Wearables in Sports: From Experimental Validation to Deep Learning Applications

Wearables im Sport: Von experimenteller Validierung zu Deep Learning-Anwendungen

Der Technischen Fakultät
der Friedrich-Alexander-Universität
Erlangen-Nürnberg

zur
Erlangung des Doktorgrades Dr.-Ing.

vorgelegt von
Johannes Link, M.Sc.
aus Würzburg
Als Dissertation genehmigt
von der Technischen Fakultät
der Friedrich-Alexander-Universität Erlangen-Nürnberg

Tag der mündlichen
Prüfung: 15.11.2023

Gutachter: Prof. Dr. Björn M. Eskofier
Prof. Dr. Jesse Davis
Acknowledgements

Upon completing my master’s in physics, I had not planned on pursuing a Ph.D. However, I stumbled upon an open position in machine learning and data analytics in sports, which greatly intrigued me. I would like to express my gratitude to Björn Eskofier, my supervisor, for giving me the opportunity to work on this fascinating topic. Thank you for your inspirational guidance and support.

Furthermore, I would like to extend my thanks to my industrial cooperation partner for this opportunity and the valuable insights into sports analytics. Special thanks go to Thomas Kautz, Markus Streicher, and Alex Hiemann, who were always there to offer their support.

I am also grateful to all my colleagues at the MaD Lab who made my experience there enjoyable. I appreciate the fun skiing trips, great team events, and fruitful discussions about research problems and solutions. I would like to thank my co-authors at the lab, Leo Schwinn, Falk Pulsmeyer, Timur Perst, and Maike Stoeve, for their collaboration. A special shoutout goes to my colleagues from the MaD running and bouldering group for the breaks from work.

I want to thank Jesse Davis for hosting and supervising me at Katholieke Universiteit Leuven (KU Leuven), which allowed me to make a brief stay abroad, an opportunity that was not feasible for most of my Ph.D. tenure due to the pandemic.

Aside from academic support, I want to express my appreciation to my friends for the enjoyable holidays and evenings that helped me unwind during stressful times. It is a privilege to still go on summer vacations every year with friends from school, and I hope it continues for a long time.

Above all, I am grateful to my parents and sisters. It was challenging to spend most of my Ph.D. tenure during the COVID-19 pandemic, and I want to thank you for your unwavering support and for always being there for me. Lastly, I want to thank my girlfriend, Ulla, for all the quality time and the shared library sessions, which made finishing my Ph.D. much easier.
Abstract

Wearable devices, such as smartwatches, fitness trackers, and specialized tracking sensors, have become increasingly popular in sports in recent years. With more wearables used, the amount of data measured is also growing rapidly. Additionally, better computational power and improved algorithms have enabled Deep Learning (DL) methods to outperform traditional machine learning algorithms in many domains. In recent years, the combination of these two advancements has led to significant progress in sports analytics.

This advancement has proven advantageous for numerous sports, although some, particularly those that are less popular, have yet to reap its benefits. Within this dissertation, I contribute to DL applications in niche sports, particularly ski jumping and ultimate frisbee.

Issues related to insufficient data are one of the primary reasons for the failure of DL projects. This includes limited data availability or insufficient data quality. Therefore, it is essential to investigate and validate the data quality properly to be able to apply a data-driven engineering approach. Within this cumulative dissertation, I present and fill a literature gap regarding the experimental validation of wearables in sports and applying deep learning to wearables’ data in sports.

This cumulative dissertation is built around three contributions. In the first contribution [P1], I, with my co-authors, systematically validated a Wearable Real-Time Tracking System (WRTTS) for ski jumping. The system consists of two wearable trackers mounted on the skis and antennas next to the ski jumping hill, making it unobtrusive while also being able to transmit data in real-time. We validated the system’s measurements of the 3D positions, the skis’ 3D orientations, and the ski jump length in comparison with multiple reference systems.

Based on the validated system, my co-authors and I developed, in my second contribution [P2], a real-time performance prediction method for ski jumping. We investigated the prediction accuracy of different Deep Neural Networks (DNNs) shortly after the take-off. Additionally, we analyzed the prediction accuracy of the best performing DNN during the flight phase.

In my third contribution [P3], I developed, with my co-authors, a Human Activity Recognition (HAR) system based on wrist-worn Inertial Measurement Units (IMUs) for ultimate frisbee. The system is based on a Convolutional Neural Network (CNN) and can classify various throwing techniques and catches. Additionally, we investigated the use of transfer learning when dealing with small-scale datasets. For this, we transferred (low-)level features from a previously developed CNN for HAR in beach volleyball using a wrist-worn IMU.

In conclusion, this thesis shows the importance of proper experimental validation of wearables and demonstrates the possibilities of DL for innovative applications in niche sports based on wearables.
Zusammenfassung


Zusammenfassend zeigt diese Arbeit die Bedeutung einer angemessenen experimentellen Validierung von Wearables und demonstriert die Möglichkeiten von DL für innovative Anwendungen im Sport auf der Basis von Wearables.
Contents

List of Abbreviations .................................................. xi

1. Introduction .......................................................... 1
   1.1. Motivation ....................................................... 1
   1.2. Objectives and Contributions ................................. 2
   1.3. Structure ......................................................... 6

2. Fundamentals ........................................................ 7
   2.1. Sports .................................................................... 7
      2.1.1. Ski Jumping .................................................. 7
      2.1.2. Ultimate Frisbee ............................................ 8
   2.2. Tracking Technologies .......................................... 9
      2.2.1. Inertial Measurement Units ............................... 9
      2.2.2. Ultra-Wideband ............................................. 10
      2.2.3. Cameras ....................................................... 10
      2.2.4. Total Stations ............................................... 11
   2.3. Deep Neural Networks .......................................... 11
      2.3.1. Neural Network Architectures ........................... 12
      2.3.2. Supervised Training ....................................... 16

3. State of the Art ....................................................... 21
   3.1. Wearables in Sports ............................................. 21
      3.1.1. Application of Wearables in Sports .................... 22
      3.1.2. Validation of Wearables ................................ 27
   3.2. Deep Learning for Wearables in Sports ..................... 29
      3.2.1. Human Activity Recognition ............................... 29
      3.2.2. Performance Prediction ................................... 31

4. Discussion of Contributions ....................................... 35
   4.1. Experimental Validation of Real-Time Ski Jumping Tracking System Based on Wearable Sensors .................................................. 35
      4.1.1. Abstract of Original Publication .......................... 35
      4.1.2. Author Contributions ...................................... 35
      4.1.3. Discussion .................................................... 36
   4.2. xLength: Predicting Expected Ski Jump Length Shortly after Take-off using Deep Learning .................................................. 37
      4.2.1. Abstract of Original Publication ......................... 37
      4.2.2. Author Contributions ...................................... 37
      4.2.3. Discussion .................................................... 37
   4.3. Wearable Sensors for Activity Recognition in Ultimate Frisbee Using Convolutional Neural Networks and Transfer Learning ................. 39
      4.3.1. Abstract of Original Publication ......................... 39
      4.3.2. Author Contributions ...................................... 40
Contents

4.3.3. Discussion ................................. 40

5. Conclusion ...................................... 43
  5.1. Impact of this Thesis ......................... 43
  5.2. Findings .................................... 44
  5.3. Outlook .................................... 44

Appendix ........................................... 47
  A. Experimental Validation of Real-Time Ski Jumping Tracking System Based on Wearable Sensors ................................................................. 47
  B. xLength: Predicting Expected Ski Jump Length shortly after Take-off using Deep Learning ................................................................. 64
  C. Wearable Sensors for Activity Recognition in Ultimate Frisbee Using Convolutional Neural Networks and Transfer Learning .......................... 80

Bibliography ..................................... 99
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>Artificial Intelligence</td>
</tr>
<tr>
<td>AoA</td>
<td>Angle of Arrival</td>
</tr>
<tr>
<td>BLSTM</td>
<td>Bi-directional Long Short-Term Memory</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional Neural Network</td>
</tr>
<tr>
<td>DGNSS</td>
<td>Differential Global Navigation Satellite System</td>
</tr>
<tr>
<td>DL</td>
<td>Deep Learning</td>
</tr>
<tr>
<td>DNN</td>
<td>Deep Neural Network</td>
</tr>
<tr>
<td>FAU</td>
<td>Friedrich-Alexander-Universität Erlangen-Nürnberg</td>
</tr>
<tr>
<td>FCNN</td>
<td>Fully Connected Neural Network</td>
</tr>
<tr>
<td>FIS</td>
<td>Fédération Internationale de Ski (International Skiing Federation)</td>
</tr>
<tr>
<td>GNSS</td>
<td>Global Navigation Satellite System</td>
</tr>
<tr>
<td>GPS</td>
<td>Global Positioning System</td>
</tr>
<tr>
<td>HAR</td>
<td>Human Activity Recognition</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>IMMU</td>
<td>Inertial-Magnetic Measurement Unit</td>
</tr>
<tr>
<td>IMU</td>
<td>Inertial Measurement Unit</td>
</tr>
<tr>
<td>kNN</td>
<td>k-Nearest-Neighbor</td>
</tr>
<tr>
<td>LOOCV</td>
<td>Leave-One-Out Cross-Validation</td>
</tr>
<tr>
<td>LPS</td>
<td>Local Positioning System</td>
</tr>
<tr>
<td>LSTM</td>
<td>Long Short-Term Memory</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean Absolute Error</td>
</tr>
<tr>
<td>MEMS</td>
<td>Micro-Electrical-Mechanical System</td>
</tr>
<tr>
<td>MLP</td>
<td>Multilayer Perceptron</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Squared Error</td>
</tr>
<tr>
<td>NB</td>
<td>Naïve Bayes</td>
</tr>
<tr>
<td>ReLU</td>
<td>Rectified Linear Unit</td>
</tr>
<tr>
<td>ResNet</td>
<td>Residual Neural Network</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root-Mean-Square Error</td>
</tr>
<tr>
<td>RNN</td>
<td>Recurrent Neural Network</td>
</tr>
<tr>
<td>SGD</td>
<td>Stochastic Gradient Descent</td>
</tr>
</tbody>
</table>
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOTA</td>
<td>State of the Art</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>TDoA</td>
<td>Time-Difference-of-Arrival</td>
</tr>
<tr>
<td>ToA</td>
<td>Time-of-Arrival</td>
</tr>
<tr>
<td>UWB</td>
<td>Ultra-Wideband</td>
</tr>
<tr>
<td>VR</td>
<td>Virtual Reality</td>
</tr>
<tr>
<td>WRTTS</td>
<td>Wearable Real-Time Tracking System</td>
</tr>
<tr>
<td>xLength</td>
<td>Expected Ski Jump Length</td>
</tr>
</tbody>
</table>
1. Introduction

1.1. Motivation

The way athletes get feedback on their performance has changed a lot in the last decade, both for professional and recreational athletes. In the past, only professional athletes received qualitative information about their performance, while recreational athletes relied on their own subjective feelings. The rise of wearables has considerably changed this. Nowadays, the vast majority of recreational runners track their runs using a smartwatch, fitness tracker, or other wearables [1]. This allows them to track basic parameters like time and distance, as well as more advanced parameters such as running power.

