Motion Capture and Visualization Laboratory (“Vislab”) at NTNU Trondheim in June 2019. A marker-based 3-D motion capture (3-DMoCap) system was used to measure participants’ movements for use in analysis and classification. Four cameras (MX400, 90 Hz, Qualisys AB) were used. Thirty-six reflective markers were placed following the Plugin-Gait (PiG) marker placement protocol [39], excluding head and fingers.Two digital video cameras (Hero 3+ Black, 25 Hz, 1080p, GoPro Inc) captured movements in the sagittal and frontal planes of the player. Two 3-axial force plates (1000 Hz, 600x400x35 mm, Kistler Nordic AB) were located under the participants’ feet to measure the ground reaction forces while playing. A platform matching the force plates’ height was placed laterally of each force plate. The experimental setup can be seen in Fig. 1.

3) Game System: The game was built in Unity (v. 5, Unity Technologies, Denmark). As time-of-flight camera technology is commonly used in exergaming [5], [40], we used the Kinect v2 (30 Hz, Microsoft Inc), set up in front of the participants, to enable gameplay. The participants played three rounds of the two parts of the game, totaling six trials for each participant. If the movement tracking from the Kinect was not satisfactory, for example when the avatar did not follow the participants’ movements, avatar movements were jittery, or if the sensor failed to identify the player at all, the trial was stopped and started again until smooth, continuous movement tracking from the Kinect was achieved.

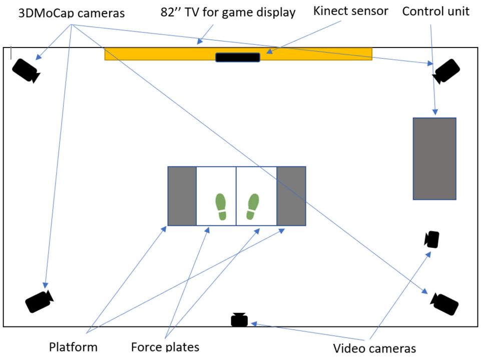


Fig. 1. Experimental setup.

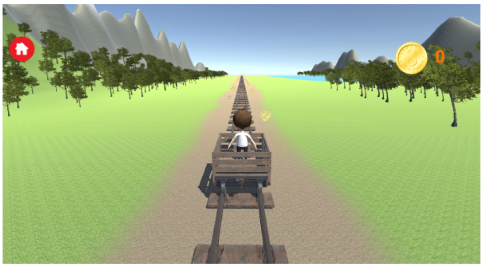
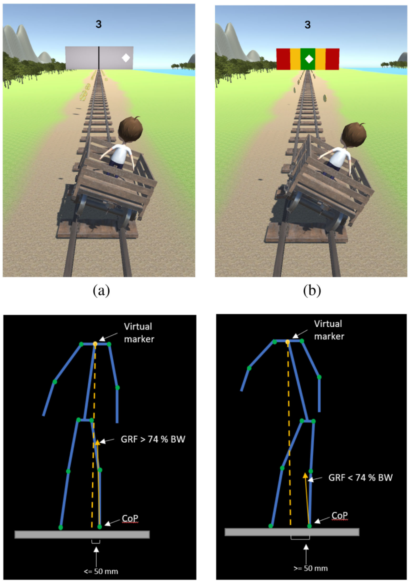


Fig. 2. Game interface.

The two parts of the game were designed to elicit different movement patterns from the players: the first aimed at having the player perform a complete, and thus correct, weight shift by moving their upper body over their weight-bearing foot. The second part was designed to make the player perform movements without moving their upper body over the weight-bearing foot, i.e., incompletely performed weight shifts. The game interface consisted of a rail cart with an avatar in it, representing the player, as shown in Fig. 2. On each side of the rail were coins which the player would try to hit with the cart as they moved along the rail. The cart tilted from side to side, following the medio-lateral leaning movements of the player. There were never more than two coins successively, and the coins appeared in random places for each participant. There were a total of approximately 100 coins in each game part, with approximately 50% of the coins on each side of the rail. The player was rewarded with points if they hit a coin with the cart, and the position of their upper body decided the amount of points rewarded in each of the game parts. There was a bar above the avatar. In part 1 the bar was grey, in part 2 the bar was multicolored as seen in Fig. 3. The grey



(a) (b)

(c) (d)

Fig. 3. (a) and (b) Two versions of the exergame. (c) and (d) Typical body postures when playing the two different exergame versions. (a) Part 1: Two-split grey bar, shown at the end of the track, with the star to the right of the dividing line, rewarding 3 points. (b) Part 2: Three-split color bar, shown at the end of the track, with the star in the middle 33%, rewarding 3 points. (c) Typical body posture when being rewarded 3 points in part 1 of the game. Here, the player is leaning their upper body over their weight-bearing foot, resulting in the distance between the virtual marker and the CoP of the weight-bearing foot being <50 mm, and the GRF Z-component being >74% of body weight. BW = body weight. GRF = ground reaction force. CoP = center pressure. (d) Typical body posture when being rewarded 3 points in part 2 of the game. Here, the player is not leaning their upper body over their weight-bearing foot, resulting in the distance between the virtual marker and the CoP of the weight-bearing foot being >50 mm, and the GRF z-component being <74% of body weight. GRF = ground reaction force. CoP = center of pressure.

bar was divided in the middle: if the star was on the line when a coin was hit, the player was rewarded 1 point. Three points (max score) were awarded if the star was as far away from the dividing line as possible, i.e., at any of the lateral parts of the grey bar as seen in Fig. 3(a). The multicolored bar was divided into three equally sized color fields: green in the middle 33%, yellow in the next 33% on each side, and red at the 33% most lateral fields. The red field rewarded 1 point, the yellow two points and the green three points, as seen in Fig. 3(b). Fig. 3(c) shows a typical posture form playing version 1, and Fig. 3(d) shows a typical posture from playing version 2 of the game.

4) Preprocessing: Joint center locations of shoulders (SHO), hips (HIP), knees (KNE) and ankles (ANK), as well as center of pressure (CoP), were extracted from the standard PiG biomechanical model from each of the six game trials for all participants. Game trials were then segmented into single mediolateral movement repetitions using the peak-finding algorithm peakutils (v 1.3.3 for Python) on the y-axis of the right SHO joint in the Qualisys coordinate system. One repetition was defined as a continuous movement starting at the most lateral point of a medio-lateral movement, ending at the most lateral point on the opposite side. Python for Windows (v. 3.8.2) was used for all analyzes. An overview of the data analysis pipeline can be seen in Fig. 4.

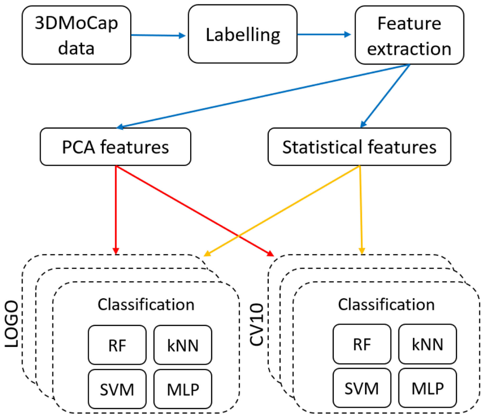


Fig. 4. Data analysis pipeline. The process, from “Feature extraction,” was repeated for all joint data combined, and for each joint subset separately. PCA = Principal component analysis, RF = random forest, SVM = support vector machine, kNN = k-nearest neighbor, MLP = multilayer perceptron, LOGO = leave-one-group-out, CV10 = tenfold cross-validation.

5) Labeling: The repetitions were subsequently assessed for the weight shift being correctly (i.e., a complete weight shift) or incorrectly (i.e., an incomplete weight-shift) performed. A physical therapist experienced in rehabilitation was consulted to determine the features of a correctly performed weight shift. The following criteria had to be met for a repetition to be deemed a correct weight shift. 1) The majority of the persons’ body weight (over 74%, as 50% on each foot means that the person is standing with equal amount of weight on their feet) must be shifted to the weight-bearing foot. 2) The upper body must be moved over the weight-bearing foot as the weight is shifted. To evaluate whether condition 2 was met, a virtual marker was calculated as the 3-D midpoint between the left and right SHO, and the distance between the y-position of this virtual marker and the y-position of the CoP was calculated. Mean distance of <50 mm was required for the repetition to be deemed correctly performed. Sample videos form all participants were consulted to ensure that these criteria captured actual incorrectly and correctly performed movement patterns. All repetitions were assessed according to these criteria and assigned a target variable for incorrect (0) or for correct performance (1). This resulted in 2821 repetitions, where 1803 were labeled 1 (correct) and 1018 0 (incorrect).

6) Feature Extraction: After the target labels were assigned, statistical features were extracted from each repetition using the TSfresh library [41] (v. 12.0) for Python. See Appendix 1 for an exhaustive list of features. Furthermore, the feature dimensions were reduced using principal component analysis (PCA). Principal components that combined explained 95% of variance in the data were retained for further analysis.

7) Classification Models and Hyperparameter Tuning: Four models were employed in this study: RF, SVM, kNN, and an artificial neural network [multilayer perceptron (MLP)]. SciKit-Learn (version 22.1) for Python was used for analysis. RF is an

TABLE I
HYPERPARAMETER VALUES FOUND TO ACHIEVE THE BEST ACCURACY FROM GRIDSEARCHCV. RF = RANDOM FOREST, SVM = SUPPORT VECTOR MACHINE, kNN = K-NEAREST NEIGHBOR, MLP = MULTILAYER PERCEPTRON, LOGO = LEAVE-ONE-GROUP-OUT, CV10 = TENFOLD CROSS-VALIDATION

	Hyperparameter 1	Hyperparameter 2	Hyperparameter 3
	Name = Value	Name = Value	Name = Value
RF	Criterion = Entropy	Min. leaf = 4	Min.samples at split = 10
SVM	C = 0.01	Gamma = 0.01	
KNN	Leaf Size = 15	Metric = Manhattan	No. neighbors = 35
MLP	#Hidden layers = 50	Alpha = 0.01	

ensemble classifier that employs a set of decision trees to predict class labels, where each tree sees a random subset of features, and uses the majority class predicted by each tree's leaf nodes to classify a sample. Ensemble classifiers have been used successfully in similar work on movement quality (e.g, [42]) and in adjacent fields such as action classification [18], [19]. SVM is a linear model that finds the optimal line (or hyperplane) to separate classes, using the line/hyperplane that yields the largest support vectors (i.e., decision boundaries) between classes. SVM is often used in action recognition, as it is a powerful classifier [18], [19]. The kNN model evaluates the *(k)* nearest data points' class for each feature and classifies the sample based on the majority of these neighbors' class. kNN is a fast and simple, yet powerful classifier that has been used in adjacent work [15], [43]. MLP is a layered network of nodes that classifies samples based on activation of nodes in the "hidden" layers between the input and the output layer, using backpropagation to adjust weights and biases in the hidden layer nodes for each iteration of training. MLP requires more training data and processing power than ML methods, but often outperforms ML methods in action classification when provided with sufficient training data [18].

The optimal combination of hyperparameter tunings for each model [44], with regard to classification accuracy, was found using grid search (threefold CV) from the SciKit-Learn-pckage. Table I shows the hyperparameter tunings (that are not default for the models in the current SciKit-Learn version) that achieved the highest accuracy for each model. These hyperparameter tunings were used in subsequent analyzes.

8) Cross-Validation and Classification Procedures: The models were trained and tested using cross-validation (CV) by LOGO, and tenfold CV (CV10). LOGO entails training the model on all the data except one participant and using this participants' data as the testing set. CV10 creates ten random subsets of the data from all participants and holds one subset out for testing in each iteration. To simulate a situation where only subsets of joints are reliably tracked, each model was also trained and tested in the same manner by using only subsets of joint data, i.e., only ankle data, knee data, hip data, or shoulder data. Thus, all models were trained and tested on 20 different versions of the data set as seen in the last step of Fig. 4.

9) Evaluation: Model performance was evaluated using the F1-score and the recall. F1-score is an accuracy measure (the harmonic mean between precision and recall), which gives more useful insight into model performance in an imbalanced dataset than standard accuracy [45]. Recall, or sensitivity, is the true positive rate and describes the ratio of correctly identified positive samples out of all sampled classified as positive by the model. This is a useful measure as it says how many of the correctly performed repetitions were actually labeled as correct, i.e., how many of the correct repetitions a model identified as a correct repetition.