Nevertheless, not only in recreational sports wearables are highly used, but also in professional settings, where athletes do not only use these in training but also in competitions. For example, in the German Handball-Bundesliga, all players are tracked using wearable devices [2]. The tracked parameters include primary indices like the distance covered and running speed but also more innovative metrics like the so-called Handball Performance Index. This performance index values every player’s action depending on their position on the field. Exemplarily, a goal scored from a closer distance is easier to throw and thus gives fewer points than a goal thrown from further away. In the end, every athlete gets a score with a maximum of 100, which makes the player’s performance very easy to interpret and compare, even for non-experts.

With more and more data being tracked, processing these large amounts of data appropriately is essential for easy-to-interpret performance measures. Otherwise, experts and non-experts will just be overwhelmed by the amount of data and can not benefit from it.

To develop such an easy-to-interpret performance metric, several steps are needed. First, the tracking system used to determine the athletes’ positions and additional parameters needs to be validated in an in-the-field setting, not just in the lab. While sensor systems developed and proposed in research publications are usually extensively validated, the accuracy and validity of commercially available sensors are often unknown to consumers. However, coaches and sports scientists rely heavily on these commercial sensors to determine factors like training load in order to improve injury prevention or performance.

Secondly, the actions of the players need to be classified. This can be done manually, which is cumbersome, and many people are needed to label all the players’ actions, or it can be done automatically. These classification algorithms are mainly based on camera or wearable sensor data and are an active research field [3–7].

Thirdly, all the data must be combined into one easy-to-understand performance metric or prediction. The Handball Performance Index is calculated using a relatively simple approach, i.e., valuing all actions by scores determined by an expert panel and adding these accordingly. However, with more and more data available, Deep Neural Networks (DNNs) can be trained to evaluate an athlete’s performance or predict the performance based on previous actions.

DNNs are a subset of machine learning. The application of machine learning has become very popular in analyzing data in recent years. The reasons for this are mainly the growth
of available data, computational power, and improved algorithms [8]. Deep Learning (DL), in particular, needs large amounts of data to perform well and has outperformed traditional machine learning algorithms in many areas [9–11].

Within this dissertation, I present three contributions [P1–P3] that cover the steps towards an easy-to-interpret performance metric. I exemplify how to properly validate a wearable tracking system and apply and investigate the use of DL algorithms in niche sports, exemplary in ski jumping and ultimate frisbee.

1.2. Objectives and Contributions

In this section, the objectives and contributions of this cumulative dissertation are presented. Figure 1 summarizes the objectives, respective publications, and contributions. The first contribution aims to validate a Wearable Real-Time Tracking System (WR-TTS) in ski jumping. The other two papers contribute new DL applications in sports based on wearables.

Objective 1: Experimental Validation of Wearable Real-Time Tracking System for Ski Jumping

Wearable tracking systems are heavily used in professional sports [12]. Their outputs are used both for entertainment and performance improvement. Parameters, such as the athlete’s speed [13], are extracted in real-time and displayed in TV broadcasting for entertainment purposes. Additionally, coaches and sports scientists use the measurements of the wearables to analyze the performance and gain additional information. With this information, they can help the athletes to improve their performance. Therefore, the tracking systems must be extensively validated before the extracted data can be used. This is crucial for coaches and sports scientists to advise athletes reasonably. However, also, to use the data for DL applications, validation is needed.

As described in more detail in Section 3.1.2, many studies validated commercially available wearables for team sports [14–23] and running [1, 24–26], covering various tracking technologies. Many team sports, such as soccer and handball, cover similar-sized areas and involve athletes with comparable speeds, making them suitable for the same tracking systems. However, for sports with a more discrete set of movements, such as ski jumping, individual tracking solutions must be developed. So far, no study has systematically validated a commercially available tracking system that is unobtrusive for use in ski jumping competitions. Therefore, the first objective of this thesis is to validate a wearable real-time tracking system for ski jumping experimentally.

My co-authors and I addressed this objective in the journal paper “Experimental Validation of Real-Time Ski Jumping Tracking System Based on Wearable Sensors” [P1]. In this work, we validated a WR-TTS for ski jumping. The system is based on Inertial Measurement Units (IMUs) and Ultra-Wideband (UWB) technology and measures the 3D trajectory, 3D velocity, orientation of the skis, and ski jump length. To my knowledge, it is the only tracking system capable of measuring these parameters. It is commercially used in all major ski jumping competitions.
1.2. Objectives and Contributions

<table>
<thead>
<tr>
<th>Wearables in Sports</th>
<th>Deep Learning Applications</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Objective 1</strong></td>
<td><strong>Publication 1</strong></td>
<td><strong>Contributions</strong></td>
</tr>
</tbody>
</table>
| Experimental Validation of WRTTS for Ski Jumping | Link et al. "Experimental Validation of Real-Time Ski Jumping Tracking System Based on Wearable Sensors". Sensors. 2021; 21(23):7780. | • Validated ski jump length measurements w.r.t. video-based system  
• Validated 3D position measurements w.r.t. cameras and theodolites  
• Investigated accuracy between WRTTS and camera measurements for skis’ 3D orientations |
| **Objective 2**     | **Publication 2**         | **Contributions** |
• Compared performance of different neural network architectures  
• Investigated accuracy of xLength during the ski jump |
| **Objective 3**     | **Publication 3**         | **Contributions** |
| Human Activity Recognition (HAR) in Ultimate Frisbee | Link et al. "Wearable Sensors for Activity Recognition in Ultimate Frisbee Using Convolutional Neural Networks and Transfer Learning". Sensors. 2022; 22(7):2560. | • Developed the first HAR system in ultimate frisbee  
• Investigated transfer learning for HAR in niche sports  
• Analyzed transfer learning for small-scale datasets |

Figure 1: Overview of this cumulative dissertation, which aims to validate wearables in sports tracking and develop new Deep Learning (DL) applications based on wearable sensor data.
The main contribution is the systematic in-the-field validation of a WRTTS for ski jumping. This includes:

- We validated the measurements of the ski jump length of the wearable tracking system with respect to the manually labeled Fédération Internationale de Ski (International Skiing Federation) (FIS) video-based system.
- We validated the 3D positions measured by the wearable system compared to camera measurements and measurements of the QDaedalus system (theodolite system).
- We investigated the accuracy of the measurements of the tracking system regarding the orientation of both skis in 3D with respect to camera measurements. This includes the pitch angle and V-style opening angle between both skis.

**Objective 2: Development of Real-Time Performance Prediction in Ski Jumping**

Some measurements of sports tracking systems may be intuitive to interpret also for non-experts, for example, that it is better to be faster during skiing downhill or take-off in ski jumping. Other parameters are harder to interpret or might not have a “perfect” value, but they might depend on the athlete’s anthropometry or technique. With modern tracking systems measuring multiple parameters, coaches may find it overwhelming to analyze the raw data.

Therefore, new innovative performance metrics are needed that make the tracking systems’ measurements more interpretable for experts and non-experts. By automatically scoring different parts of the athlete’s performance, the coaches can focus on the weaknesses of the athletes instead of having to analyze every parameter individually, such as the highly sampled position and velocity data during a ski jump.

As presented in more detail in Section 3.2.2, current literature on performance prediction in sports mainly covers the prediction of match results in various team sports [27–46]. Additional studies predict a ball’s trajectory based on its previous trajectory [47–51] or based on the athlete’s pose and movement [52–54]. Furthermore, studies investigated the correlation of different parameters in ski jumping with the ski jump length [55–58]. However, no study exists proposing a performance prediction in ski jumping. Therefore, the second objective of this dissertation is the development of a real-time performance prediction in ski jumping.

My co-authors and I addressed this objective in the journal paper “xLength: Predicting Expected Ski Jump Length Shortly after Take-Off Using Deep Learning” [P2]. In [P2], my co-authors and I contributed a novel DL approach for performance prediction in ski jumping. It is based on more than 2300 ski jumps of more than 200 athletes. The main contributions are:

- We developed a new innovative performance prediction in ski jumping called xLength. Using DL, the trajectory, speed, and skis’ 3D orientations are combined with metadata to predict the expected ski jump length. The metadata includes ski jump hill parameters, wind, start gate, and biological sex.
Objective 3: Human Activity Recognition (HAR) in Ultimate Frisbee

Sports analytics is part of all popular sports nowadays. This has several benefits for coaches and athletes. For example, by monitoring the actions of the athletes, high training loads can be detected, and the training can be adjusted for injury prevention. While many sports already benefit from sports analytics, less-known sports like ultimate frisbee were not covered yet.

To develop reliable activity recognition systems, large amounts of data are needed. Acquiring these may be easy for common sports but hard for less common sports, especially to achieve a homogeneous dataset over different ages, skill levels, etc. To combat this, transfer learning might help to generalize to new athletes with a skill or age profile not represented in the training data.

As described in more detail in Section 3.2.1, many studies investigated HAR in various sports [59–66]. Additionally, multiple studies found that DL approaches typically outperform traditional machine learning methods [6, 67–70]. However, no study proposed an HAR system for different frisbee throwing techniques. Therefore, the third objective of this dissertation is the development of an HAR system for ultimate frisbee.

My co-authors and I addressed this objective in “Wearable sensors for activity recognition in ultimate frisbee using convolutional neural networks and transfer learning” [P3]. In [P3], my co-authors and I contributed a DL approach for HAR in ultimate frisbee. To my knowledge, this is the first HAR system for ultimate frisbee. We conducted a study with 14 athletes. Thereby, an IMU mounted at the athlete’s wrist of the dominant hand measured acceleration and angular velocity. The main contributions are:

• We developed the first HAR system for ultimate frisbee. For this, we trained a Convolutional Neural Network (CNN) to classify seven different throwing techniques, catches, and a null class for other activities, such as running.

• We investigated the use of transfer learning for activity recognition in sports based on wearables. Exemplary, we pre-trained a CNN on beach volleyball data and fine-tuned it on ultimate frisbee data.

• We used the dataset conducted during the study without data augmentation techniques to investigate the generalization capabilities of transfer learning for small-scale datasets in HAR.
1. Introduction

1.3. Structure

The thesis is structured as follows. In Chapter 2, I present the fundamentals related to the contributions of this thesis. Section 2.1 introduces the basics of ski jumping and ultimate frisbee. Section 2.2 presents the basics of different tracking technologies, followed by the fundamentals of Deep Learning in Section 2.3. Chapter 3 presents and discusses the current literature and points out the respective research gaps. In Chapter 4, I present the contributions of this dissertation and discuss them in the context of the State of the Art (SOTA). Finally, in Chapter 5, the findings within the thesis are concluded, and an outlook is given. The contributed publications can be found in the Appendix.
2. Fundamentals

This chapter introduces fundamental terminology and methodology as a basis for the following chapters. It starts with a short introduction to ski jumping and ultimate frisbee, followed by an introduction and explanation of different tracking technologies. Lastly, the fundamentals of Deep Learning (DL), which is used to process the data acquired with wearables, will be introduced.

2.1. Sports

2.1.1. Ski Jumping

The contributions [P1] and [P2] cover the use of wearables in ski jumping. I will shortly introduce the basics of ski jumping here to provide context for the research covered in these publications.

Ski jumping is an individual sport originating from Norway [71]. The athletes, equipped with skis, go down the inrun to accelerate and jump off at the take-off table. The goal is to jump on designated ski jumping hills as far as possible. The ski jump is divided into the inrun, take-off, early flight phase, stable flight, landing preparation, and landing. Due to its impact on the horizontal and vertical velocity in the early flight phase, the take-off is considered the essential part [55–57]. The athletes reach around \( 25 \text{ m/s} \) depending on the starting gate and ski jumping venue during the take-off. Apart from the ski jump length, the flight style and landing are rated by judges and affect the final score.

Since the ski jump length heavily depends on the wind conditions, a wind compensation factor was introduced to make the competition fairer [72]. In addition to the wind compensation factor, a gate compensation factor was introduced to compensate for variations in the starting gate. Due to varying wind conditions, the judge might change the starting gate so that the athletes are either faster in bad wind conditions or slower when having good wind conditions. The final score is calculated from the ski jump length, judges’ scores, wind compensation, and gate compensation factors.

Ski jumping hills exist in different sizes. The so-called hill size is a measure of the size of a ski jumping hill. The hill size is defined as the distance between the edge of the take-off table to the end of the landing area, measured on the ski jumping hill. The end of the landing area is defined as the point where the ski jumping hill still has \( 32^\circ \) inclination. Ski jumping on ski jumping hills with a hill size over \( 185 \text{ m} \) is called ski flying. The largest ski flying hills have a hill size of \( 240 \text{ m} \).

Ski jumping is primarily popular in Europe and Japan. Due to the high risks and expensive venues required for the sport, it is only practiced by a small number of professional athletes and is not widely accessible as a grassroots activity.
2. Fundamentals

Figure 2: The three main throwing techniques in ultimate frisbee are visualized here. These are from left to right, the forehand, overhead, and backhand throw. Additionally, the sensor placement during the study is shown. The Inertial Measurement Unit (IMU) (white) is fixed with a wristband (black) at the dominant hand.

2.1.2. Ultimate Frisbee

The contribution [P3] applies wearables to different throwing techniques in ultimate frisbee. To contextualize the research described in this publication, I will shortly present the basics of ultimate frisbee.

Ultimate frisbee, or shortly ultimate, is a relatively new sport invented in 1968 [73]. It involves two teams of seven players each. Similar to American football, the goal is to catch the frisbee in the opposing endzone. However, in contrast to American football, ultimate frisbee is a non-contact sport, and the player currently holding the frisbee must keep one of their feet planted on the ground. Additionally, if covered by an opponent, the players only have ten seconds to pass the frisbee to the next player. Exceeding this limit or the frisbee touching the ground leads to a turnover.

The "Spirit of the game" is essential to ultimate frisbee. This means that fair play is the most important rule for everyone. This goes so far that ultimate frisbee is usually played without referees.

Also relatively unknown in Germany, according to the Sports and Fitness Industry Association [74], 5.1 million people played it in the U.S. alone in 2012, making it the country where ultimate frisbee is most popular.

Various throwing techniques result in different release points to throw around a direct opponent and differently curved trajectories. Figure 2 shows the three basic throwing techniques: forehand, overhead, and backhand throw. Additionally, depending on the release angle, the frisbee travels different curves. Therefore, throws can additionally be separated as inside-out, outside-in, and flat.
2.2. Tracking Technologies

Tracking technologies for sports applications can be separated into two major groups: wearable and non-wearable systems. Wearable systems consist of sensors attached to the athlete or the athlete’s equipment. Wearable trackers exist in various forms of appearance depending on the use case. In team sports like soccer or handball, trackers are usually worn with a vest underneath the jersey, such as Catapult\(^1\), Kinexon\(^2\), and STATSports\(^3\), and are mainly worn by professional athletes. Especially in endurance running, smartwatches are widespread also amongst amateur athletes, such as Garmin\(^4\) and Polar\(^5\).

Wearable sensors are based on different technologies or combinations of technologies and, therefore, can measure diverse parameters. IMUs are used to measure acceleration and angular velocity. This can be used, for example, to count the steps by extracting the repetitive movement. Ultra-Wideband (UWB) is common to measure an athlete’s position; therefore, anchors are needed in addition to wearable trackers. Another method to track athletes’ position is Global Positioning System (GPS), where no additional anchors are needed, but the position is determined through satellite communication. However, since no wearables using GPS are used within this dissertation, the working principle of GPS will not be further explained. The remaining, namely IMU and UWB, will be described in detail in the following subsections.

In contrast, non-wearable systems are stationary, for example, cameras or marker-based motion capture systems, such as Qualysis Arqus\(^6\). Depending on the use case, each tracking technology has advantages and disadvantages, which will be discussed in the following sections.

2.2.1. Inertial Measurement Units

IMUs are cheap, lightweight sensors integrated into most wearables. They are also integrated into mobile phones and can be used, for example, for gesture recognition or image stabilization. IMUs measure acceleration and angular velocity and therefore consist of two components, the accelerometer and gyroscope. The accelerometer measures the acceleration of the sensor in 3D, and the gyroscope the angular velocity around three axes. Inertial-Magnetic Measurement Units (IMMUs) are IMUs with an additional magnetometer integrated to measure magnetic fields. IMUs integrated in wearables are realized as Micro-Electrical-Mechanical Systems (MEMSs).

The angular velocity \(\omega\) is measured using the Coriolis force \(\vec{F}_c = -2m(\omega \times \vec{v})\) of a vibrating mass \(m\) that moves with speed \(\vec{v}\). The Coriolis force can be determined by measuring the deviation of the movement path of the vibrating mass. This operating mode is easy to produce and very small. To obtain measurements of angular velocity in 3D usually, three gyroscopes are integrated into one IMU. By integrating the angular velocity over time, the rotation of an object can be determined.

\(^1\) Catapult Group International Ltd, Melbourne, Australia
\(^2\) Kinexon GmbH, Munich, Germany
\(^3\) STATSports Group Limited, Newry, Northern Ireland
\(^4\) Garmin Ltd., Schaffhausen, Switzerland
\(^5\) Polar Electro Oy, Kempele, Finland
\(^6\) Qualisys AB, Gothenburg, Sweden
2. Fundamentals

The acceleration measured by the accelerometer consists of two parts: the gravitational acceleration \( g \) and the inertial acceleration \( \ddot{a}_i \). In this way, the orientation of the sensor with respect to the gravitation can also be determined when at rest or in constant velocity. To determine the inertial acceleration \( \ddot{a}_i \), the gravitational acceleration must be extracted from the measurements when in motion. As the gyroscopes, usually, three accelerometers are integrated into an IMU to measure acceleration in 3D.

Since the acceleration and angular velocity can be measured by the IMU independently of reference stations, the volume covered by an IMU is extensive. However, in general, MEMS IMUs are not suited to track the position accurately. This is because by double integrating the acceleration measured to position, all measurement errors accumulate over time, leading to a drift between the true and calculated position [75]. However, they are often used in sensor fusion algorithms together with other measurement methods, such as UWB, leading to more accurate and robust positioning solutions. Sensor fusion algorithms leverage different sensors’ complementary strengths to compensate for individual sensors’ limitations.

Within this thesis, IMUs are used within all contributions [P1–P3]. They are integrated into the Wearable Real-Time Tracking System (WRTTS) to track ski jumping and measure athletes’ movement patterns for activity recognition in ultimate frisbee.

2.2.2. Ultra-Wideband

UWB is a wireless communication technology that uses a large portion of the radio spectrum to transmit data over short distances with low power consumption. One application field of UWB is real-time localization. A UWB system consists of beacon nodes (anchors) and mobile nodes. UWB is an emerging technology used in industrial applications, for example, to track items in logistics [76].

Different measurement principles can be used for localization using UWB, which depending on the requirements, are chosen. The most common are Time-of-Arrival (ToA), Time-Difference-of-Arrival (TDoA), and Angle of Arrival (AoA). Electromagnetic waves travel with the speed of light, so the distance between the mobile node to the anchors can be determined by measuring the time the signal travels. Combining the distance measurements to at least three anchors leads to one distinctive position used in ToA. For TDoA, instead of the absolute time, the time difference between the arrival times at the nodes is used. For AoA, the anchors consist of multiple beacon nodes. The difference in arrival time between these nodes is used to determine the direction and the angle from which the signal of the mobile node was received.

In contrast to IMUs, the volume covered by UWB is limited due to the anchors. However, UWB can be an accurate standalone positioning solution.

UWB is a component of the wearables for ski jumping, validated in this dissertation [P1]. The wearables’ data is also used for a DL-based ski jump length prediction [P2].

2.2.3. Cameras

Cameras map the 3D world onto a 2D projection. They are widely used not only in sports contexts but also in surveillance in public areas.
A camera needs to be calibrated to determine real-world coordinates from pixel coordinates. Calibrating a camera is the process of determining the orientation and location of the camera as well as lens distortion parameters. To determine 3D real-world coordinates or distances, at least two calibrated cameras with shared field-of-view are needed.

The advantages of cameras are that they are unobtrusive and, in contrast to wearable sensors, multiple athletes or joints can be tracked simultaneously with just one camera. However, they cover only a limited volume (field of view) and the camera data processing is much more computationally expensive and complicated compared to, for example, IMUs.

In this dissertation, calibrated cameras are used as a reference system to validate the measurements of the WRTTS during ski jumping [P1].

### 2.2.4. Total Stations

A total station is a combination of a theodolite and a distance measure. This enables measuring the distance and vertical and horizontal angles between an object and the total station. Total stations are commonly used in surveying, for example, on construction sites. Within this thesis, a QDaedalus system is used [77, 78]. Thereby, a charge-coupled device camera replaces the eye-piece of a total station to automate the measuring process.

By combining multiple QDaedalus stations, the 3D position of an object can accurately be triangulated. The advantages of the QDaedalus system are that it is unobtrusive and can cover large volumes. One disadvantage is that a clear view between the object to track and the QDaedalus stations is necessary.

Within this dissertation, the QDaedalus total stations are used as a reference system to validate the skis’ 3D positions during ski jumping measured with the WRTTS [P1].

### 2.3. Deep Neural Networks

Wearables are increasingly used in professional sports and by amateur athletes, for example, in fitness trackers or smartwatches. As a result, the amount of data available for analysis in this field is rapidly growing. Machine learning is one of the most common approaches for analyzing large datasets. However, with traditional machine learning (e.g., Support Vector Machine (SVM), random forest), meaningful features must be defined, which is complicated and requires domain experts. With Deep Neural Networks (DNNs), a machine learning subdomain, no prior feature definition and feature selection are necessary. Also, preprocessing is often not needed. In addition, DNNs outperformed traditional machine learning approaches in many application areas, like image classification [9] or disease diagnosis [10, 11]. Because of these reasons, DNNs are more and more applied to analyze wearables’ data.