TABLE II
PERCENT F1-SCORE ACHIEVED ON JOINT SUBSETS [SHOULDER (SHO), HIP (HIP), KNEE (KNE), AND ANKLE (ANK) JOINTS]. MODELS ARE RANDOM FOREST (RF), SUPPORT VECTOR MACHINE (SVM), K-NEAREST NEIGHBOR (kNN), AND ARTIFICIAL NEURAL NETWORK (MLP). THE FEATURE REPRESENTATIONS (FEATS) ARE STATISTICAL (STAT) AND PCA (PRINCIPAL COMPONENTS). CV = CROSS-VALIDATION: LOGO = LEAVE-ONE-GROUP-OUT, CV10 = TENFOLD. SD = STANDARD DEVIATION. M = MEAN. THE HIGHEST AVERAGE RECALL ACHIEVED BETWEEN JOINT SUBSETS (COLUMNS), AND HIGHEST AVERAGE BETWEEN THE MODELS (ROWS) ARE HIGHLIGHTED IN BOLD FONT. THE HIGHEST RECALL ACHIEVED WITHIN JOINT SUBSETS IS HIGHLIGHTED IN GREEN

	FEATS	CV	SHO (SD)	HIP (SD)	KNE (SD)	ANK (SD)	M(SD)
RF	Stat	LOGO	79.7 (10.8)	78.9 (11.4)	69.9 (18.5)	73.1 (10.0)	75.4 (12.7)
		CV10	79.2 (10.5)	77.1 (13.0)	75.9 (12.3)	74.8 (11.0)	77.4 (11.7)
	PCA	LOGO	77.3 (9.3)	76.7 (10.3)	76.2 (9.3)	75.3 (10.8)	76.8 (9.9)
		CV10	77.0 (10.2)	76.5 (9.9)	77.6 (9.9)	75.9 (9.2)	77.1 (9.8)
SVM	Stat	LOGO	76.3 (11.1)	68.0 (17.2)	64.7 (17.2)	64.7 (17.2)	70.0 (15.7)
		CV10	77.9 10.7)	72.8 (14.0)	72.5 (14.9)	72.5 (14.8)	74.7 (13.6)
	PCA	LOGO	76.0 (10.7)	67.5 (18.4)	64.5 (17.0)	61.9 (20.4)	69.2 (16.6)
		CV10	77.4 (10.7)	72.4 (13.4)	72.6 (14.5)	69.1 (11.7)	73.8 (12.6)
KNN	Stat	LOGO	79.8 (10.3)	76.9 (12.5)	75.9 (9.1)	75.9 (9.1)	77.7 (10.2)
		CV10	79.2 (8.5)	75.9 (9.8)	75.2 (11.0)	75.2 (11.9)	77.0 (10.1)
	PCA	LOGO	78.9 (10.0)	77.5 (11.6)	77.1 (8.7)	74.2 (9.2)	77.3 (9.9)
		CV10	78.0 (9.0)	77.0 (9.6)	76.2 (10.1)	75.7 (9.5)	77.0 (9.6)
MLP	Stat	LOGO	79.6 (10.0)	78.1 (10.1)	74.7 (12.7)	77.5 (10.8)	77.8 (10.9)
		CV10	79.9 (8.7)	77.7 (8.7)	76.2 (9.7)	76.6 (9.0)	**78.2 (9.0)**
	PCA	LOGO	79.7 (10.5)	77.9 (9.9)	76.3 (9.3)	77.5 (10.3)	78.1 (10.0)
		CV10	79.3 (8.2)	77.3 (9.2)	77.4 (9.5)	77.2 (8.9)	78.0 (8.9)
M(SD)			**78.4 (1.3)**	75.5 (3.4)	73.9 (4.0)	73.4 (4.4)	

TABLE III
PERCENT RECALL ACHIEVED ON JOINT SUBSETS [SHOULDER (SHO), HIP (HIP), KNEE (KNE), AND ANKLE (ANK) JOINTS]. MODELS ARE RANDOM FOREST (RF), SUPPORT VECTOR MACHINE (SVM), K-NEAREST NEIGHBOR (kNN) AND ARTIFICIAL NEURAL NETWORK (MLP). THE FEATURE REPRESENTATIONS (FEATS) ARE STATISTICAL (STAT) AND PCA (PRINCIPAL COMPONENTS). CV = CROSS-VALIDATION: LOGO = LEAVE-ONE-GROUP-OUT, CV10 = 10-FOLD. SD = STANDARD DEVIATION. M = MEAN. THE HIGHEST AVERAGE RECALL ACHIEVED BETWEEN JOINT SUBSETS (COLUMNS), AND HIGHEST AVERAGE BETWEEN THE MODELS (ROWS) ARE HIGHLIGHTED IN BOLD FONT. THE HIGHEST RECALL ACHIEVED WITHIN JOINT SUBSETS IS HIGHLIGHTED IN GREEN

	FEATS	CV	SHO (SD)	HIP (SD)	KNE (SD)	ANK (SD)	M (SD)
RF	Stat	LOGO	87.5 (11.2)	88.3 (10.3)	79.1 (26.0)	82.9 (14.7)	84.4 (15.6)
		CV10	84.8 (8.9)	82.2 (14.1)	82.8 (13.9)	82.0 (12.3)	82.9 (12.3)
	PCA	LOGO	89.4 (7.3)	89.0 (8.4)	87.9 (14.8)	88.4 (10.4)	88.7 (10.2)
		CV10	87.7 (7.2)	87.5 (8.1)	89.2 (7.9)	88.5 (9.7)	88.2 (8.2)
SVM	Stat	LOGO	78.2 (14.4)	66.5 (21.9)	29.2 (64.7)	67.4 (29.2)	60.3 (32.5)
		CV10	79.0 (11.0)	70.6 (16.7)	73.5 (18.1)	73.5 (18.1)	74.2 (16.0)
	PCA	LOGO	77.8 (14.0)	66.0 (22.6)	66.6 (28.9)	61.5 (25.5)	68.0 (22.8)
		CV10	78.3 (11.2)	70.0 (15.9)	73.6 (18.0)	66.4 (12.7)	72.1 (14.4)
KNN	Stat	LOGO	87.3 (8.6)	83.9 (10.9)	84.5 (10.6)	84.5 (10.6)	85.1 (10.2)
		CV10	86.9 (6.4)	82.0 (8.6)	84.3 (15.1)	84.3 (15.1)	84.4 (11.3)
	PCA	LOGO	87.2 (8.6)	86.8 (8.2)	88.1 (10.0)	84.3 (10.3)	86.6 (9.3)
		CV10	85.9 (6.9)	85.4 (7.8)	86.5 (12.4)	88.1 (8.8)	86.5 (9.0)
MLP	Stat	LOGO	92.1 (7.9)	95.7 (5.1)	87.0 (19.4)	93.3 (8.6)	92.0 (10.3)
		CV10	90.8 (6.0)	94.9 (4.3)	89.0 (10.9)	92.7 (8.7)	91.8 (7.5)
	PCA	LOGO	94.5 (5.8)	96.3 (3.8)	90.9 (11.7)	96.4 (5.5)	94.5 (6.7)
		CV10	94.1 (4.4)	96.5 (3.0)	93.4 (6.6)	96.4 (3.2)	**95.1 (4.3)**
M(SD)			**86.3 (5.3)**	83.9 (10.1)	80.3 (14.9)	83.2 (10.4)	

IV. RESULTS

Results for *each joint subset* are presented with F1-score in Table II and recall in Table III. Results for *joint subsets combined* are shown with F1-score in Fig. 5 and recall in Fig. 6.

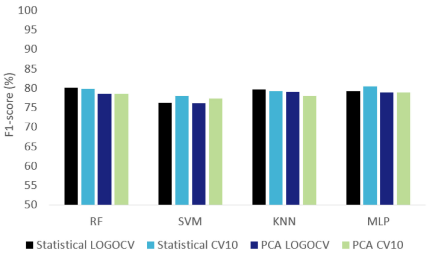


Fig. 5. F1-score achieved using different feature representations and CV methods on all joint subsets combined. RF = random forest, SVM = support vector machine, kNN = k-nearest neighbor, MLP = multilayer perceptron, LOGO = leave-one-group-out, CV10 = tenfold cross-validation. PCA = principal components.

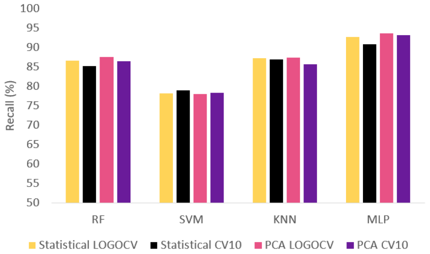


Fig. 6. Recall achieved using different feature representations and CV methods on all joint subsets combined. RF = random forest, SVM = support vector machine, kNN = k-nearest neighbor, MLP = multilayer perceptron, LOGO = leave-one-group-out, CV10 = tenfold cross-validation. PCA = principal components.

A. F1-Score

The four models reached different levels of F1-score on different joint subsets of the data. Table II shows the F1-scores for each subset of joints in classifying correct repetitions, as well as the average performance of each joint subset. All models achieved similarly good results, with a mean of 75.3% (SD 11.3) for the F1-score. RF slightly outperformed other models on hip and knee joint subsets, while MLP performed best on shoulder and ankle joint subsets. Overall, the performance variation in using different feature representations or cross-validation methods was small. Somewhat surprisingly, the SVM achieved the overall lowest performance in terms of F1-score. All joint subsets also had high average F1-scores, with over 70%, but the SHO subset achieved the highest average with 78.4% (SD 1.3).

Fig. 5 shows the F1-score achieved by using all joint subsets combined, using different feature representations and cross-validation methods. These are results from all joint data only F1-score on joint subsets can be seen Appendix 2. Results show good performance from all models, with 78.5% (1.3 SD) F1-score on average. Different feature representations and cross-validation models are not affecting performance to any noteworthy degree.

B. Recall

Table III shows the recall achieved by the models on joint subsets of the data using different feature representations and CV methods. On average, the models achieved 83.3% (SD 17.6) recall (see Table III). The MLP outperformed the SVM and kNN models by 10%–25%, and was around 10% better than the RF model. Lowest recall was by SVM on the knee joint subset with statistical features and LOGO CV, with only 29.2%. On average, the SHO joint subset achieved the highest recall with 86.3% (SD 8.7) but other joint subsets also achieved high recall with >80%.

Fig. 6 shows the recall achieved by different feature representations and cross-validation methods. These are results from all joint subsets combined joint subset recall results can be seen Appendix 2. The MLP slightly outperformed the other models, with an excellent average of 92.6% (SD 1.1) recall. RF and KNN achieved comparable results, with an average of 86.5% (SD 0.8) recall and 86.9% (SD 0.7) recall, respectively. SVM was the overall lowest performing model in recall of correct repetitions, with an average of 78.4% (SD 0.3). Feature representation and CV methods showed only small differences, but PCA with LOGO was the marginally best configuration in three out of four models.

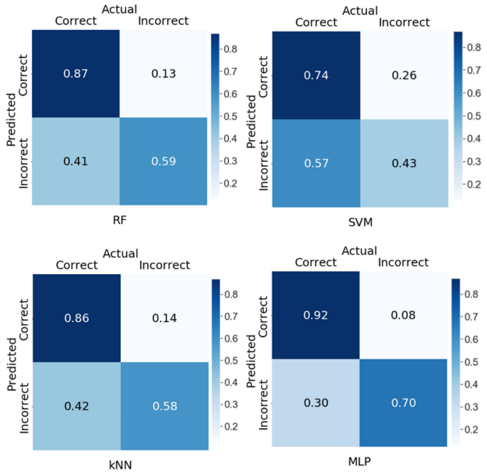


Fig. 7. Confusion Matrices for all models, with ratios of (clockwise from top left) true positive, false positive, true negative, and false negative predictions. Darker blue = higher ratio of samples predicted to belong in quadrant. Going clockwise from top left, the panels are for random forest (RF), support vector machine (SVM), multilayer perceptron (MLP), and k-nearest neighbor (kNN).