Machine learning algorithms, in general, and DNNs can be divided into three main types based on their training process. During supervised learning, the DNN receives the correct output for every input sample. The deviation between the correct output and the output of the DNN is used to update the weights and train the DNN. In unsupervised learning, no correct output is provided to the corresponding input sample. Unsupervised learning is often...
2. Fundamentals

used for clustering or anomaly detection. The third option is reinforcement learning, where an intelligent agent interacts with its environment. The agent tries to maximize its rewards, which depend on the actions. Possible applications for reinforcement learning are natural language processing and autonomous driving. Since only supervised learning is used within this dissertation, the remainder of this work focuses on this paradigm.

In this dissertation, I use DNNs for a classification and a regression problem. The classification task is Human Activity Recognition (HAR) in ultimate frisbee [P3], and the regression task is a ski jump length prediction [P2].

2.3.1. Neural Network Architectures

DNNs can consist of various layer types. Depending on the layer type, the network architecture is named differently. The most common architecture are Multilayer Perceptrons (MLPs), which consist of fully-connected layers [79, 80] and, therefore, are also called Fully Connected Neural Networks (FCNNs). Another widespread type, especially in image processing, are Convolutional Neural Networks (CNNs), which consist of convolutional layers followed by fully-connected layers [81, 82]. Recurrent Neural Networks (RNNs) consist of layers, where nodes can have connections to themselves affecting subsequent inputs to the same nodes [83–85]. Therefore, RNNs are very common in analyzing time series data. All these different architectures will be introduced in the following.

2.3.1.1. Multilayer Perceptron

An MLP consists of three different types of layers. Firstly, the input layer is where the input data is fed in. One or more hidden layers follow this layer type. An MLP with multiple hidden layers is considered a DNN. The last layer of the MLP is the output layer which calculates the output of the MLP.

Figure 3 visualizes an MLP with two hidden layers. Thereby, every node of one layer is connected with every node of the previous and subsequent layer. The connections between the nodes represent the weights.

Every layer in an MLP consists of multiple nodes (also called neurons). Every node computes a weighted sum of its inputs $z_i$

$$a_j = \sum_{i=1}^{N} w_{ji} z_i + b_j,$$

where $N$ is the number of nodes in the previous layer, $w_{ji}$ the weight of the connection from node $i$ to node $j$ and $b_j$ the bias.

The output or activation $z_j$ of every node is calculated by a non-linear activation function $g(\cdot)$ of the weighted sum $a_j$

$$z_j = g(a_j).$$
2.3. Deep Neural Networks

![Diagram of a Multilayer Perceptron (MLP) with two hidden layers. Every node of one layer is connected with every node of the following layer. The connections between nodes illustrate the weights $w_{ji}$.](image)

The weights $w_{ji}$ are determined during the training of the neural network (see Section 2.3.2). Popular activation functions are the Rectified Linear Unit (ReLU) which is defined as:

$$g_{\text{ReLU}}(a_j) = \max(0, a_j) = \begin{cases} a_j & \text{if } a_j > 0 \\ 0 & \text{otherwise} \end{cases},$$

the sigmoid function

$$g_{\text{sigmoid}}(a_j) = \frac{1}{1 + e^{-a_j}},$$

and the hyperbolic tangent

$$g_{\text{tanh}}(a_j) = \tanh(a_j).$$

The softmax function is commonly used as the activation function for the output layer of classification tasks. It is calculated as

$$g_{\text{softmax}}(a_j) = \frac{e^{a_j}}{\sum_{i=1}^{C} e^{a_i}},$$

where $C$ is the number of classes. This network output can be interpreted as a class probability.

Hornik (1991) showed that one hidden layer could approximate any continuous function. However, this is not applied since this hidden layer would require many nodes. Arranging the nodes in multiple layers is, therefore, the much more efficient and also more applicable way.

For inputs with large dimensions (e.g., images), the computational cost of MLPs gets enormous. This is due to many connections, meaning weights between the input nodes and nodes of the fully connected layer. This is especially problematic since the number of arithmetic operations grows quadratically with the number of nodes.
2.3.1.2. Convolutional Neural Network

Another popular neural network architecture is a CNN [81, 82]. Due to their structure, they are heavily used for image processing and time series data. One advantage of CNNs is that the spatial respectively temporal proximity in data is conserved. This can be used in subsequent layers to combine multiple low-level features in close proximity to high-level features. These high-level features are again combined to detect local patterns.

Another advantage over MLPs is that CNNs have fewer trainable parameters, reducing the computational cost. Additionally, the convolutions are invariant against translations that enable, for example, the detection of an object anywhere in an image.

CNNs usually contain several convolutional layers followed by a flattening layer and fully connected layer(s). The flattening layer transforms the multi-dimensional feature maps into a one-dimensional vector that can be used as an input for the following fully connected layer.

Figure 4 shows the convolution process within a convolutional layer. A filter of size $3 \times 3$ is convoluted with the $4 \times 4$ input resulting in a $2 \times 2$ feature map. The weights of the filter are learned during the training process. By only learning the filter weights, the number of parameters in a convolutional layer is independent of the input dimensions. One convolutional layer usually consists of multiple filters.

![Figure 4: Convolution of a 3x3 filter with a 4x4 input, resulting in a 2x2 feature map. In a convolutional layer, the 3x3 filter weights are learned.](image)

Another layer type often used in CNNs are pooling layers. They are usually used after convolutional layers to reduce the size of the feature maps. The pooling layer reduces and combines the information of a subset of the feature map into one value. Common is using the maximum value, so-called max pooling, or the average value, so-called average pooling. Local pooling combines a small number of nodes, for example, $2 \times 2$, and global pooling combines all nodes into one. Pooling layers reduce computational costs, and, for example, to detect an object in an image, the exact pixel location of the object is often unnecessary.
2.3. Deep Neural Networks

2.3.1.3. Recurrent Neural Network

RNNs are a form of neural networks used for sequential or time series data [83–85]. They have connections between nodes that can create circles, and nodes can have connections to themselves. This self-connection works as an internal state or memory and affects the network’s output. That means that the network might produce different outputs for the same input depending on the previous inputs. RNNs are commonly used for natural language processing [87, 88], and speech recognition [89, 90].

RNNs can process inputs of variable length and produce outputs of variable length. So depending on the use case, one output may produce many outputs (one-to-many), many inputs just one output (many-to-one), or many inputs many outputs (many-to-many).

Figure 5 shows one hidden recurrent node with input and output nodes. Additionally, the unfolded version is shown. It is essential to mention that the unfolded version does not show layers as in the MLP but different steps in a sequence.

In contrast to MLPs, the training process of an RNN also propagates through time (see unfolded RNN). This introduces several problems. Common problems of RNNs are exploding gradients and vanishing gradients [91]. If the gradient gets too tiny (vanishing gradient), the network’s weights are almost not updated, and the network learns very slowly or stops learning. The opposite is the exploding gradients problem, where the gradient becomes too large, leading to an unstable model that is unable to learn.

A particular form of an RNN is a Long Short-Term Memory (LSTM). It was developed by Hochreiter and Schmidhuber (1997) and overcame the problem of vanishing gradients. Intuitively, the LSTM learns which information in a sequence is important to remember and which information can be safely forgotten.

Figure 6 shows an unfolded LSTM cell with its components. The \( \sigma \) stands for the sigmoid function (Equation 4), tanh for the hyperbolic tangent function (Equation 5), and \( \times \) for the element-wise product. One LSTM cell consists of an input gate, an output gate, and a forget gate. The cell at a step in a sequence \( t \) has three inputs the input vector \( x_t \) from the previous
Figure 6: Diagram of an unfolded LSTM unit over several inputs \( x_t \), with \( h_t \) being the hidden state, and \( c_t \) the cell state. Furthermore, \( \sigma \) stands for the sigmoid function, tanh for the hyperbolic tangent, and \( \times \) for the element-wise product.

layer, the cell state vector \( c_{t-1} \) from the previous step in the sequence, and the hidden state vector \( h_{t-1} \), which is also the previous output vector of the LSTM cell.

The forget gate concatenates the hidden state from the previous step \( h_{t-1} \) and the recent input \( x_t \) to decide which elements of the cell state \( c_{t-1} \) to forget and which to keep. The input gate also combines the hidden state from the previous step \( h_{t-1} \) and the recent input \( x_t \) and decides which of the data to store in the cell state \( c_t \). Finally, by combining the previous hidden state \( h_{t-1} \), the updated cell state \( c_t \), and the input \( x_t \), the output of the LSTM is determined. This output is also the updated hidden state \( h_t \). Therefore, LSTMs can learn tasks where memory needs to be stored over very long sequences.

Within this dissertation, I compared the performance of an FCNN, two CNNs, and an LSTM to predict the ski jump length [P2]. Additionally, I trained a CNN for HAR in ultimate frisbee [P3].

2.3.2. Supervised Training

During the training of a neural network, the nodes’ weights are adapted so that the network predicts the desired output for a given input-output set. This is achieved by minimizing a loss function.

2.3.2.1. Loss function

Different loss functions are used depending on the task the model should learn. The loss function measures the difference between the predicted output of the network and the target
output. For regression problems, the Mean Squared Error (MSE) is a very common loss function and is calculated as

$$\text{MSE}(\hat{Y}, Y) = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2,$$  \hspace{1cm} (7)

where $\hat{Y}$ is the network’s prediction, $Y$ is the true value, and $N$ is the number of samples in the training (sub-)set.

For multi-class classification, the categorical cross-entropy is a widespread loss function

$$\text{CE}(\hat{Y}, Y) = - \sum_{i=1}^{C} y_i \cdot \log(\hat{y}_i).$$  \hspace{1cm} (8)

Within this dissertation, I used the MSE as a loss function for the ski jump length prediction and the categorical cross-entropy for the classification of activities in ultimate frisbee.

### 2.3.2.2. Backpropagation

A neural network learns by adapting its weights according to a loss function. The weights are adapted using backpropagation. The following section describing the backpropagation process is an adapted version of Bishop (1995).

Using the chain rule, we can calculate the gradient of a loss function $E$ with respect to a weight $w_{ji}$ as

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial a_j} \frac{\partial a_j}{\partial w_{ji}}.$$  \hspace{1cm} (9)

For the second factor, using Equation 1 we obtain

$$\frac{\partial a_j}{\partial w_{ji}} = z_i.$$  \hspace{1cm} (10)

Introducing $\delta_j$

$$\delta_j = \frac{\partial E}{\partial a_j}$$  \hspace{1cm} (11)

we get

$$\frac{\partial E}{\partial w_{ji}} = \delta_j z_i.$$  \hspace{1cm} (12)

We see that the derivative of the gradient of a loss function $E$ with respect to a weight $w_{ji}$ can be determined by only calculating $\delta_j$ for each node and multiplying it with the input $z_i$. 

By using $g' = \frac{\partial g}{\partial a_k}$, we get for the error of the output nodes

$$\delta_k \equiv \frac{\partial E}{\partial a_k} = g'(a_k) \frac{\partial E}{\partial z_k}. \quad (13)$$

For the hidden nodes, the error can be determined using the chain rule for partial derivatives and by summing over the nodes $k$ to which node $j$ sends its output to

$$\delta_j = g'(a_j) \sum_k w_{kj} \delta_k. \quad (14)$$

This backpropagation formula is applied during training to calculate the derivative of the loss function with respect to the weights.

The weights are then updated according to

$$w_{ji}^{\text{new}} = w_{ji}^{\text{old}} + \Delta w_{ji} = w_{ji}^{\text{old}} - \eta \delta_j z_i, \quad (15)$$

where $\eta$ is the so-called learning rate. This technique is called gradient descent, and the weights are iteratively adapted, stepping in the opposite direction of the gradient towards the minimum.

Typically, the backpropagation is not done for every sample individually but for a random subset, called a batch, of samples. Then the previous formula results in

$$w_{ji}^{\text{new}} = w_{ji}^{\text{old}} - \sum_n \eta \delta_j^n z_i^n, \quad (16)$$

where $n$ are all samples included in the batch. This combination of batch training and gradient descent is called Stochastic Gradient Descent (SGD).

This dissertation uses the Adam optimizer [94] to train the DNNs. The Adam optimizer is an extension of SGD and adapts the learning rate according to the initial moments of the gradients.

One cycle of backpropagation over all the training samples is called an epoch. The most straightforward approach to stop the training process is when a predefined number of epochs is reached. This, however, has the disadvantage that this predefined number is randomly chosen and may be too small or too large. Therefore, more sophisticated approaches were developed. One is to check whether the loss on the validation set reaches a defined threshold. Another is to check if the validation loss is on a plateau or increases again. I terminated the training process for both DL applications within this dissertation when the loss function reached a plateau, which is explained in more detail in my contribution [P3].

### 2.3.2.3 Data Splitting

The data is usually split into a training, validation, and test set during the training process. The network’s weights are adapted by calculating the loss function and backpropagating the errors on the training set. Afterward, the network’s performance with these weights is evaluated on a
2.3. Deep Neural Networks

Validation set. This is repeated for multiple epochs. Finally, the performance of the network is evaluated on the test set.

Another method of data splitting is cross-validation, where the data is split into $k$ folds. Then the network is trained on $k - 1$ folds and evaluated on the remaining one. This is repeated $k$ times, with every fold being the test set once. Cross-validation has the advantage that every sample is used for training and evaluation. One drawback of cross-validation is the much higher computational cost. This is especially true for the Leave-One-Out Cross-Validation (LOOCV), where $k$ equals the number of samples.

When the training process is run for too few epochs or the network architecture is too small (layer- or node-wise), the network cannot correctly represent the training data. This leads to a high training loss and is called underfitting. The opposite of this is overfitting, which occurs when the network architecture is too complex for the training dataset. Another reason for overfitting is when the training is run for too many epochs. Overfitting can be detected when the training loss is much smaller than the validation loss. This means that the network cannot generalize to unseen data samples.

There are several techniques to combat overfitting. One is dropout, where a specific rate of the network’s weights is randomly selected and set to zero during training. This equals dropping these connections between the nodes.

Another method against overfitting is data augmentation. Through artificially created data, the training dataset is increased, and the generalization capabilities of the network are improved. Depending on the data, different data augmentation techniques are available. For images, for example, it is common to crop, rotate, shift, and flip images as data augmentations.

In the case of HAR in ultimate frisbee [P3], we applied data augmentation to IMU data by randomly shifting the time window of the peak detection, random rotations in 3D, and adding normally distributed noise to the input data.

In my contribution of a ski jump length prediction [P2], the different time series were strongly dependent on each other and contained physical processes, making it challenging to augment the data. Therefore, I doubled the dataset by mirroring the left and right sides of the time series data of the ski jump trajectories and added normally distributed noise to the input data.

Apart from the weights of all nodes, additional parameters need to be chosen and optimized, such as the number of layers and nodes. The process of optimizing these is called hyperparameter tuning.

One problem with hyperparameter tuning is that it leads to overfitting the dataset and an optimistically biased performance estimation of the DNN on the dataset [95]. To overcome this problem, nested cross-validation can be used. It consists of an inner loop of cross-validation and an outer loop of cross-validation. In the inner loop, a model is trained on the training set, and hyperparameters are optimized on the test sets. In the outer loop, the model’s performance is evaluated on the test set, leading to a more realistic performance estimate for a later inference of unseen data.

Within this dissertation, I used a leave-three-subjects-out cross-validation for the HAR in beach volleyball [P3] to ensure comparability with the original publication [67]. To evaluate
the network’s performance on the ultimate frisbee dataset, I performed a leave-one-subject-out cross-validation. For the prediction of the ski jump length [P2], I used a nested 5-fold cross-validation.

2.3.2.4. Transfer Learning

Transfer learning describes the process of training a neural network on one domain (source domain) and applying it to another domain (target domain). The idea is that the same (low-level) features essential in the source domain might also be relevant in the target domain.

Transfer learning can improve the network’s performance through better generalization capabilities and convergence and reduces the training time. However, transfer learning can also decrease performance, so-called negative transfer [96].

Transfer learning can be applied by transferring all layers or only some of the layers (e.g., low-level features). Additionally, these transferred layers can be excluded from the training process afterward (frozen) or included (fine-tuned).

In computer vision, pre-trained weights are very popular due to the deep networks, which would take much longer to train.

Within this dissertation, I investigated transfer learning for a CNN trained on a beach volleyball dataset [67] to an ultimate frisbee dataset [P3]. I analyzed the influence of transfer learning on the CNN’s performance and training time depending on the number of frozen layers.
3. **State of the Art**

This chapter presents the SOTA approaches relevant to this dissertation. Figure 7 shows an overview of this SOTA chapter and how it is structured. I first focus on the application and validation of wearables in sports in Section 3.1 and then introduce different Deep Learning (DL) methods that have been applied to the data of wearables in sports in Section 3.2. For every section, I will start with a general overview of different sports and then focus on the ones covered in this dissertation. For every subsection, I will point out limitations of the current SOTA and which research gaps this introduces.

<table>
<thead>
<tr>
<th>3.1 Wearables in Sports</th>
<th>3.1.1 Application of Wearables in Sports</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.1.2 Validation of Wearables</td>
</tr>
<tr>
<td>3.2 Deep Learning</td>
<td>3.2.1 Human Activity Recognition</td>
</tr>
<tr>
<td></td>
<td>3.2.2 Performance Prediction</td>
</tr>
</tbody>
</table>

Figure 7: Overview of the structure of this SOTA chapter. First, different applications of wearables in sports are presented, followed by the validation of wearables in sports. The subsequent section handles the application of DL on wearable data in sports, including HAR and performance prediction.

### 3.1. **Wearables in Sports**

The use of wearable sensors in sports is a relatively recent development, with research in this area only beginning to emerge in the last two decades. Figure 8 shows the number of publications per year dealing with "wearable" and "sports". We can see that this area of research began in the early 2000s, and the number of publications per year has considerably risen over the last two decades.

Due to the large number of publications in this field, I will present a selection of the most relevant ones to gather an overview of the SOTA and how my contributions extended this.
3. State of the Art

Figure 8: Number of publications per year dealing with wearables in sports. This plot was created using the Web of Science search engine for "wearable" and "sports". The number of results from a topic search was counted per year (as of May 31, 2023).

3.1.1. Application of Wearables in Sports

Nowadays, wearables have been applied to almost every sport. Table 1 shows an overview of different sports with the respective number of publications applying wearables. The table was created using the Web of Science search engine, performing a topic research using "wearable" and the specific sport without filtering for publication years (as of May 31, 2023). Looking at the table, several points need to be mentioned. Firstly, we see that numerous different sports have been covered. Secondly, the vast majority of studies are in the field of running. Additionally, there are other very popular sports with many publications, like soccer, basketball, and swimming. Finally, the number of publications for less popular sports like ultimate frisbee or field hockey is very low, so they do not benefit from the current technological developments yet. I want to emphasize that for ultimate frisbee, one of the two publications is my contribution [P1], and two of the seven publications in ski jumping are my contributions [P1], and [P2].

These studies cover, amongst others, the automatic detection of pitching and throwing actions in baseball [97], monitoring of bad posture during dribbling in basketball [98], observing kinematic changes with fatigue in running [99]. Other studies investigated the computation of throwing speed in handball [100], the quantification of alpine skiing quality [101], and the determination of training and competition load in volleyball [102]. This exemplary selection shows the diverse use of wearable technology in sports, covering a variety of sports and exploring various topics.
Table 1: Overview of studies using wearables in different sports. The number of publications per sport was determined using the Web of Science search engine. Topic research was performed using "wearable" and the specific sport independent of the publication year (as of May 31, 2023).

<table>
<thead>
<tr>
<th>Sports</th>
<th>Number of Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpine skiing</td>
<td>9</td>
</tr>
<tr>
<td>American Football</td>
<td>20</td>
</tr>
<tr>
<td>Badminton</td>
<td>21</td>
</tr>
<tr>
<td>Baseball</td>
<td>36</td>
</tr>
<tr>
<td>Basketball</td>
<td>60</td>
</tr>
<tr>
<td>Cricket</td>
<td>15</td>
</tr>
<tr>
<td>Cross-country skiing</td>
<td>12</td>
</tr>
<tr>
<td>Field Hockey</td>
<td>6</td>
</tr>
<tr>
<td>Football/Soccer</td>
<td>202</td>
</tr>
<tr>
<td>Golf</td>
<td>20</td>
</tr>
<tr>
<td>Handball</td>
<td>13</td>
</tr>
<tr>
<td>Ice Hockey</td>
<td>26</td>
</tr>
<tr>
<td>Lacrosse</td>
<td>8</td>
</tr>
<tr>
<td>Rowing</td>
<td>44</td>
</tr>
<tr>
<td>Rugby</td>
<td>61</td>
</tr>
<tr>
<td>Running</td>
<td>1114</td>
</tr>
<tr>
<td>Ski Jumping</td>
<td>7</td>
</tr>
<tr>
<td>Swimming</td>
<td>56</td>
</tr>
<tr>
<td>Table Tennis</td>
<td>18</td>
</tr>
<tr>
<td>Tennis</td>
<td>43</td>
</tr>
<tr>
<td>Ultimate Frisbee</td>
<td>2</td>
</tr>
<tr>
<td>Volleyball</td>
<td>26</td>
</tr>
</tbody>
</table>
3. State of the Art

3.1.1. Ski Jumping

Due to my contributions [P1] and [P2], that cover the use of wearables in ski jumping. I will present the current SOTA of the application of wearables in ski jumping. Several studies tracked athletes in ski jumping with different technologies and measured various parameters, with some of these systems being more and others less obtrusive.