C. Classification of Incorrect Repetitions

Even though classification of correctly performed weight-shift repetitions may be sufficient for many applications, being able to accurately identify incorrect repetitions is important in a feedback perspective. An exergame system often needs to be able to identify e.g. an incomplete weight shift, and provide feedback to the player on how the movement pattern can be adjusted to achieve a complete weight shift. We analyzed the current models' ability to identify samples labeled as incorrect. This is not captured in metrics such as F1-score and recall, as they attenuate the influence of true negative samples. As seen in Fig. 7, incorrect samples were not classified with as high accuracy as correct samples, although the MLP achieved 70%

TABLE IV
PERFORMANCE OF EACH MODEL AND CROSS-VALIDATION METHOD IN MEAN TIME CONSUMPTION FOR TRAINING AND PREDICTION. RF = RANDOM FOREST, SVM = SUPPORT VECTOR MACHINE, kNN = K-NEAREST NEIGHBOR, MLP = MULTILAYER PERCEPTRON, LOGO = LEAVE-ONE-GROUP-OUT, CV10 = TENFOLD CROSS-VALIDATION

		Training time (s, SD)	Prediction time (ms, SD)
kNN	LOGO	0.8 (0.1)	1738 (476)
	CV10	0.1 (0.1)	153 (15)
SVM	LOGO	9.9 (0.7)	843 (220)
	CV10	0.8 (0.1)	72 (5)
RF	LOGO	8.3 (0.3)	15 (4)
	CV10	2.8 (0.2)	15 (1)
MLP	LOGO	7.0 (1.2)	2 (0.5)
	CV10	3.3 (0.1)	1 (0.5)

accuracy. Overall, models were able to classify about half of the incomplete weight shifts correctly.

V. Discussion and Limitations

In this article, we investigated the level of recall and F1-score the employed ML/DL models achieved in classification of correctly performed weight-shifting exercise repetitions, naturally elicited while playing a balancing exergame.

A. Correct Weight Shifts

Classification of correctly performed whole-body movement patterns is found to be feasible for all models used in this study, arriving at results ranging between 70%–80% F1-score (Table II) and 75%–95% recall (Table III). The best performing models in our study achieve over 90% recall and around 80% F1-score, which demonstrates that these models could be used in real-world applications for medio-lateral balance exercises. Although there are few directly comparable studies, our results show that using MLP or RF for classification of correct repetitions is in line with the state-of-the-art activity classification systems as reported in [18], [19], and [46]. Even though some of the models did not perform at a satisfactory level, we showed that the best performing models are promising in settings where it is useful to be able to receive feedback on movement pattern quality without having a clinician present, such as in home exercise.

The recall achieved by all models show that 90%–95% (see Fig. 6) of the correctly performed repetitions were, in fact, identified as such, which in an exergame situation would imply rewarding the player for close to all correctly performed repetitions. In other words, only a rather low number of correct repetitions were missed by the models. This is an indication that the models accurately captured and represented the movement features of a correct weight shift, without using manually designed rules or thresholds. This work echoes the results in [20] and [46].

The different classification models performed with slightly different results, as seen in Figs. 5 and 6. When it comes to computational performance, the models performing best on average, RF and MLP, were also the most efficient in training and prediction in terms of time usage (see Table IV). kNN was very fast in training, but slowest in prediction with >1.5 s used for each LOGO iteration, which is likely due to kNN having to build the model for each datapoint. As expected, SVM was the slowest in terms of training time, as well as being slow in prediction time. The distance-based models (kNN and SVM) often perform worse in terms of classification accuracy when the number of features is large compared to the number of samples [47], as a complex feature space makes it difficult to define decision boundaries that separate classes. The high MLP performance is likely due to the manner MLP models adjust the weights and biases in an iterative manner for a given classification problem by using gradient descent [48]. As such, MLP models also intuitively assess importance of different features during training. This is similar to what RF models do: features with high importance for the given classification task are used in early splits. Furthermore, features are used in a random fashion in the different decision trees, which contributes to high performance despite a complex feature space. This is also possibly the situation with the current dataset. The overall high recall can be attributed to the high quality of the data; low levels of noise have been shown to improve model performance [18], [49]. These results suggest that RF is likely the model that should be considered in similar applications for the following reasons: 1) RF achieves high recall; 2) RF is considered a “white box”, e.g., it is possible to extract the decision making process in situations where transparency in the decision process is required; 3) the computational cost of prediction in RF is low, especially compared to MLP. These three features are likely of importance for a ML/DL system to be usable in e.g., a clinical or rehabilitation exercise setting. However, as the No Free Lunch theorem suggest, and as is shown in these results, there is no one model that is universally “best” for all problems (e.g., joint subsets). The model that performs best on average might not always be the best performing model in all problem subsets [50]. This indicates that it is necessary to evaluate the specific problem at hand, and how different models perform with the given data types, available computational power and noise level.

Results from the two cross-validation experiments are promising with respect to classification of previously unseen movement patterns. The models' performance did not worsen when classifying movement patterns from a participant that the models were not trained on. This is evident as the LOGO method performs similarly to the CV10 method, which holds out random subsets of all participants' data. Such similarity might be explained in two ways. 1) Participants performed the correct movement patterns similarly. 2) The models were indeed not overfitting, but truly and accurately captured and represented the features for correctly performed movement patterns to a good enough degree to identify unseen data with high accuracy. The practical implication of such models is that people who have not been playing a game using these assessment models before, will receive rewards when performing weight-shifting movements correctly. This is in line with the findings in [18]. Authors of [35] similarly found that using different neural network configurations with LOGO cross-validation produced good results. This further supports our findings that a person can use such a game system even though the employed model for assessing movement pattern quality has not seen his/her movement patterns before.

When looking at results from separate joint subsets, shoulder movement patterns produced the best results in both F1-score and recall. This suggests that the shoulder movement pattern is the most relevant in assessment of weight shifting, and should be included to ensure high classification accuracy. Overall, using joint subsets, our models also achieved a level of performance (about 75% F1-score and 83% recall) comparable to other classification models using joint subsets [18]. One might argue that using any of these joint subsets could provide accurate rewards in weight-shifting exergames. Whole-body movement patterns still achieve slightly better results than joint subsets, both in terms of F1-score and recall, indicating that whole body movement patterns might still be a preferred setup if the primary goal is to achieve the best quality assessment possible. However, if the available tracking method only allows for accurate tracking of subsets of joints, using subsets is nonetheless a worthy alternative (even a preferred one if and when any cost benefit consideration renders the whole body tracking setup unsuitable) as it still achieves a very good classification accuracy of correct movement patterns using those subsets.

Regarding feature representation, there is no clear indication of any of the methods producing superior classification results. This suggests that statistical features are representing the exercise repetitions well, and that the principal components explaining 95% of the variance in the feature data sufficiently represent the latent information in the statistical features. PCA might be preferable over statistical features in future use, as they are lower dimensional and thus more computationally efficient.

B. Incorrect Weight Shifts

Being able to identify and provide feedback on erroneous movement patterns is useful in serious exergaming situations like rehabilitation, as exergames could be used to guide rehabilitation exercises without the presence of a clinician. The player would then need feedback on how to improve their movement pattern (such as having a larger range of motion, or moving faster) in order to perform the exercise in a efficient manner. In earlier work, where samples were labeled by error class, error types were classified with 85%–95% accuracy [42], [51]. The results from classification of incorrect repetitions in the current study support this notion that classification models needs to be trained on erroneous movement patterns that are labeled by error type, in order to construct representations of the error types in the features. Hence, actively classifying incorrect samples should be the goal of classification systems aiming for use in feedback during exergaming in rehabilitation settings. The current dataset does not contain enough samples of different error types, and is therefore not suited for such analyzes. Furthermore, the movement patterns in the erroneous repetitions probably vary significantly between participants, making it challenging to find robust representations of incorrect repetitions in the features. This also indicates that the features in the current study might not capture the information required for the models to represent an incorrect repetition, as some incorrect repetitions might have very similar movement patterns to correct repetitions. Still, the MLP is able to classify incorrect samples with 70% accuracy, as seen in Fig. 7, indicating that DL models might be usable for such tasks in future work.

C. Limitations

There are some limitations to this study that are necessary to keep in mind. Because this study included 12 participants only moving in a single plane, it is important to keep limitations of applicability of our results in mind. The movement performed is restricted to a medio-lateral weight shifting exercise, which is (ideally) confined to movement in the frontal plane of the body, so movements in other planes or in combinations of planes might be more difficult to classify correctly. Even though our results are promising, further research should be conducted to investigate the performance of these ML/DL models in more complex and challenging settings. Furthermore, data from other motion capture tools that are commonly available should be evaluated as this might impact classification performance.

VI. Conclusion

In conclusion, this study shows that RF and MLP are able to identify correctly performed weight-shifting repetitions with high recall and F1-score. In the development of exergame systems we should consider using the best models presented here for evaluating movement patterns, especially when aiming to reward players for correctly performing exercise repetitions in weight-shifting exercises. We showed that training ML/DL models using labeled training data is a feasible option for identifying correctly performed movement patterns, which can subsequently be used to reward players in an accurate manner during exergaming. This is an important improvement of many existing exergame systems that are based on comparisons to templates, or assessments using coarse rules and thresholds. Moreover, implementing a self-learning approach based on our work can allow a system to learn new movements without requiring a priori explicit identification of their templates. Trusting that the game system is actually rewarding the correct movements is a prerequisite for using exergames in serious settings like physical rehabilitation or independent exercise for older adults. If the game system is trusted, the threshold for using exergame systems might be lower for both users and clinicians, making it possible to benefit from higher motivation and adherence in the rehabilitation process. In future work, the implementation of the present classification models into game systems would be an interesting next step, possibly testing differences in rewards and/or feedback compared to rule-based or template-based systems. Exploring features is also a natural next step. The results of this study also warrant further investigation into how well these models perform in patient populations with more variable movement patterns, and in classification of error types. Furthermore, other movement patterns are also interesting to examine for classification accuracy, especially more complex movements that combine movements in various anatomical planes.

APPENDIX A
FEATURES

TABLE V
FEATURES CALCULATED FROM TSFRESH

Variable	Parameters/Units
Variance	
Standard deviation	
Mean	
Maximum	
Minimum	
Sum of values	
Count below mean	
Count above mean	
Sum reoccurring values	
Longest strike above mean	
Has duplicate values	True, False
Kurtosis	
Skewness	
Complexity invariance distance	True, False
Absolute sum of changes	
Change quantiles	Var,mean
Max Langevin Fixed Point	
Fourier Transform Coefficient	Abs,angle,real,imag
Fourier Transform Aggregated	Skew,centroid, kurtosis,variance
Mean absolute change	
Quantile	Q 0.1-0.9
Spektral Welch Density	Coeff 2,5,8
Large sd	R 0.01,0.05,0.25
Variance larger than sd	True,False
Binned entropy	Max bins 10
Number crossing m	-1,0,1
Range count	Max 1, min-1
Value count	0
Ratio beyond r sigma	0.5, 1.5, 5
Linear trend	P-value,intercept,slope
Aggregate linear trend	Max, min, mean
Quantile	
Has duplicate minimum	
Has duplicate maximum	
First location of minimum	
Last location maximum	
Last location minimum	
Has duplicate maximum	
Has duplicate minimum	
First location minimum	
Quantile	0.1-0.9
Autocorrelation	Lag 1-9
Agg autocorrelation	Mean,median,var
Partial autocorrelation	Lag 1-9
Absolute energy	
Continous wavelet transform	Width, peaks
Autoregressive AR(k)	2,3,4
Count above mean	
Augmenter dickey fuller	p-value, teststat
Energy ratio by chunks	
Friedrich Coefficients	
% of reoccurring values	
Value to time series length	Ratio
Number of peaks	
Mean second derivative central	
Index mass quantile	

APPENDIX B
JOINT SUBSET CLASSIFICATION RESULTS

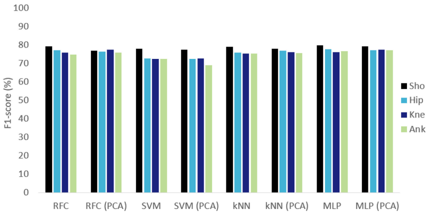


Fig. 8. F1-score achieved using different feature representations with CV10 on joint subsets RA = random forest, SVM = support vector machine, kNN = k-nearest neighbor, MLP = multilayer perceptron.