Chardonnens et al. (2012) used four Inertial Measurement Units (IMUs) attached to the thighs and shanks of the athletes to measure the duration of different ski jumping phases automatically. This included the take-off release, take-off, and early flight phase. They evaluated the system indoors and outdoors against a motion capture system and manual labeling by coaches, respectively. They found an error comparable to the variability of the coaches’ assessments. In a follow-up study, Chardonnens et al. (2013) used five IMUs attached to the lower limbs to measure the inter-segment coordination of the lower limbs. They then correlated these to the achieved ski jump length using the continuous relative phase. They found that the athletes with their thighs leading their shanks for longer achieved longer ski jump lengths. Additionally, their movement between the thighs and sacrum was more symmetric. Using an extended setup consisting of seven IMUs, Chardonnens et al. (2013) demonstrated how to measure the kinematics during the entire ski jump. Therefore, the athletes wore underwear suits with pockets for the IMUs at the sacrum, thighs, and shanks. One IMU was also mounted on each ski behind the binding. They validated the measurements by comparing the skis’ orientation with the landing hill’s inclination. Additionally, they compared various measured angles (e.g., hip angle and shank-ski angle) with values from literature and found similar values, although this does not provide adequate validation. Furthermore, with the multiple IMUs at the athlete’s body, the system is obtrusive and not applicable in competitions. Additionally, the system lacked the measurement of in-flight positions and velocities.

Chardonnens et al. (2014) used the same setup as in [105] consisting of seven IMUs. The authors combined and extended their prior works [103–105] to one system capable of calculating various parameters during take-off and stable flight. The system automatically detects the flight phases as in [103] and measures different parameters depending on the phase. During the take-off, the position, velocity, and forces of the center of mass perpendicular to the take-off table are measured. Additionally, during the stable flight, the aerodynamic forces are measured. They compared these measurements between different athletes and found a correlation between some parameters and the ski jump length.

A further extended setup was presented by Logar and Munih (2015), who used ten IMUs, attached to the body and skis, as well as force-plates to measure joint forces and moments during the inrun and take-off. They investigated simulated jumps indoors with marker-based motion capture as a reference. Further, they investigated jumps on a ski jumping hill with a force plate in the take-off table.

In contrast to these extensive setups that are obtrusive and not applicable in competitions, previous work by Groh et al. (2014) used one IMU behind the binding of each ski to measure the 3D orientations during the whole jump. For this, they integrated the angular rates using a quaternion-based approach and achieved a Root-Mean-Square Error (RMSE) of 2.0° for the right ski and 9.3° for the left ski in comparison with a high-speed camera. They explained this difference through the different camera perspectives of the two skis. While the IMUs...
on the skis were unobtrusive, the system lacked the measurement of in-flight positions and velocities. Further investigating the possibilities of such reduced setups, Groh et al. (2017) used Inertial-Magnetic Measurement Units (IMMUs) mounted before the ski’s bindings to determine the take-off velocity and the ski jump length. Their pipeline consisted of a hierarchical Hidden Markov Model (HMM) to segment the different phases of ski jumping from the start of the inrun to the landing. The velocity was determined by integrating the acceleration data and incorporating additional knowledge about the constraints through the inrun. They calculated the ski jump length with a motion model of accelerated movements and the parameters of the ski jumping hill. They compared the results to a light barrier system and manually labeled video-based ski jump length measurements. Thereby, they reached an accuracy of $-0.78 \text{ m/s} \pm 1.18 \text{ m/s}$ for the take-off velocity and $0.8 \text{ m} \pm 2.9 \text{ m}$ for the ski jump length. In a follow-up study, Groh et al. (2018) proposed a system to measure the landing momentum in ski jumping. The system consisted of three IMMUs mounted on the skis, two on the left and one on the right. Additionally, they used two force plates under each binding as a reference system. Their study with four jumps showed promising results for the landing momentum estimation based on IMMUs.

Apart from IMUs, pressure insoles have also been applied to ski jumping. A study by Bessone et al. (2019) estimated the ground reaction forces during the landing of ski jumping using wireless pressure insoles. They found that the ground reaction force is independent of performing a telemark or parallel leg technique. In the second part, they equipped one athlete with an additional eleven IMUs attached to the body and skis of the athlete to investigate a correlation between kinematics and kinetics during landing exploratively. The authors mentioned the low sampling rate of the pressure insoles of 100 Hz as a limitation of this study. Additionally, they noted that having only one subject for the second part of the study could influence the results since the landing biomechanics depend on gender, expertise, and age [112]. Furthermore, they emphasized that the IMUs mounted under the ski jumping suit could be moved during dressing, resulting in incorrect outcomes. This could also affect similar study setups such as by Chardonnens et al. (2013).

Fang et al. (2020) proposed a pipeline to reconstruct the trajectory of the ski jump. For this, they used a Global Positioning System (GPS) logger on the helmet and the IMU and magnetometer of an iPhone 7 attached to the athlete’s right arm to measure the position and velocity. They applied an extended Rauch-Tung-Striebel smoother [114] to the raw data and included geometric information of the ski jumping hill’s shape. They compared their pipeline to simulated results. They only compared the determined ski jump length for the real measurements but not the trajectory. Compared to video-based recordings, they found a RMSE below 1.5 m for the estimated ski jump length.

Other previous works incorporated a Differential Global Navigation Satellite System (DGNSS) with camera-based pose estimation to estimate the kinematics and kinetics during the inrun and flight phase [113]. Details of the agreement between both systems can be found in Section 3.1.2. Additionally, they made a proof of concept for a sensor fusion approach between the pose estimation data and DGNSS data. For the proof of concept of the sensor fusion, the authors compared two jumps of the same athlete with similar wind conditions. They then investigated the influence of parameters such as knee angle and hip angle. In additional studies using DGNSS, researchers determined the aerial phase in ski jumping and the steady glide phase [116],...
They compared the correlation between different parameters of the steady glide phase, such as lift-to-drag-ratio, vertical and horizontal acceleration, and the ski jump length as a performance metric. They found that depending on the hill size, the correlation between the parameters differs significantly, and the athletes should set a different focus depending on the hillsize.

In addition to the studies presented, various other studies used camera-based systems to investigate ski jumping. Among other things, they investigated the take-off \cite{57, 118}, flight styles \cite{119}, ski jumping phases \cite{103, 120, 121}, dynamics \cite{122}, and aerodynamic forces \cite{123}. Since this dissertation focuses on wearables, I do not present their methods and results.

3.1.1.2. Ultimate Frisbee

Due to my contribution \cite{P3}, which covers the use of wearables in ultimate frisbee, I will present the current SOTA of the application of wearables in ultimate frisbee. Even though the number of athletes playing ultimate frisbee is much higher than for ski jumping \cite{74, 124}, only very few studies exist where wearable sensors are applied in ultimate frisbee.

In a study, Slaughter and Adamczyk (2020) investigated cutting maneuvers within competitive ultimate frisbee games. The athletes wore seven IMUs at the lower body (hips, legs, feet) to reconstruct joint angles, ground contact, pose, and position of the athletes. They automatically detected cuts, extracted the cutting angle and speed, and found that the speed and acceleration did not change with the cutting angle. However, athletes from more competitive teams had higher speed and acceleration during cutting.

Other studies did not use body-worn IMUs but mounted the wearable on the frisbee. Weizman et al. (2020) mounted a gyroscope on a frisbee to investigate the flight behavior. They showed that during the flight, the wobbling of the frisbee gets less, similar to a damped vibration. Additionally, the change of the angular velocity and torques over time were presented, but no further processing was done. Similarly, Solomon et al. (2014) equipped a frisbee with a force-sensitive resistor and an IMU to give the athlete optical and haptic feedback. With the force-sensitive resistor, they determined the grip strength and, with the IMU, the angle of release, along with flight information such as rotation speed and time of flight. Even though different throwing techniques were included in the study, no classification of these was performed.

To conclude, wearables are nowadays a commonly used tool in sports to track the performance of athletes. Many studies exist that use wearables to investigate ski jumping. However, only tracking systems for specific research studies have been developed for ski jumping to investigate particular variables. Additionally, they are mostly not applicable in practice since they are obtrusive. The high risks inherent in ski jumping make unobtrusive tracking systems crucial for athletes to maintain their focus without interruption during competitions and training. This can be achieved by wireless data transmission and wearables attached to the skis instead of body-worn sensors. For ultimate frisbee, only two studies existed with a wearable on the frisbee that rather focused on describing the measured signals over time. Apart from automatic cutting detection and retrospective analysis, no activity recognition system existed for ultimate frisbee \cite{125} and not for different throwing techniques. However, HAR is needed to automate the process of performance monitoring, technique comparison (between athletes and within
3.1 Wearables in Sports

3.1.2 Validation of Wearables

When new tracking methods are developed in a research study, they are compared with an alternative or gold standard. However, these tracking methods are usually prototypes to demonstrate a new measurement procedure or algorithm, not a commercially available product. Even though this prototype might be appropriately validated, error sources might be added during the final product’s development process. One reason for this could be the use of other components or adaptations to the algorithms due to the optimization of power consumption. Another reason might be adaptations for real-time data analysis instead of retrospective analysis. Therefore, it is essential to also systematically validate the accuracy of commercially available products, which is also done in multiple studies.

Wearables are commonly used to track, among other parameters, the position of athletes. In their systematic review, Pino-Ortega et al. (2022) compared various studies validating commercially available Local Positioning Systems (LPSs) with other tracking technologies. After applying different exclusion criteria, they reviewed seven studies investigating the accuracy of LPSs in team sports. One was conducted indoors to test the validity for ice hockey tracking [21]. The other six were conducted outdoors, and the athletes performed several exercises to mimic soccer [14, 16–20]. The studies compared the LPSs to GPS tracking [16, 19, 20], camera-based tracking [18, 21], or GPS and camera-tracking [14, 17].

Similar to the studies included in the review, Roell et al. (2019), Wundersitz et al. (2015), and Heuvelmans et al. (2022) also validated wearable sensors for team sport-specific movements. Roell et al. (2019) compared IMU measurements of the acceleration with a motion capture system. Wundersitz et al. (2015) also used a motion capture system as a reference and validated the measurements of a trunk-mounted accelerometer. The athletes performed team sport-specific movements, and the researchers compared the respective peak impacts. Heuvelmans et al. (2022) compared a system of nine IMUs measuring the joint kinematics of the lower extremity to a motion capture system.