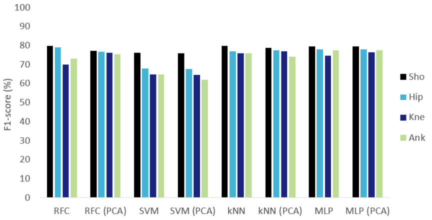


Fig. 9. F1-score achieved using different feature representations with LOGO on joint subsets RA = random forest, SVM = support vector machine, kNN = k-nearest neighbor, MLP = multilayer perceptron.

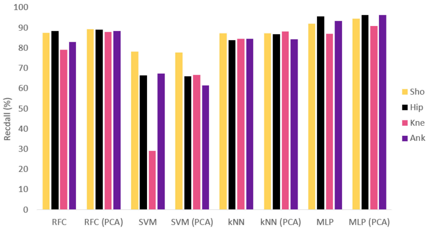


Fig. 10. Recall achieved using different feature representations with LOGO on joint subsets RA = random forest, SVM = support vector machine, kNN = k-nearest neighbor, MLP = multilayer perceptron.

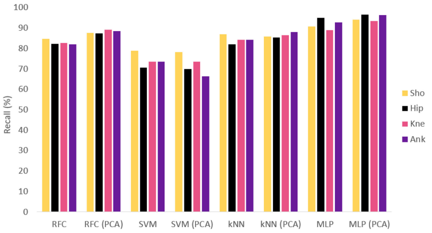


Fig. 11. Recall achieved using different feature representations with CV10 on joint subsets RA = random forest, SVM = support vector machine, kNN = k-nearest neighbor, MLP = multilayer perceptron.

ACKNOWLEDGMENT

The authors would like to thank A. Øvreness and M. A. Berge for assisting in data collection, C. Velvin for game development, and R. Stock for valuable discussions on weight shifting in physical rehabilitation.

REFERENCES

[1] C. Girard, J. Ecalle, and A. Magnan, "Serious games as new educational tools: How effective are they? A meta-analysis of recent studies," *J. Comput. Assist. Learn.*, vol. 29, no. 3, pp. 207–219, 2013.

[2] A. Ahmed and M. J. Sutton, "Gamification, serious games, simulations, and immersive learning environments in knowledge management initiatives," *World J. Sci., Technol. Sustain. Develop.*, vol. 14, no. 2/3, pp. 78–83, 2017.

[3] M. Graafland, J. M. Schraagen, and M. P. Schijven, "Systematic review of serious games for medical education and surgical skills training," *Brit. J. Surgery*, vol. 99, no. 10, pp. 1322–1330, 2012.

[4] T. M. Fleming *et al.*, "Serious games and gamification for mental health: Current status and promising directions," *Front. Psychiatry*, vol. 7, Jan. 2017, Art. no. 215.

[5] N. Skjæret, A. Nawaz, T. Morat, D. Schoene, J. Lægdheim, and B. Vereijken, "Exercise and rehabilitation delivered through exergames in older adults: An integrative review of technologies, safety and efficacy," *Int. J. Med. Inform.*, vol. 85, no. 1, pp. 1–16, 2016.

[6] S. Deterding, "Gamification: Designing for motivation," *Interactions*, vol. 19, no. 4, 2012, Art. no. 14. [Online]. Available: https://dl.acm.org/citation.cfm?doid=2212877.2212883

[7] N. Gal, D. Andrei, D. I. Nemeş, E. Ndşan, and V. Stoicu-Tivadar, "A Kinect based intelligent e-rehabilitation system in physical therapy," *Studies Health Technol. Inform.*, vol. 210, pp. 489–493, 2015.

[8] W. Zhao, A. M. Reinthal, D. D. Espy, and X. Luo, "Rule-based human motion tracking for rehabilitation exercises: Realtime assessment, feedback, and guidance," *IEEE Access*, vol. 5, pp. 21382–21394, 2017.

[9] M. Pasch, N. Bianchi-Berthouze, B. van Dijk, and A. Nijholt, "Movement-based sports video games: Investigating motivation and gaming experience,*" Entertainment Comput.*, vol. 1, no. 2, pp. 49–61, 2009.

[10] L. H. Skjaerven, K. Kristoffersen, and G. Gard, "An eye for movement quality: A phenomenological study of movement quality reflecting a group of physiotherapists' understanding of the phenomenon," *Physiotherapy Theory Pract.*, vol. 24, no. 1, pp. 13–27, 2008.

[11] A. Lacroix, T. Hortobágyi, R. Beurskens, and U. Granacher, "Effects of supervised vs. unsupervised training programs on balance and muscle strength in older adults: A systematic review and meta-analysis," *Sports Med.*, vol. 47, no. 11, pp. 2341–2361, 2017.

[12] J. D. Smeddinck, M. Herrlich, and R. Malaka, "Exergames for physiotherapy and rehabilitation: A medium-term situated study of motivational aspects and impact on functional reach," in *Proc. ACM CHI'15 Conf. Human Factors Comput. Syst.*, 2015, vol. 1, pp. 4143–4146. [Online]. Available: https://dx.doi.org/10.1145/2702123.2702598

[13] J. R. Beard and D. E. Bloom, "Towards a comprehensive public health response to population ageing," *Lancet*, vol. 385, no. 9968, pp. 658–661, 2015.

[14] F. Ofli, G. Kurillo, Š. Obdržálek, R. Bajcsy, H. B. Jimison, and M. Pavel, "Design and evaluation of an interactive exercise coaching system for older adults: Lessons learned," *IEEE J. Biomed. Health Inform.*, vol. 20, no. 1, pp. 201–212, Jan. 2016.

[15] A. W. Lam, D. Varona-Marin, Y. Li, M. Fergenbaum, and D. Kulić, "Automated rehabilitation system: Movement measurement and feedback for patients and physiotherapists in the rehabilitation clinic," *Human-Comput. Interact.*, vol. 31, no. 3/4, pp. 294–334, 2016.

[16] M. Pirovano, E. Surer, R. Mainetti, P. L. Lanzi, and N. Alberto Borghese, "Exergaming and rehabilitation: A methodology for the design of effective and safe therapeutic exergames," *Entertainment Comput.*, vol. 14, pp. 55–65, 2016.

[17] E. J. Lyons, "Cultivating engagement and enjoyment in exergames using feedback, challenge, and rewards," *Games Health J.*, vol. 4, no. 1, pp. 12–18, 2015.

[18] L. Lo Presti and M. La Cascia, "3D skeleton-based human action classification: A survey," *Pattern Recognit.*, vol. 53, pp. 130–147, May 2016.

[19] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Commun. Surv. Tut.*, vol. 15, no. 3, pp. 1192–1209, 2013.

[20] A. Depari, P. Ferrari, A. Flammini, S. Rinaldi, and E. Sisinni, "Lightweight machine learning-based approach for supervision of fitness workout," in *Proc. IEEE Sensors Appl. Symp., Conf.* 2019, vol. 5, pp. 1–6.

[21] E. K. Vonstad, X. Su, B. Vereijken, J. H. Nilsen, and K. Bach, "Classification of movement quality in a weight-shifting exercise," *in Proc. CEUR Workshop.*, 2018, vol. 2148, pp. 27–32.

[22] J. Wiemeyer and A. Kliem, "Serious games in prevention and rehabilitation—A new panacea for elderly people?," *Eur. Rev. Aging Physical Activity*, vol. 9, no. 1, pp. 41–50, 2012.

[23] E. Flores, G. Tobon, E. Cavallaro, F. I. Cavallaro, J. C. Perry, and T. Keller, "Improving patient motivation in game development for motor deficit rehabilitation," in *Proc. Int. Conf. Adv. Comput. Entertainment Technol.*, Jan. 2008, pp. 381–384. [Online]. Available: https://portal.acm.org/citation.cfm?doid=1501750.1501839

[24] M. van Diest, C. C. Lamoth, J. Stegenga, G. J. Verkerke, and K. Postema, "Exergaming for balance training of elderly: State of the art and future developments," *J. Nanoeng. Rehabil.*, vol. 10, no. 1, 2013, Art. no. 101.

[25] I. J. M. de Rooij, I. G. L. van de Port, and J.-W. G. Meijer, "Effect of virtual reality training on balance and gait ability in patients with stroke: Systematic review and meta-analysis," *Physical Therapy*, vol. 96, no. 12, pp. 1905–1918, 2016.

[26] E. F. Ogawa, T. You, and S. G. Leveille, "Potential benefits of exergaming for cognition and dual-task function in older adults: A systematic review," *J. Aging Physical Activity*, vol. 24, no. 2, pp. 332–336, 2016.

[27] Y. Gao and R. L. Mandryk, "The acute cognitive benefits of casual exergame play," in *Proc. Conf. Human Factors Comput. Syst.*, 2012, pp. 1863–1872.

[28] M. Adcock, F. Sonder, A. Schättin, F. Gennaro, and E. D. De Bruin, "A usability study of a multicomponent video game-based training for older adults," *Eur. Rev. Aging Physical Activity*, vol. 17, no. 1, pp. 1–15, 2020.

[29] E. D. Mekler, F. Brühlmann, A. N. Tuch, and K. Opwis, "Towards understanding the effects of individual gamification elements on intrinsic motivation and performance," *Comput. Human Behav.*, vol. 71, pp. 525–534, 2017. [Online]. Available: https://dx.doi.org/10.1016/j.chb.2015.08.048

[30] S. Göbel, S. Hardy, V. Wendel, F. Mehm, and R. Steinmetz, "Serious games for health - personalized exergames," in *Proc. ACM Multimedia Int. Conf.*, 2010, pp. 1663–1666. [Online]. Available: https://dl.acm.org/citation.cfm?doid=1873951.1874316

[31] K. Gerling, I. Livingston, L. Nacke, and R. Mandryk, "Full-body motion-based game interaction for older adults," in *Proc. ACM Annu. Conf. Human Factors Comput. Syst.*, 2012, pp. 1873–1882. [Online]. Available: https://dl.acm.org/citation.cfm?doid=2207676.2208324

[32] M. Antunes, R. Baptista, G. Demisse, D. Aouada, and B. Ottersten, "Visual and human-interpretable feedback for assisting physical activity," in *Proc. Comput. Vision Workshops*, 2016, pp. 115–129.

[33] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 556–571.

[34] X. Yu and S. Xiong, "A dynamic time warping based algorithm to evaluate kinect-enabled home-based physical rehabilitation exercises for older people," *Sensors (Switzerland)*, vol. 19, no. 13, 2019, Art. no. 2882.

[35] Y. Liao, A. Vakanski, and M. Xian, "A deep learning framework for assessing physical rehabilitation exercises," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 2, pp. 468–477, Feb. 2020.

[36] N. A. Borghese, M. Pirovano, P. L. Lanzi, S. Wüest, and E. D. de Bruin, "Computational intelligence and game design for effective at-home stroke rehabilitation," *Games Health J.*, vol. 2, no. 2, pp. 81–88, 2013.

[37] L. Tao *et al.*, "A comparative study of pose representation and dynamics modelling for online motion quality assessment," *Comput. Vision Image Understanding*, vol. 148, pp. 136–152, 2016.

[38] L. D. M. Carvalho and V. Furtado, "Using machine learning for evaluating the quality of exercises in a mobile exergame for tackling obesity in children," in *Proc. SAI Intell. Syst. Conf.*, 2016, vol. 16, pp. 373–390. [Online]. Available: https://link.springer.com/10.1007/978-3-319-56994-9

[39] *Plug-in Gait Reference Guide*, Vicon Motion Systems Ltd., Oxford, U.K., 2016.

[40] A. Da Gama, P. Fallavollita, V. Teichrieb, and N. Navab, "Motor rehabilitation using kinect: A systematic review," *Games Health J.*, vol. 4, no. 2, pp. 123–135, 2015.

[41] M. Christ, A. W. Kempa-Liehr, and M. Feindt, "Distributed and parallel time series feature extraction for industrial big data applications," in *Proc. ACML Workshop Learn. Big Data*, Nov. 2016, pp. 1–17. [Online]. Available: https://arxiv.org/abs/1610.07717

[42] P. E. Taylor, G. J. Almeida, J. K. Hodgins, and T. Kanade, “Multi-label classification for the analysis of human motion quality,” in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2012, pp. 2214–2218.