Team sport-specific movements and movements like single-leg squats were used by Dahl et al. (2020) to validate an IMU system. The system comprised eight IMUs attached to the body and was compared to a motion capture system. They investigated the accuracy of measuring the joint angles of the lower extremity. Luteberget et al. (2018) validated a commercially available LPS, based on IMMU, Ultra-Wideband (UWB), and Bluetooth in an indoor setting. They used a motion capture system to compare athletes’ positions, distance, and speed during team sport-specific exercises. Benson et al. (2020) validated a commercially available jump counter to estimate jump load in youth basketball players. The system consisted of an IMMU attached to the belly of the athletes. In rugby, Zanetti et al. (2014) validated an armband measuring the energy expenditure, and Carey et al. (2021) validated the head impacts measured by a wearable. With a focus on lacrosse, Buice et al. (2018) validated a sensor to measure head accelerations during sports. They compared linear accelerations and angular velocities with a reference system and found that the impact location heavily affects the system’s accuracy.
Apart from team sports, further studies investigated the accuracy of wearables in walking or running \([1, 24–26]\). Navalta et al. (2020) validated the heart rate measurements of smart sports bras. Compared to a heart rate monitor, the commercial products were worn during rest, walking, and running on a treadmill. The accuracy of estimating the energy expenditure and maximum oxygen uptake by commercially available wrist-worn devices was investigated by Passler et al. (2019). They compared smartwatches and fitness trackers to a gold standard during a lab and a field session. Most heart rate validation studies are performed in the lab on a treadmill. Therefore, Navalta et al. (2020) investigated the performance in the wild by conducting a study during trail running with varied intensities. They tested a smartwatch, a smart earbud, a smartwatch with the corresponding chest strap, a smart armband, and a smart ring. An even more realistic setting was chosen by Pobiruchin et al. (2017), who conducted surveys before and after a half-marathon regarding the type of wearables used and the distance tracked by the devices. Additionally, they investigated the adoption of wearables of different ages, sex, and fitness level groups and compared the accuracy of the devices in a real-world setting.

In addition to these studies, heart rate measurements of smartwatches or fitness trackers from various companies have been validated in many additional studies extensively \([135–142]\). Cosoli et al. (2020) reviewed validation studies for wrist-worn and chest strap wearables. They concluded that most validation studies use different test protocols, making it hard to compare different results. However, with the enormous and still growing number of smartwatches on the market, comparability for applications in research or sports is essential.

All previously presented validation studies are mainly in team sports or running. One study closer to ski jumping, in terms of speeds and accelerations, was performed by Gilgien et al. (2014). They investigated the accuracy of Global Navigation Satellite Systems (GNSSs) in alpine skiing. Therefore, they compared different combinations of differential and non-differential solutions, satellite systems, and signal frequencies. The GNSS antenna was mounted on the athlete’s helmet, and the receiver was carried in a backpack. They found that only one combination could achieve sub-decimetre position accuracy consistently.

A similar setup was used by Elfmark et al. (2021) for ski jumping. Using DGNSS and a camera-based system within their study, they validated the DGNSS measurements with the camera system. Firstly, they compared the two systems’ agreement in the overlap near the take-off table. They used only one camera and, therefore, only could compare the position, velocity, and acceleration in 2D. Thereby, they found a Mean Absolute Error (MAE) between both systems between 0.02 m and 0.10 m for different 5 m segments. The authors stated that this shows the validity and usability of these methods in research and practice. However, since the DGNSS consists of a wired antenna mounted on the athlete’s helmet and a receiver in a backpack carried under the ski jumping suit, the DGNSS is obtrusive and not suited for use in practice or even competitions. Other disadvantages are the system’s lack of real-time position determination and data transmission.

To conclude, commercially available wearables have been validated in many studies. This includes wearables based on GPS, UWB, IMU, IMMU, and DGNSS. They are mainly used to measure athletes’ positions, joint angles, or impacts. Also, heart rate measurements of
wearables, especially smartwatches, have been tested extensively. They all share that they focus mainly on team sport-specific movements or running.

However, the circumstances in ski jumping differ from these other sports. Especially regarding the athletes’ speed, which is maximum around 35 km/h in soccer and greater than 90 km/h in ski jumping. Additionally, the tracking system must cover a greater distance in ski jumping than in team sports.

The only validation study for a tracking system in ski jumping lacks real-time positioning, wireless data transmission, and unobtrusiveness, which makes it not applicable in practice. To my knowledge, no validation study for commercially available wearable sensors for ski jumping exists. Additionally, no study introduces a tracking system with the same capabilities as the one validated in my contribution [P1]. The Wearable Real-Time Tracking System (WRTTS) for ski jumping validated and used for performance prediction within this dissertation is based on the three previously presented publications by Groh et al. [108–110] and used a combination of IMU and UWB.

3.2. Deep Learning for Wearables in Sports

Since wearables are increasingly used in sports analytics, the amount of data gathered is constantly and rapidly growing. This creates new opportunities for analyzing data in sports using DL.

Before the rise of DL in many application fields [9–11], mainly traditional machine learning algorithms were used to analyze sports data. This includes Support Vector Machine (SVM) [145], k-Nearest-Neighbor (kNN) [146], decision trees [147], and random forests [148].

3.2.1. Human Activity Recognition

As mentioned, before the rise of DL in many application fields [9–11], mainly traditional machine learning methods were used for HAR in sports. Among others, studies were conducted in soccer [59], rugby [60], basketball [61], hockey [62], table tennis [63], skateboarding [64], and snowboarding [65, 66].

There are many studies comparing the performance of DL methods and traditional machine learning. Kautz et al. (2017) were among the first to apply DL to HAR in sports and showed that DL outperforms traditional machine learning algorithms. For this, they equipped beach volleyball athletes with an IMMU at the wrist of the dominant hand. The data were recorded during beach volleyball training. Afterward, a deep Convolutional Neural Network (CNN) was trained to distinguish between ten volleyball-specific actions (e.g., block, underhand serve, jump serve). They compared the performance of the CNN with a SVM, kNN, Gaussian Naïve Bayes (NB), decision tree, random forest, and a VOTE classifier based on the previous classifiers. The CNN reached the highest accuracy of 83.2% and outperformed the others by at least 16.0%. Also comparing DL and traditional machine learning, Rassem et al. (2017) investigated the performance of two CNNs, three Long Short-Term Memorys (LSTMs), and two Multilayer Perceptrons (MLPs) for classifying cross-country gears based on a 3D accelerometer. The standard forward LSTM had the smallest classification error, and all DL models outperformed
3. State of the Art

the MLPs. Similarly, Brock et al. (2017) automatically classified ski jumping errors using nine IMUs attached to the athlete’s arms, legs, skis, and pelvis. They compared the performance of a SVM, HMM, and CNN and found that the CNN outperformed the feature-based methods.

Another study by Anand et al. (2017) used a smartwatch’s acceleration and gyroscope data to classify shots in different swing sports. This included tennis, badminton, and squash. The performances for the different sports slightly differed, with badminton having the worst performance. Again the CNN and Bi-directional Long Short-Term Memory (BLSTM) outperformed the feature-based approach. Also covering a swing sport, Tabrizi et al. (2020) investigated forehand stroke classification in table tennis. In contrast to the previous paper, they did not use a wrist-worn sensor but mounted the sensor to the center of the racket. The sensor measured Euler angles and consisted of an accelerometer, gyroscope, and magnetometer. They compared the performance of a CNN, LSTM, and SVM, classifying the strokes into topspin, push, and basic shots. The LSTM achieved the best performance, followed by the CNN. Jiao et al. (2018) investigated the performance of different CNNs and a SVM in the classification of golf swings. They trained a standard CNN, a VGG, an Inception, and Residual Neural Network (ResNet) architecture on two strain gauge sensors, an accelerometer, and a gyroscope mounted to the golf club. All CNN architectures outperformed the SVM, representing the traditional machine learning methods.

Since many studies rely only on lab settings, Stoeve et al. (2021) compared the performance of a SVM with different DL architectures on HAR in soccer in laboratory and real-world scenarios. The investigated DL architectures were a CNN, a LSTM, and a convolutional LSTM. They found that the SVM could not reliably classify the activities in the investigated settings. The CNN outperformed all other architectures, demonstrating that CNNs can generalize and perform well in real-world settings. Also applying HAR in soccer, Cuperman et al. (2022) used five IMUs mounted on the athlete’s legs and pelvis to classify jogging, sprinting, passing, shooting, and jumping. For this, they investigated the performance of different CNNs, in combination with LSTMs and BLSTMs. Additionally, they investigated different weight-sharing methods for the inputs of different IMUs.

Apart from the sports context, DL was used for HAR in many different domains. For example, for activities in healthcare [149, 150], smart homes [151, 152], and autonomous driving [153].

Concluding, DL algorithms typically outperform other methods, including traditional machine learning, for HAR in sports. However, one drawback of DL algorithms is that they rely on lots of data. This is cumbersome to acquire, especially for supervised learning, where data needs to be labeled manually. Therefore, various studies have been investigating techniques to overcome this problem. One possibility is transfer learning. It has been applied to DL in multiple domains such as image classification [154–157] and image segmentation [158, 159].

Cook et al. (2013) reviewed studies covering transfer learning for activity recognition. They grouped the current literature into instance transfer, feature-representation transfer, parameter transfer, and relational-knowledge transfer as described by Pan and Yang (2010). However, Cook et al. (2013) reviewed the literature in a stage where DL has not yet gained much attention, so it did not cover HAR using DL.
One of the first DL approaches to transfer low-level features of a CNN to HAR in another domain using wearables was published by Morales and Roggen (2016). They used two different datasets of IMU-recorded activities, one with activities of daily living and one in an industrial setting. They investigated the transfer between users, domains, sensor modalities, and sensor locations. Thereby, they found that the choice of the source dataset is essential, and the convolutional features should be trained on complex datasets since, otherwise, transfer learning might decrease the performance.

Rokni et al. (2018) tackled the problem of dropping performance when applying an HAR model to unseen data from new users. Their approach was to train a CNN on many users as a source domain. Then they reused the lower layers and only re-trained the upper layer on very few labeled activities of a new user to personalize the HAR model. They evaluated this approach to activities of sports and daily living. They showed that this personalized CNN outperformed the traditional machine learning approaches but did not show the performance of the not-personalized CNN.

In a real-world application, however, it is not feasible to ask the users of a wearable to record and label their own dataset. Soleimani and Nazerfard (2021) worked on this by a semi-supervised approach with a labeled dataset in the source domain and an unlabeled dataset in the target domain. They then used a generative adversarial network to generate new data in the target domain. The generated data was then used to train the classifier in the target domain.

As stated, transfer learning might lead to negative transfer, which is dropping performance. Chen et al. (2019) investigated the person-specific discrepancy and the task-specific consistency. The person-specific discrepancy is the distribution shift in data between persons, and the task-specific consistency is the similarity in the same task. They proposed a new semi-supervised method consisting of multiple losses to reduce the person-specific discrepancy, preserve the task-specific consistency, and minimize the prediction error. They evaluated their approach on an HAR task, muscular movement recognition, and intention recognition.

Chen et al. (2020) proposed a multi-task model by extending the HAR task with a user recognition task. Both models share parameters, and an attention mechanism is used to spotlight important features of the other module. Through the proposed method, the performance of the HAR task should be improved by user-specific features and the performance of the user recognition task by activity-specific features.

To conclude, many studies have investigated HAR based on DL in various sports. The DL approaches typically outperform the traditional machine learning approaches. Even though many sports are covered, some disciplines, such as ultimate frisbee, have not yet been covered with feature-based or DL approaches. Additionally, transfer learning has been applied to HAR using DL but, to my knowledge, not to wearable-based HAR in sports. Especially less popular sports have not been covered yet, where transfer learning might especially be interesting since less data is needed.