[43] S. Gaglio, G. L. Re, and M. Morana, “Human activity recognition process using 3-D posture data,” *IEEE Trans. Human-Mach. Syst.*, vol. 45, no. 5, pp. 586–597, Oct. 2015.

[44] M. Feurer and F. Hutter, “Hyperparameter optimization,” in *Automated Machine Learning*. Cham, Switzerland: Springer, 2019, pp. 3–33.

[45] C. J. Van Rijsbergen, *Information Retrieval*. 2nd ed. London, U.K.: Butterworth, 1979.

[46] E. Zdravevski *et al.*, “Improving activity recognition accuracy in ambient-assisted living systems by automated feature engineering,” *IEEE Access*, vol. 5, pp. 5262–5280, 2017. [Online]. Available: https://ieeexplore.ieee.org/document/7883880/

[47] C. J. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[48] D. P. Kingma and J. Lei Ba, “Adam: A method for stochastic optimization,” in *Proc. 3rd Int. Conf. Learn. Representations (ICLR 2015)*, 2015, *arXiv:1412.6980*.

[49] A. Atla, R. Tada, V. Sheng, and N. Singireddy, “Sensitivity of different machine learning algorithms to noise,” *J. Comput. Sci. Colleges*, vol. 26, no. 5, pp. 96–103, 2013.

[50] R. Caruana and A. Niculescu-Mizil, “An empirical comparison of supervised learning algorithms,” in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 161–168.

[51] A. Yurtman and B. Barshan, “Automated evaluation of physical therapy exercises using multi-template dynamic time warping on wearable sensor signals,” *Comput. Methods Programs Biomed.*, vol. 117, no. 2, pp. 189–207, 2014.

Chapter 10

Paper III

Estimation of Ground Reaction Force from Kinematic Data in Balance Exergaming

Authors: Elise Klæbo Vonstad, Kerstin Bach, Beatrix Vereijken, Xiaomeng Su, Jan Harald Nilsen.

Published in Journal of NeuroEngineering and Rehabilitation. Volume 19, issue 18, 2022. DOI: 10.1186/s12984-022-00998-5

RESEARCH

Estimation of Ground Reaction Force from Kinematic Data in Balance Exergaming

Elise Klæbo Vonstad[1*]
, Kerstin Bach[1]
, Beatrix Vereijken[2]
, Xiaomeng Su[1]
and Jan Harald Nilsen[1]

[*] Correspondence:
elise.k.vonstad[at]ntnu.no
[1] Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway
Full list of author information is available at the end of the article

Abstract

Background: Balance training exercise games (exergames) are a promising tool for reducing fall risk in elderly. Exergames can be used for in-home guided exercise, which greatly increases availability and facilitates independence. Providing biofeedback on weight-shifting during in-home balance exercise improves exercise efficiency, but suitable equipment for measuring weight-shifting is lacking. Exergames use kinematic data as input for game control: using this data to estimate weight-shifting would be a great advantage, which might be feasible using machine learning (ML) models. Therefore, the aim of this study was to investigate the performance of ML models in estimation of weight-shifting during exergaming using kinematic data.

Methods: Twelve healthy older adults (mean age 72(±4.2), 10 F) played a custom exergame that required repeated weight-shifts. Full-body 3D motion capture (3DMoCap) data and standard 2D digital video (2D-DV) was recorded. Weight shifting was directly measured by 3D ground reaction forces (GRF) from force plates, and estimated using a linear regression model, a long-short term memory (LSTM) model and a decision tree model (XGBoost). Performance was evaluated using coefficient of determination (R^2) and root mean square error (RMSE).

Results: Results from estimation of GRF components using 3DMoCap data shows a mean (±1SD) RMSE (% total body weight, BW) of the vertical GRF component (F_z) of 4.3 (2.5), 11.1 (.4.5), and 11.0 (4.7) for LSTM, XGBoost and LinReg, respectively. Using 2D-DV data, LSTM and XGBoost achieve mean RMSE (±1SD) in F_z estimation of 10.7 (9.0) %BW and 19.8 (6.4) %BW, respectively. R^2 was $>$.97 for the LSTM in the F_z component using 3DMoCap data, and $>$.77 using 2D-DV data. For XGBoost, F_z R^2 was $>$.86 using 3DMoCap data, and $>$.56 using 2D-DV data.

Conclusion: This study demonstrates that an LSTM model can estimate 3-dimensional GRF components using 2D kinematic data extracted from standard 2D digital video cameras. The F_z component is estimated more accurately than F_y and F_x components, especially when using 2D-DV data. Weight-shifting performance during exergaming can thus be extracted using kinematic data only, which can enable effective independent in-home balance exergaming.

Keywords: Weight Shifting; Balance Training; Exergaming; Ground Reaction Force; Deep Learning; Long Short-Term Memory Networks; XGBoost

Background

Being able to maintain or regain balance is a corner stone for sustained independence in daily life of older adults. Balance, or postural control, is a complex motor skill that depends on coordination and function of multiple bodily systems [1]. As we age, our postural control deteriorates gradually, increasing the risk of falls and decreasing community mobility and quality of life. These are major factors of increased risk of disability and mortality in elderly [2]. Targeted balance exercise improves postural control, and exercises typically included in exercise programs for balance training are for example leaning, reaching and weight shifting [3]. These types of exercises have been shown to reduce fall risk [4, 5] by improving dynamic stability during gait [6] as well as anticipatory and reactive balance ability[3]. Research has shown that technological tools that provide visual biofeedback and guidance can improve the potential effect of such exercises [7, 8]. By using exercise games (so-called exergames), biofeedback can be provided in a motivational and fun manner [9, 10]. In weight-shifting exercises, biofeedback is provided typically by using force-sensing equipment placed under the person's feet or inside the shoes. One of the most accurate types of force measurement equipment are piezoelectric force plates. These return three-dimensional ground reaction force (GRF) vectors, which are precise representations of the magnitude and directions of the force exerted on the plates by the person's feet.

Even though force plates are effective to provide biofeedback in balance exercising, they are rarely used outside laboratory settings as they are costly and resource demanding to use. More user friendly substitutes, such as the Wii Balance Board (Nintendo Co Ltd, Japan) have been developed and are used in exergames for balance training, but they are less accurate and register limited information only [11, 12]. More recent exergames for balance training started using kinematic data from depth-sensing cameras such as the Kinect (Microsoft Inc). However, using kinematic data as a proxy for kinetic information is problematic due to insufficient accuracy in the kinematic data provided [13]. Accurate and useful information about exercise performance is vital if independent exercise in older adults is to be effective. At the same time the equipment necessary to provide this information has to be easy to use and resource friendly, without sacrificing accuracy.

We know from previous research that GRF can be successfully estimated in other movements such as gait using machine learning (ML) methods. In [14] a Long-Short Term Memory (LSTM) model was employed, achieving estimates of GRF components within 12 % RSME, and in [15, 16] feed-forward artificial neural networks (ANN) gave a RMSE of GRF forces of <10 % in all three components. These studies all use features that are based on computation of a biomechanical model from a 3DMoCap system, which is not feasible to use in an in-home setting for elderly. It also requires physical measurements of the body of the person playing, as well as an additional computational layer for the calculation of the biomechanical model.

Others use data from inertial measurement units (IMU's) as input to neural networks, as seen in [17], or to an inverse dynamics model as seen in [18]. Although successful in estimation, with an error rate of <15% [18] and <10%[17], IMU-based approaches require procedures (e.g., full-body device placement) that are not feasible to implement in daily-life settings. These studies also employ a biomechanical model as input to their estimation procedures.

LSTM is a form of neural network where sequential data is processed recurrently and important features are ”remembered” for future predictions/estimations [19]. LSTMs are also relatively quick in estimation, allowing for real-time estimates which is a requirement when giving feedback during exergaming. Another approach, widely used because of its powerful method of representing the relationships in the data, is decision tree-based methods. Recently, a version of decision trees, called ”extremely boosted gradient trees” (XGBoost, [20]), has been shown to outperform other regression methods [21], including in estimation of forces in a biomechanical setting [22]. In addition, decision trees are inherently transparent in their decision making process, which is a highly valuable feature. This can provide information about which joints are important in estimating GRF, which might inform decisions on relevant motion tracking tools in this context.

Furthermore, it has recently been shown that standard digital 2D video can be used to extract 2D kinematic data of joint positions (e.g. [23, 24, 25]). This makes it possible to use devices such as smartphones, tablets or web-cameras to capture movements. These pose estimation systems could provide accessible, easy to use tools for motion capture in exergaming, and our recent research showed that such models are comparable in accuracy to the gold-standard motion capture systems [26]. However, a remaining challenge for using 2D data to estimate 3D GRF components is that in this case, the 2D-DV camera is placed directly behind the participant, meaning that the movements in the anterio-posterior (X) dimension is missing from the data set, affecting the estimation of the F_x-component of the GRF data.

In sum, the technology to capture movements accurately while playing weight-shifting exergames is available, but the tools to provide GRF information during weight-shifting exergaming, without needing force-sensing equipment, have not been tested and validated yet.

We propose utilizing positional data of joint centers in combination with machine learning models to estimate 3D GRF components during balance training. This would remove the need for any physical measurements of the person playing, and achieving this using a standard digital video camera only would make the system very easy to use and suitable for in-home guided exercise. Therefore, the aim of this paper is two-fold: 1) to investigate the performance of an LSTM model and an XGBoost model for estimation of ground reaction forces using kinematic data during balance exergaming; and 2) to compare performance between using 3D and 2D kinematic data.

Methods

Participants and protocol

Twelve healthy older adults were recruited from local exercise groups. Mean age was 72 ± 4.2 years, 10 were female. Exclusion criteria were physical or cognitive injuries/impairments that affected their balance and gait ability, and age <50 or age >80 years. Data was collected at the Movement Capture and Visualization Laboratory at the Norwegian University of Science and Technology in Trondheim, Norway in June 2019.

The Exergame

A custom exergame for balance training was used in this study, using Kinect (v2, Microsoft Inc) to track participants' movements for input to the game. The exergame was designed to elicit medio-lateral weight shifts from the user: An avatar representing the user was shown in a rail cart on a train rail, as seen in Figure 1. Along each side of the rail there were coins that the user should try to hit by tilting the cart sideways, which was achieved by shifting their body weight over to the foot that on the side of the coin (Figure 2). There were never more than two coins consecutively on one side. There were approximately 100 coins in total, with 50 % appearing on each side.

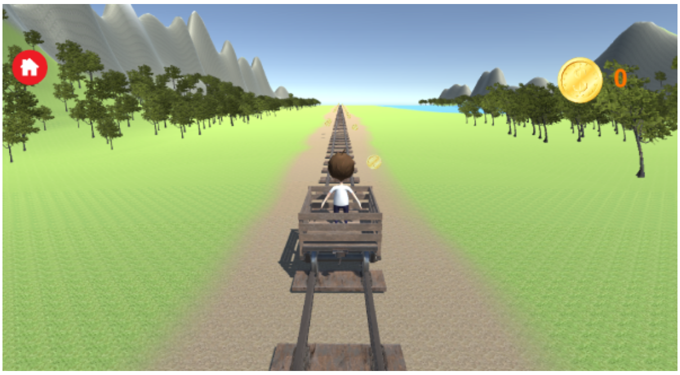


Figure 1 Game interface

Equipment

A four-camera (MX400, 90Hz, Qualisys Inc, Sweden) setup was used for capturing 3D motion data (3DMoCap) from participants. The Plug-in-Gait Full Body (PiG-FB, [27]) marker setup, excluding head and hands, was used. Two digital cameras (GoPro Hero Black 3+, 25 Hz, GoPro Inc) placed 200cm behind and to the side of the player were used to capture player movements simultaneously with the 3DMoCap system. To capture force data, two force plates (60x5x40 cm, 1000Hz, Kistler AB) were used, one under each foot of the player. The experimental setup can be seen in Figure 3.

Preprocessing

To extract joint center positions from 2D-DV data, the DeepLabCut(DLC, [24]) framework was used. The 3DMoCap data was gap-filled and the joint center positions were extracted using the standardized PiG-FB biomechanical model implemented in Nexus (v. 2.9, Vicon Motion Systems Ltd). The joint center positions extracted from both data sources were ankles, knees, hips, shoulders, elbows and wrists. From the 3DMoCap system the anterio-posterior (X), medio-lateral (Y) and vertical (Z) positions relative to the Qualisys global coordinate system origin were extracted, and in the 2D-DV data the vertical (Y) and medio-lateral (X) positions

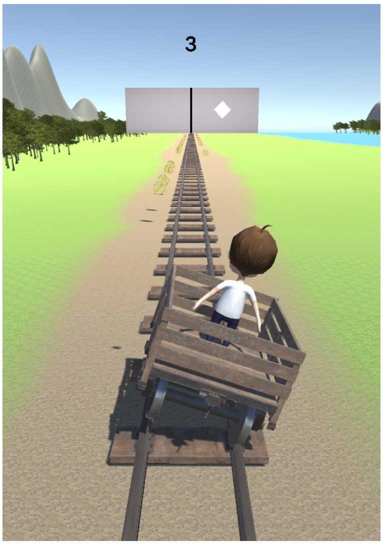


Figure 2 Cart tilting sideways to hit coin.

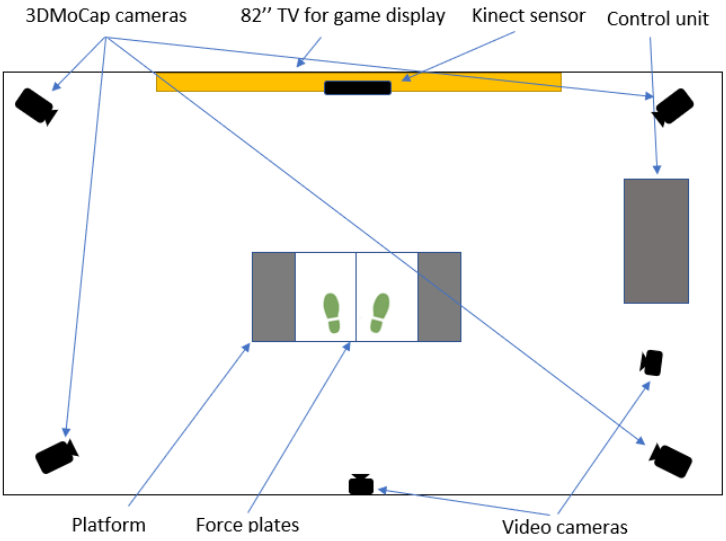


Figure 3 Experimental setup.

relative to the 2D-DV camera origin were extracted. This resulted in 36 input features from the 3DMoCap system, and 24 features from the 2D-DV system. The data was then normalized to the [0,1] range. Data was synchronized by resampling joint center data from digital video using the 3DMoCap data frequency as reference. Force components F_x (anterio-posterior), F_y (medio-lateral) and F_z (vertical) were extracted from the force plate data. GRF components were scaled to body weight (BW) for each time frame. The video data of ankles was occluded in participants 4, 8, 9, and 10, resulting in missing ankle data for these participants. 3DMoCap data from participants 1 and 2 was corrupted, and not used in further analyses.

Machine Learning Models

Python v. 3.7.10 was used for all analyses and evaluation. Sci-Kit Learn [28] was used for multivariate linear regression (LinReg), GridSearchCV and feature importance, and for evaluation of model performances; the Keras framework [29] was used to build the LSTM model; and XGBoost was implemented using the XGBoost package for Python (https://github.com/dmlc/xgboost).

Multivariate linear regression (LinReg) is a method for modeling a linear relationship between independent feature variables $x_1, x_2, x_3 ... x_n$ and target variables $\beta_1, \beta_2, \beta_3 ... \beta_n$ [30]. LinReg uses a least squares optimization method of finding the optimal line that represents the relationship between the features and the target variable, and has been used in countless estimation and prediction situations. In a multivariate LinReg, there are several (n) target variables; the model then fits a hyperplane in n dimensions to represent the relationship between all the feature variables and the target variables [30]. This model was used as a baseline model for reference purposes, as it is considered a go-to model that performs very well in a wide variety of applications in data analysis [30].

In contrast to linear models, decision trees are able to represent non-linear relationships between features and target variables. Decision trees find attribute values that separate the data well at descending nodes in the tree, with a new split of a different attribute in the next node, ending in a leaf node that is essentially an output node. In regression trees each leaf node has a continuous score, and the final estimation from each tree is the mean of these scores [30]. Decision trees are very powerful, and XGBoost is an improved version of decision tree models that combines random forest technique of feature bagging, and a gradient decent method to reduce boosting error - hence the name ”gradient boosting”. This improves performance of the model, and XGBoost has been shown to perform very well on a wide range of non-linear estimation tasks [20]. We therefore hypothesize that it can perform well in the current task.

Long short-term memory model (LSTM) is a version of a recurrent neural network. The network consists of units that data is passed through recurrently. Using different types of gates (input, forget and output gates [31]), it is able identify information that is important to ”remember” from the previous input data [32], thereby enabling it to learn long-range dependencies in the data [19]. In the current problem setting, remembering past relationships between the kinematic data and the GRF component data might be fruitful as the limb orientation and posture of the person at each time step directly affects the GRF component magnitude. LSTM has drawn much attention in several fields where time series forecasting and recognition is important [33]. Stacked LSTM is a version of LSTM models that utilizes several layers of LSTM nodes, which has been shown to improve performance over single layer LSTMs [34]. A schematic of the stacked LSTM model we implemented in this study can be seen in Figure 4. Here, there is one dense input layer, three hidden layers of 512 nodes each, a dropout layer (0.2), and a dense output layer of 6 nodes with sigmoid activation: one for each dimension in the force data for each force plate.

Parameters and Optimization

Hyperparameters for the XGBoost model were tuned using GridSearchCV with five cross-validation iterations, and the most optimal hyperparameter settings were found. The hyperparameter grid searched can be found in Table 1. The hyperparameter values in bold font were the ones found to yield the highest performance, and were used in training the final XGBoost model.

Optimization of the LSTM network was conducted using Adam optimizer [35] with an initial learning rate of 0.0001, decay steps 10000, and decay rate 0.96. The model was trained for 200 epochs, with a minimum rate of improvement of loss (mean squared error, MSE) of 0.0003 for three consecutive epochs.

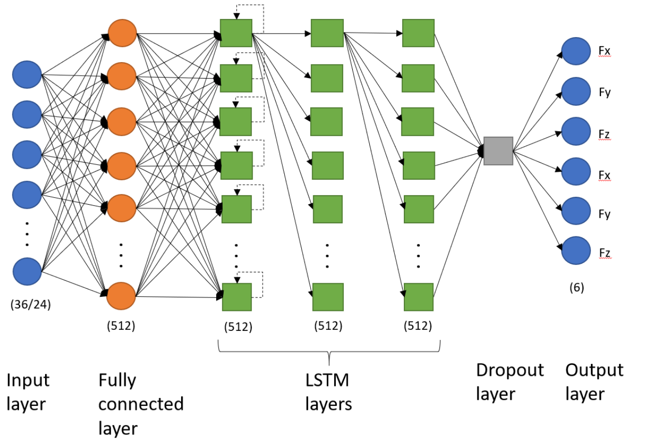


Figure 4 Schematic of the stacked long-short term memory (LSTM) model we implemented. For clarity, not all connections are shown; all layers re fully connected, and all LSTM units have the recurrent connection depicted in the first LSTM layer. The input layer consists of 36 nodes for 3DMoCap data and 24 nodes for 2D-DV data before feature selection.

A leave-one-group-out cross validation was performed on all models, where one group was the data from one participant, which served as the test set in each iteration. This was performed on the joint data from 3DMoCap and 2D-DV systems. For evaluation, mean of left and right foot (1SD) root mean square error (RMSE), and mean (1SD) coefficient of determination (R^2) for the different cross-validation splits was computed.

Results

The results showing feature selection and subsequent estimation performance of LSTM, XGBoost and LinReg using 3DMoCap and 2D-DV data, are presented as RMSE in Table 2 and R^2 in Figure 7. Figure 8 shows illustrative example graphs of estimation performance of the three models using 3D and 2D data, over a randomly selected sequence (1000 frames) from one person during one trial of play.

Furthermore, the contribution of each joint center to estimation performance was computed using a permutation procedure. Here, the data in each feature is shuffled in a random manner, which breaks the real-world relationship between the feature and the target. The resulting difference in estimation performance between using the shuffled and un-shuffled feature is indicative of how much the model depends on this feature [36]. This is then repeated for all features, and inform about which features, i.e. joint centers, are most important to the estimation performance. Results from the feature importance analysis, using 3DMoCap data, showed that eight joint centers contributed with 82.9% of the information needed to estimate GRF components. These joint centers were right and left wrist, right elbow, left knee, and torso joint centers (left and right shoulders, and left and right hip joints). The models were subsequently retrained using these joints.

Using 2D-DV data, there were also eight joint centers that had a total contribution of 78%: Left wrist, shoulder, hip, knee and ankle, and right shoulder, knee, and ankle. The relative contributions of all joint centers can be seen in Figures 5 and 6.

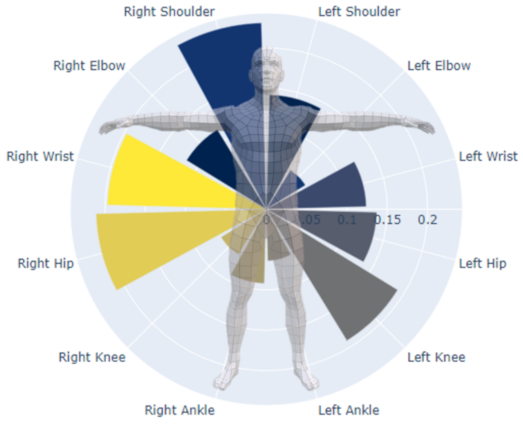


Figure 5 Overview of the joint centers' total impact (fraction of R^2) on estimation performance when using 3DMoCap data.

Estimation error

Prediction performance is presented in Table 2, with the mean (±1SD) RMSE (% BW) for the three models using 3DMoCap and 2D-DV data for the three force components. The LSTM model outperforms both XGBoost and LinReg when using

Table 1 Hyperparameter space searched in GridSearchCV for the XGBoost model after feature selection. Values in **bold** were used in further analyses.

Hyperparameter	Values searched
Learning rate	.001, **.005**, .01, .05, .10, .15
Max depth	5, 7, 9, **12**, 15
No. Estimators	50, 100, **200**, 500, 700
Min. child weight	1, 3, **5**, 7
Gamma	0.0, **0.1**, 0.2

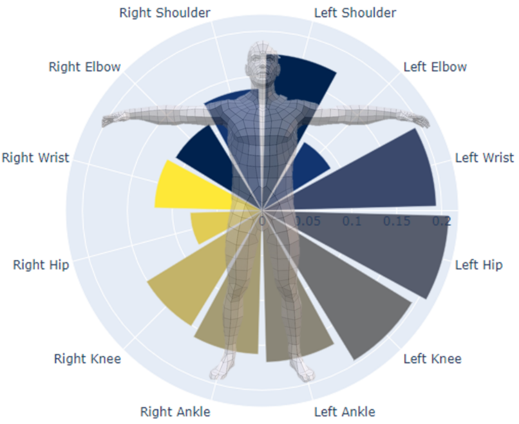


Figure 6 Overview of the joint centers' total impact (fraction of R^2) on estimation performance when using 2D-DV data.

both 3DMoCap and 2D-DV data. The XGBoost model achieves at the same level as LinReg using both 3DMoCap and 2D-DV data. Lowest mean RMSE (4.3% BW) was achieved by the LSTM model on the F_z component using 3DMoCap data; highest (23.5% BW) was the LinReg model in the F_y component using 2D-DV data. RMSE was generally higher using 2D-DV data than when using 3DMoCap data.

Table 2 Mean (±1SD) RMSE (% BW) achieved by the three models from estimation of all three components of GRF.

	3DMoCap			2D-DV		
	LSTM	XGBoost	LinReg	LSTM	XGBoost	LinReg
F_x	10.3 (6.2)	17.3 (3.6)	17.6 (3.8)	12.7 (7.6)	18.6 (3.8)	18.2 (3.8)
F_y	7.1 (4.6)	13.4 (3.9)	19.0 (2.7)	10.4 (6.0)	20.2 (4.0)	23.5 (8.5)
F_z	4.3 (2.5)	11.1 (4.5)	11.0 (4.7)	10.7 (9.0)	19.8 (6.4)	18.6 (6.4)

Model fit

As shown in Figure 7, the LSTM R^2 is consistently higher than in the XGBoost and LinReg model using both MoCap and 2D-DV data. Using the MoCap data, the mean (±1SD) LSTM R^2 was .589 (.34), .796 (.31), and .971 (.05) in the F_x, F_y, and F_z components, respectively, and XGBoost R^2 was -.246 (.27), .114 (.36), and .863 (.16), respectively. The LinReg model achieved a mean R^2 of -.168 (.28), -.054 (.21), and .856 (.17), respectively. Using 2D-DV data, all models achieved slightly lower R^2. LSTM achieved mean (± 1SD) R^2 of .379 (.55) in F_x, .579 (.58) in F_y and .770 (.45) in F_z. XGBoost mean (± 1SD) R^2 in F_x was -.313 (.26), -.234 (.53) in F_y, and .564 .(.31) in F_z. Here, the LinReg results were mean (± 1SD) R^2 of -.266 (.39), -.950 (2.23), and .617 (.28) for the F_x, F_y, and F_z components, respectively.

Estimation plots

In Figure 8, example plots from the left foot are presented that show the estimated component values by the XGBoost, LSTM, and LinReg models over a random set of 1000 frames, along with the ground truth component values. The LSTM model estimates all three components very well, both using MoCap and 2D-DV data. The F_x component seems to be the least accurate, although the LSTM model estimates the major changes in BW here as well. The XGBoost model also estimates F_z very well, but this is not seen to the same degree in F_y and F_x. In F_x and F_y the XGBoost model is able to follow the major trends in the data, although from these plots it seems like rapid changes in force are not estimated well. The LinReg model is able to estimate F_z fairly well, but not with the level of detail seen in the LSTM or XGBoost model. F_x and F_y components, however, are not estimated as well by the LinReg model.

Test/train error

The XGBoost RMSE from using test and train data is presented in Figure 9 and 10. The test error is consistently about 3× higher than the train error, using both 3DMoCap and 2D-DV data. For XGBoost, the mean (±) train/test RMSE was 8.8 (0.5)/19.8 (3.7) %BW, respectively, using 2D-DV data and 5.2 (0.1)/15.9 (3.0) % BW, respectively, using 3DMoCap data. For LSTM, the mean (±) train/test RMSE using 2D-DV data was 9.8 (0.6)/ 11.5 (7.6) %BW, respectively, and 11.85 (0.2)/ 13.6 (2.9) % BW, respectively, using 3DMoCap data.

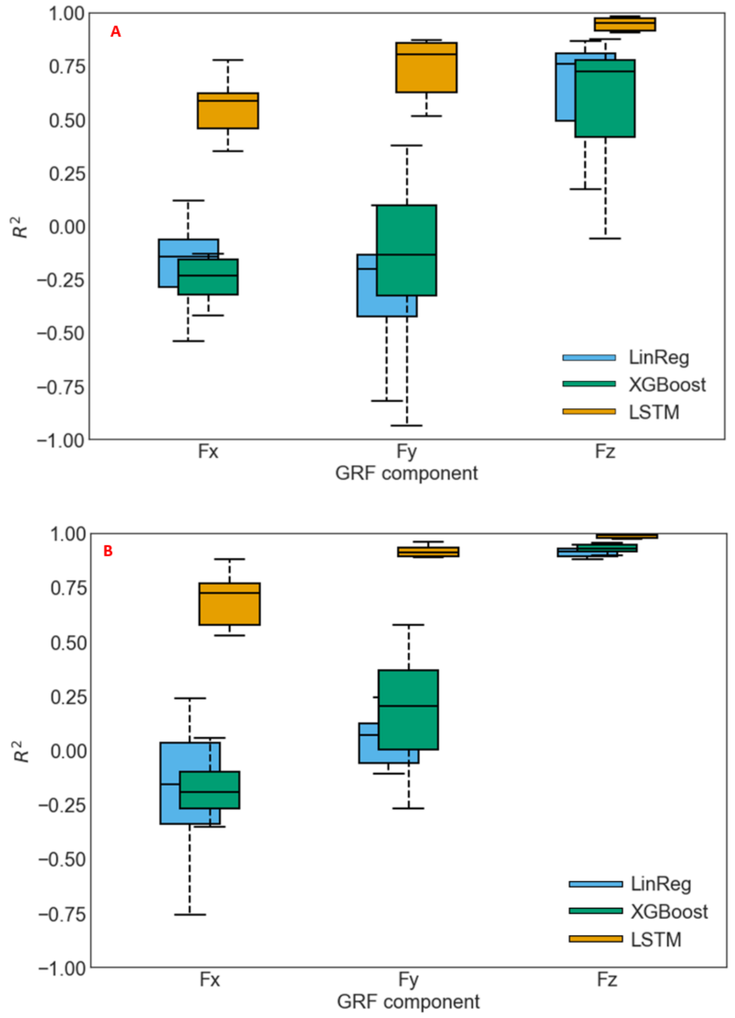


Figure 7 Box plots showing median R^2 from LSTM, XGBoost, and LinReg models in all three GRF components. Plot (a) shows results using 2D-DV data, and plot (b) MoCap data.

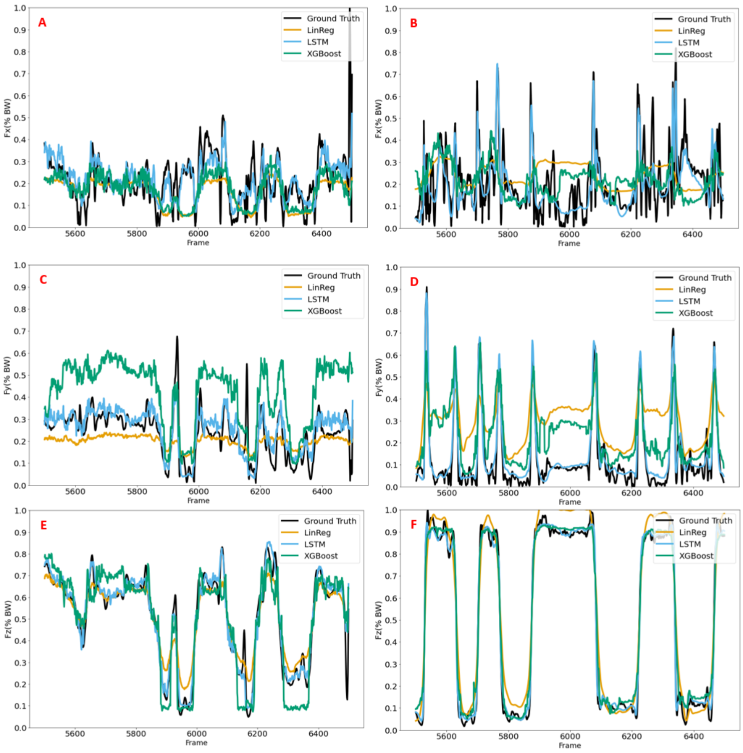


Figure 8 Example of estimation performance on the left side from XGBoost (green), LSTM (blue), and LinReg (orange) models in each GRF component over 1000 frames, along with the ground truth GRF (black). Plots (a),(c), and (e) show results from one 2D-DV dataset, and plots (b), (d) and (f) show one MoCap dataset.

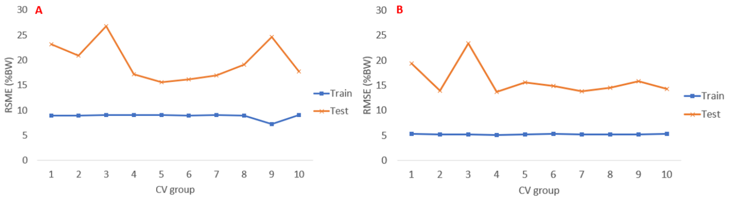


Figure 9 XGBoost model test/train RMSE (%BW) from each cross-validation iteration using 2D-DV (A) and 3DMoCap data (B).

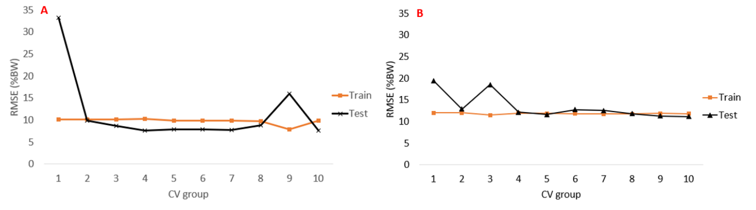


Figure 10 LSTM model test/train RMSE (%BW) from each cross-validation iteration using 2D-DV (A) and 3DMoCap data (B).

Discussion

This study investigated two facets of estimation of GRF components in balance training. First, we assessed the overall estimation performance of an LSTM and an XGBoost model on GRF components, comparing it to a baseline LinReg model's performance. Second, the performance of the LSTM and XGBoost models in estimating 3D GRF data using 2D joint data was examined. Overall, the LSTM model performance was very good, considering that joint position data was the only input data used for estimation. The LSTM RMSE was <11% BW for all GRF components when using 3DMoCap data, and R^2 was moderate to high (>.58 and >.79) for F_x and F_y, and excellent (>.97) for F_z. This shows that the LSTM model was able to accurately estimate the F_z component, while achieving only slightly less accurate results in the F_x and F_y components. The boxplots in Figure 7 also show that the F_z estimation was very stable around the median. This was the case in all three models.

The most promising facet of our results is that our method does not require information about the person playing or any calculations using the input data to represent the person - i.e., no biomechanical model is needed. This makes our method less computationally expensive, and easier to implement in an in-home setting. Still, our findings on estimation of GRF from kinematic data are in line with related literature in gait analysis, such as Mundt et al (2020) [14], S. Ooh et al (2018)[15], and Choi et al (2013)[16]. The movement pattern is different, so a direct comparison of results is not feasible. These studies used 3DMoCap data to calculate biomechanical features such as joint angles [15, 37], body segment velocities [16], and foot contact events [18], which are not obtainable using only joint position data. This demonstrates the strength in our results: our method use the joint center positions directly, skipping both practical and computational steps that complicate the process. This makes our method more accessible and easy to use, while being as accurate as more complicated methods.

Regarding performance using 2D-DV data, our findings support using this modality for estimation of F_z during balance exergaming. This is a step in the right direction regarding in-home use of exergaming, as a standard digital camera that most people already possess can provide accurate information about weight shifting performance during exergaming. However, our findings also show that when the context requires 3D GRF data, the use of 3D kinematic data is preferred to ensure estimation accuracy in all three GRF components. This is also true when the context requires model performance that has a <10% BW error requirement in other components than F_z.

LinReg also performs surprisingly well in F_z, with comparable RMSE and R^2 to LSTM and XGBoost, although both the LSTM and XGBoost models are better at estimating the small changes in force that occurs between lateral weight-shifts (i.e., when the person is standing with the majority of their BW on one foot).

The F_z component is arguably the most informative of the three directions in balance training, as it represents the vertical force - i.e., the weight that is being pushed straight downwards onto the surface. In practice, this informs about how much body weight the person places on each leg, which is an indication of how well the person is performing a weight shift during exercise. However, F_x and F_y information may also be relevant to measure accurately as the force exerted in these directions contribute to postural control. For example, force magnitude, directional accuracy, and variability in F_y and F_x in relation to a (externally or internally induced) disturbance in posture can be informative about balance ability [38, 39]. In medio-lateral weight-shifting the F_x component might not be as critical to measure as the F_z component measures the same movement in this context. In contrast, control over anterior-posterior movement (and thus F_y) is important to maintaining a steady and stable sideways movement pattern, to prevent large anterior-posterior movements during weight-shifting exercises and potentially create destabilizing conditions. This means that even though F_z provides the main information about sideways weight-shifting performance, F_y can inform about the variability and stability in a weight-shifting movement.

The feature importance information from the XGBoost model showed different joints to be important based on the type of data used. When using 3D data, more joints from the right side contributed to estimation performance, while more joint on the left side were important when using 2D data. From these results we were not able to elucidate any systematic or clear pattern in joint importance, which might be caused by the limited set of movements performed in this study. This might be an interesting avenue to explore further using a data set richer in terms of movements.

The high R^2 achieved could be a sign of overfitting by the LSTM model [40]. However, the 10-fold CV process showed a stable fit using test data, which can be seen in the low spread of the LSTM model in Figure 7 as well. Results from test/train errors also support this, as the difference between test/train errors is low, as seen in Figure 10. Even more reassuring is the fact that the CV process was not a holdout of random pieces of data, but a holdout of all the data from each person. Thus, estimation of GRF was performed on previously unseen data from a person with an unknown movement pattern.

The XGBoost model, however, does indeed seem to suffer from overfitting, which presents itself as higher RMSE when estimating based on unseen data compared to training data [41](Figure 9). This is likely caused by either too much noise in the data (especially in the 2D-DV data), where the limited tree depth (max depth = 12) does not allow for the tree to fully model the real relationship in the data, or that the current data set is too sparse. Even though XGBoost inherently possesses features that are known to prevent overfitting, our findings indicate that this was not successful here.

Limitations There are some limitations to be aware of in the current study. The movement pattern performed by participants was limited to to sideways leaning, and there were a low number of participants. The data was collected in a laboratory setting, and the models used require training data to be usable in a real-world setting.

Conclusion

In conclusion, the LSTM model performed very well, especially in F_z. 3DMoCap data produced the best results, but F_z estimation using 2D-DV data was also very good. These findings show that it is feasible to develop exergames that provides weight-shifting biofeedback by only using 2D joint position data from a standard digital video camera. With the support of a standard camera, an exergame in balance training can incorporate the LSTM model to provide real-time biofeedback on weight-shifting performance. This warrants further investigation into how such systems can be integrated into exergames for in-home or in balance exercise, as it opens up broad opportunities for providing accurate feedback in a simple, yet accurate manner. The LSTM model and 2D-DV input data combination has the potential to facilitate more effective and motivating in-home balance training by incorporating accurate feedback on weight-shifting performance in exergames.

Appendix

Acknowledgements

Authors thank the participants for providing their valuable time and effort, and Martin A. Berge for his assistance in data collection.

Abbreviations

Abbreviation	Meaning
2D	Two-dimensional
2D-DV	Digital Video
3DMoCap	Three-dimensional Motion Capture
AI	Artificial Intelligence
BW	Body Weight
CV	Cross-Validation
DL	Deep Learning
DLC	DeepLabCut
GRF	Ground Reaction Force
IMU	Inertial Measurement Unit
LinReg	Linear Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
PiG-FB	Plug-in-Gait FullBody
R^2	Coefficient of Determination
RMSE	Root Mean Square Error
SD	Standard Deviation
XGBoost	Extremely Boosted Gradient Trees

Availability of data and materials

The data set used and/or analysed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Participants were given written and oral information about the project, and all gave their written consent upon arrival at the laboratory. The project was approved by the Norwegian Regional Ethics Committee and the Norwegian Centre for Research Data.

Competing interests

The authors declare that they have no competing interests.

Funding

This study received to external funding.

Consent for publication

Not applicable.

Authors' contributions

E.K.V., J.H.N., and X.S. conceived the work, and all authors contributed in the design. E.K.V. acquired the data, and was joined by X.S., B.V., J.H.N., and K.B. in analysis and interpretation of data. E.K.V. drafted the manuscript, and all authors contributed in substantial revisions.

Author details

[1] Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway.
[2] Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, Trondheim, Norway.

References

1. Horak, F.B.: Postural orientation and equilibrium: What do we need to know about neural control of balance to prevent falls? Age and Ageing **35-S2**, 7–11 (2006). doi:10.1093/ageing/afl077
2. Rubenstein, L.Z.: Falls in older people: Epidemiology, risk factors and strategies for prevention. Age and Ageing **35**(SUPPL.2), 37–41 (2006). doi:10.1093/ageing/afl084
3. Sherrington, C., Whitney, J.C., Lord, S.R., Herbert, R.D., Cumming, R.G., Close, J.C.T.: Effective exercise for the prevention of falls: A systematic review and meta-analysis. Journal of the American Geriatrics Society **56**(12), 2234–2243 (2008). doi:10.1111/j.1532-5415.2008.02014.x
4. Gillespie, L.D., Robertson, M.C., Gillespie, W.J., Sherrington, C., Gates, S., Clemson, L.M., Lamb, S.E.: Interventions for preventing falls in older people living in the community. Cochrane Database Syst Rev **9**(9), 007146 (2012). doi:10.1002/14651858.CD007146.pub3
5. Shubert, T.E.: Evidence-based exercise prescription for balance and falls prevention: A current review of the literature. Journal of Geriatric Physical Therapy **34**(3), 100–108 (2011). doi:10.1519/JPT.0b013e31822938ac
6. Rogers, H.L., Cromwell, R.L., Grady, J.L.: Adaptive changes in gait of older and younger adults as responses to challenges to dynamic balance. Journal of Aging and Physical Activity **16**(1), 85–96 (2008). doi:10.1123/japa.16.1.85
7. Zijlstra, A., Mancini, M., Chiari, L., Zijlstra, W.: Biofeedback for training balance and mobility tasks in older populations: A systematic review. Journal of NeuroEngineering and Rehabilitation **7**(1), 1–15 (2010). doi:10.1186/1743-0003-7-58
8. Laver, K.K.E., Lange, B., George, S., Deutsch, J.J.E., Saposnik, G., Crotty, M.: Virtual reality for stroke rehabilitation. Cochrane Database of Systematic Reviews **11**(11), 008349 (2017). doi:10.1002/14651858.CD008349.pub4.www.cochranelibrary.com
9. Smeddinck, J.D., Herrlich, M., Malaka, R.: Exergames for Physiotherapy and Rehabilitation: A Medium-term Situated Study of Motivational Aspects and Impact on Functional Reach. Proceedings of the ACM CHI'15 Conference on Human Factors in Computing Systems **1**, 4143–4146 (2015). doi:10.1145/2702123.2702598
10. Sveistrup, H.: Motor rehabilitation using virtual reality. BioMed Central (2004). doi:10.1186/1743-0003-1-10
11. Leach, J.M., Mancini, M., Peterka, R.J., Hayes, T.L., Horak, F.B.: Validating and calibrating the Nintendo Wii balance board to derive reliable center of pressure measures **14**(10), 18244–18267. doi:10.3390/s141018244
12. Bartlett, H.L., Ting, L.H., Bingham, J.T.: Accuracy of force and center of pressure measures of the Wii Balance Board. Gait and Posture **39**(1), 224–228 (2014). doi:10.1016/j.gaitpost.2013.07.010
13. Obdrzalek, S., Kurillo, G., Ofli, F., Bajcsy, R., Seto, E., Jimison, H., Pavel, M.: Accuracy and robustness of Kinect pose estimation in the context of coaching of elderly population. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, pp. 1188–1193 (2012). doi:10.1109/EMBC.2012.6346149
14. Mundt, M., Koeppe, A., Bamer, F., David, S., Markert, B.: Artificial Neural Networks in Motion Analysis—Applications of Unsupervised and Heuristic Feature Selection Techniques. Sensors **20**(16), 4581 (2020). doi:10.3390/s20164581
15. Oh, S.E., Choi, A., Mun, J.H.: Prediction of ground reaction forces during gait based on kinematics and a neural network model. Journal of Biomechanics **46**(14), 2372–2380 (2013). doi:10.1016/j.jbiomech.2013.07.036
16. Choi, A., Lee, J.M., Mun, J.H.: Ground reaction forces predicted by using artificial neural network during asymmetric movements. International Journal of Precision Engineering and Manufacturing **14**(3), 475–483 (2013)
17. Jacobs, D.A., Ferris, D.P.: Estimation of ground reaction forces and ankle moment with multiple, low-cost sensors. Journal of NeuroEngineering and Rehabilitation **12**(1), 90 (2015). doi:10.1186/s12984-015-0081-x
18. Karatsidis, A., Bellusci, G., Schepers, H., de Zee, M., Andersen, M., Veltink, P.: Estimation of Ground Reaction Forces and Moments During Gait Using Only Inertial Motion Capture. Sensors **17**(12), 75 (2016). doi:10.3390/s17010075
19. Hochreiter, S., Hochreiter, S., Schmidhuber, J.: Long Short-term Memory. Technical report (1995). `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.51.3117`

20. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016). doi:10.1145/2939672.2939785. https://dl.acm.org/doi/10.1145/2939672.2939785
21. Dietterich, T.G.: Ensemble Methods in Machine Learning. Multiple Classifier Systems **1857**, 1–15 (2000). doi:10.1007/3-540-45014-9
22. Aljaaf, A.J., Hussain, A.J., Fergus, P., Przybyla, A., Barton, G.J.: Evaluation of machine learning methods to predict knee loading from the movement of body segments. Proceedings of the International Joint Conference on Neural Networks **2016-Octob**, 5168–5173 (2016). doi:10.1109/IJCNN.2016.7727882
23. Chen, Y., Tian, Y., He, M.: Monocular human pose estimation: A survey of deep learning-based methods. Computer Vision and Image Understanding **192**(December 2019), 102897 (2020). doi:10.1016/j.cviu.2019.102897
24. Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M.: DeepLabCut: Markerless Pose Estimation of User-Defined Body Parts with Deep Learning. Nature Neuroscience **21**(September) (2018). doi:10.1038/s41593-018-0209-y
25. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S.-E., Sheikh, Y.A.: OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–1 (2019). doi:10.1109/tpami.2019.2929257
26. Vonstad, E.K., Su, X., Vereijken, B., Bach, K., Nilsen, J.H.: Comparison of a deep learning-based pose estimation system to marker-based and kinect systems in exergaming for balance training. Sensors (Switzerland) **20**(23), 1–16 (2020). doi:10.3390/s20236940
27. Vicon Motion Systems Limited: Plug-in Gait Reference Guide. Technical report (2016). `https://docs.vicon.com/display/Nexus25/PDF+downloads+for+Vicon+Nexus?preview=/50888706/50889377/Plug-inGaitReferenceGuide.pdf`
28. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2012). doi:10.1007/s13398-014-0173-7.2
29. Chollet, F.: Keras. GitHub (2015). `https://github.com/fchollet/keras`
30. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning. Springer Texts in Statistics, vol. 103. Springer, New York, NY (2013). http://link.springer.com/10.1007/978-1-4614-7138-7
31. Greff, K., Srivastava, R.K., Koutnik, J., Steunebrink, B.R., Schmidhuber, J.: LSTM: A Search Space Odyssey. IEEE Transactions on Neural Networks and Learning Systems **28**(10), 2222–2232 (2017). doi:10.1109/TNNLS.2016.2582924
32. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning, p. 800. MIT Press, ??? (2016)
33. Lecun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015). doi:10.1038/nature14539
34. Sagheer, A., Kotb, M.: Unsupervised Pre-training of a Deep LSTM-based Stacked Autoencoder for Multivariate Time Series Forecasting Problems **9**(1). doi:10.1038/s41598-019-55320-6
35. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 1–15 (2015)
36. Breiman, L.: Random Forests, 1–33 (2001). doi:10.1017/CBO9781107415324.004
37. Mundt, M., Koeppe, A., David, S., Bamer, F., Potthast, W., Markert, B.: Prediction of ground reaction force and joint moments based on optical motion capture data during gait. Medical Engineering & Physics (2020). doi:10.1016/j.medengphy.2020.10.001
38. Horak, F.B., Hemy, S.M.: Postural Perturbations: New Insights for Treatment of Balance Disorders. Physical Therapy **77**, 517–533 (1997)
39. Önell, A.: The vertical ground reaction force for analysis of balance? Gait and Posture **12**(1), 7–13 (2000). doi:10.1016/S0966-6362(00)00053-9
40. Cheng, C.L., Shalabh, Garg, G.: Coefficient of determination for multiple measurement error models **126**, 137–152. doi:10.1016/j.jmva.2014.01.006
41. Mitchell, T.M.: Machine Learning, p. 432. McGraw Hill, ??? (1997). https://www.cs.cmu.edu/~tom/mlbook.html

ISBN 978-82-326-5908-1 (printed ver.)
ISBN 978-82-326-6623-2 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU
Norwegian University of
Science and Technology