3.2.2. Performance Prediction

Performance prediction in sports covers various topics such as predicting match results, performance prediction of individual athletes, or predicting the future trajectory within a movement.
3. State of the Art

One very active field of performance prediction is the prediction of match results. This can be done previous to matches but also in real-time during the match. Various studies investigated the prediction of soccer match results using logistic regression [167–169], kNN [167, 170–172], SVM [27, 167, 168], NB [27, 168], and Random Forest [27, 167, 168]. Also different DL approaches have been applied, for example, Fully Connected Neural Networks (FCNNs) [27, 38], Recurrent Neural Networks (RNNs) [40], and LSTMs [40]. Additionally, to be able to compare the match prediction accuracy, an open database was developed [173].

Also, for other sports, match result prediction studies have been performed. This includes, among others, American football [41–45], basketball [28–33, 46], rugby [34, 35], and ice hockey [36, 37, 39]. Some studies included Twitter data to improve the prediction accuracy in addition to the match and player statistics [41, 168, 174].

However, all these studies investigate the prediction of match results and are not based on wearables. The contribution in this dissertation [P2] is a real-time performance prediction in ski jumping. In the case of ski jumping, the performance prediction corresponds to the prediction of the ski jump length, which can also be interpreted as a trajectory prediction.

There are several studies covering ball trajectory prediction. Huang et al. (2011) developed an analytical rebound model between a table tennis ball and a table to predict the trajectory and spin of the ball. Sato et al. (2019) developed a system to predict and visualize the ball's landing point in volleyball. They used 22 motion capture cameras to track the volleyball, equipped with retroreflective markers. The detected 3D positions are then used to predict the landing point. The landing point and the current ball position were projected to the ground with 12 projectors. They found that the system improved beginners’ landing point prediction, enhancing their enjoyment.

Knibbe et al. (2015) used a depth camera to track and predict ball positions in real-time during juggling by a Kalman Filter. However, their application field was quite different compared to my contribution [P2] since it should be used to project images on the balls while juggling for an interactive mixed reality. Pan and Niemeyer (2017) also used a Kalman filter to predict the ball trajectory. With a motion capture system, they tracked the ball, head, and hands. The ball trajectory was then predicted and visualized in Virtual Reality (VR). The user could then catch the real ball using the visualization in VR.

Other studies covered predicting an athlete’s action within a match based on previous features. Hoang et al. (2015) predicted the classification of baseball pitches into fastball or nonfastball using machine learning. They used features recorded by Major League Baseball and predicted the class by linear discriminant analysis. Similarly, Bransen and Davis (2021) proposed a real-time prediction model for the penalty direction in soccer. The system was based on the penalty taker’s in-match performance indicators. They found, among others, that players who performed well preferably aimed for the left as a right-footed player and vice versa. They also found differences between men and women choosing a side based on their in-match performances.

Various studies showed that humans can predict human movements or ball trajectories in sports within the movement [177–180]. Shim et al. (2005) showed that skilled and novice tennis players could predict the course of a tennis serve by observing the movement of the
opponent player before the ball is hit. They also showed that the skilled players outperformed the novice players in anticipating the serve direction. Jackson and Mogan (2007) further investigated the anticipation skill by showing partly occluded tennis serve videos. By analyzing the prediction accuracies of different conditions, they examined which body parts contain the most information about the serve direction. Additionally, Tenenbaum et al. (2000) investigated the anticipation of tennis serve direction in relation to the duration of the subjects’ observation of the serve motion. They found that the longer the serve could be observed, the better the prediction of the serve direction.

The fact that humans can predict, for example, the trajectory of a serve in tennis by watching the opponent’s body shows that there are some features in the movement that contain information about the future movement and action. Researchers try to automatically extract these features to predict the ball trajectories or athletes’ movements like human capability.

To achieve this, researchers proposed a real-time mixed-reality martial arts training system [181, 182]. It consisted of a VR headset and an RGB camera mounted on top of the VR headset. It was built for an athlete with its coach. The athlete wearing the VR headset saw a 3D model of the coach together with the predicted pose 0.5 s in the future. The prediction pipeline consisted of multiple neural networks. Firstly, a ResNet was used to estimate the pose of the coach in 2D, then an LSTM predicted the pose of the coach in 2D. Finally, the 2D prediction was recovered to a 3D model by a residual network. Similarly, Itoh et al. (2016) proposed an augmented reality system to predict and visualize the future trajectory of a ball in real-time. The system used a see-through head-mounted display for visualization and two infrared cameras for tracking the ball.

Wu and Koike (2020) proposed a system to predict a serve’s table tennis landing point in real-time before the ball is hit. For this, they used a single camera and a ResNet to estimate the pose of the athlete. A series of these poses was then fed into an LSTM to predict the landing point on the table tennis table. The landing point was then visualized with a projector. Similarly, Suda et al. (2019) proposed a system to predict the ball trajectory based on the movement of the setter player before the toss. They used the data from two Kinect depth cameras to track the 3D positions of joints and the ball. This data was input into a FCNN to predict the ball position 0.3 s in the future. Shimizu et al. (2019) proposed a shot direction prediction in tennis based on a player’s pose and position. They used OpenPose [183] to extract the positions of the joints and input these together with the player positions into an LSTM.

Similar to the previously presented studies, the Omron Forpheus [184, 185] is a table tennis robot able to play against professional players. It uses multiple high-speed cameras and motion sensors to track the ball, racket, and players. When playing doubles, they claim to be able to estimate the empathy and cooperation between team members by measuring their motions and emotions.

Also, apart from the sport context, many studies have investigated the prediction of human poses using DL [186–190].

To conclude, many studies have been conducted on performance prediction in sports. However, the publications mainly deal with predicting match outcomes or ball trajectory prediction. To my knowledge, no previous study covers the prediction of the ski jump length, particularly
3. State of the Art

not a real-time prediction. However, multiple retrospective studies investigated the correlation of different parameters with the athletes’ performance, that is, the ski jump length [35–38].

In addition, the existing literature on predicting human pose or ball trajectory is mainly based on camera data rather than wearable sensor data. My contribution [P2] differs from current SOTA in these respects.
4. Discussion of Contributions

In this dissertation, I contribute to validating a wearable tracking system and present two Deep Learning (DL) approaches to process data of wearable tracking systems.

The papers in the appendix include a detailed discussion of the results of the presented works. However, I will briefly discuss and place the publications in a larger scientific context in this chapter. Additionally, I will present their respective impact and future perspectives.

4.1. Experimental Validation of Real-Time Ski Jumping Tracking System Based on Wearable Sensors

In the publication “Experimental Validation of Real-Time Ski Jumping Tracking System Based on Wearable Sensors” [P1] presented in Appendix A, my co-authors and I systematically investigated the accuracy of a Wearable Real-Time Tracking System (WRRTS) by comparing it with different reference systems.

4.1.1. Abstract of Original Publication

For sports scientists and coaches, it’s crucial to have reliable tracking systems to improve athletes. Therefore, this study aimed to examine the validity of a wearable real-time tracking system (WRRTS) for the quantification of ski jumping. The tracking system consists of wearable trackers attached to the ski bindings of the athletes and fixed antennas next to the jumping hill. To determine the accuracy and precision of the WRRTS, four athletes of the German A or B National Team performed 35 measured ski jumps. The WRRTS was used to measure the 3D positions and ski angles during the jump. The measurements are compared with camera measurements for the in-flight parameters and the official video distance for the jumping distance to assess their accuracy. We statistically evaluated the different methods using Bland–Altman plots. We thereby find a mean absolute error of 0.46m for the jumping distance, 0.12m for the in-flight positions, and 0.8° and 3.4° for the camera projected pitch and V-style opening angle, respectively. We show the validity of the presented WRRTS to measure the investigated parameters. Thus, the system can be used as a tracking system during training and competitions for coaches and sports scientists. The real-time feature of the tracking system enables usage during live TV broadcasting.

4.1.2. Author Contributions

J.L. and S.G. conceptualized the study, developed the methodology, wrote and tested the software, performed validation and formal analysis, and conducted the investigation and data curation. J.L. wrote the original draft and created visualizations, while J.L., S.G., and B.M.E. contributed to the writing, reviewing, and editing of the manuscript. B.M.E. supervised the project and was responsible for project administration and funding acquisition.
4. Discussion of Contributions

4.1.3. Discussion

The goal of this publication was to scientifically validate a WR TTS to investigate if it achieves sufficient accuracy for the application in training, competitions, and retrospective analysis. This aim was successfully achieved by analyzing the accuracy of the skis’ 3D positions, the skis’ 3D orientations, and ski jump length compared to camera measurements, total stations, and official video-based ski jump length determination, respectively. A company developed the tracking system based on previous studies [108–110].

Regarding the current literature, the contributed paper validates the only commercially available non-intrusive tracking system for ski jumping. Many Ultra-Wideband (UWB)-based wearables were validated in different study settings, as discussed in Section 3.1.2. However, due to the high speeds and large volume to cover, ski jumping is very different from team sports, where most of the validation studies were examined. Therefore, this publication contributes to the validation of wearables in sports, which goes beyond the current SOTA.

One limitation of the presented study is that it has been conducted on only one ski jumping hill. However, since the athlete is flying in an open volume without obstacles scattering or absorbing the electromagnetic waves of UWB, the accuracy should be similar on other ski jumping hills. Additionally, the construction standards for ski jumping hills are fixed [191], so there are no significant differences in the geometry of different ski jumping hills and, therefore, should also not affect the accuracy of the WR TTS.

The WR TTS has been in use in competitions already before this validation study. After this systematic validation, the system is used in all major ski jumping competitions. All athletes are forced to use the wearable in the official training jumps and competitions for some events [192]. The athlete’s team receives the data, which is analyzed by the coaches and sports scientists. Furthermore, the skiing federations use the data for different analyses, for example, to investigate the security of ski jumping. Additionally, the system is used for real-time application during TV broadcasting. The speed at take-off, the speed at 30 m after take-off, the respective speed change, the speed at the landing, and the skis’ opening angle at 30 m after take-off were visualized during TV broadcasting.

In the future, the system will be applied further during ski jumping competitions. This will result in an even larger dataset than the one in [P2]. This will make many more investigations possible, for example, analyzing the effect of different features of the ski jump on the ski jump length to improve the understanding of ski jumping and help the athletes to perform better as in [55–58]. Similar to [116], other studies could develop generic definitions for flight phases for better comparability between studies. Also, different methods of big data analysis and DL could be applied to such large datasets, for example, to identify athlete-individual patterns in ski jumping by classifying the ski jumps. These are only some future ideas, and many more investigations could be done based on the WR TTS.

More generally, properly systematically validating commercially available tracking systems is crucial to sports analytics. This is often neglected, and sports scientists and coaches use measured parameters with too little skepticism. In the future, more studies should systematically validate tracking systems for sports applications. This is essential to gain new knowledge through the data of tracking systems.
4.2. xLength: Predicting Expected Ski Jump Length Shortly after Take-off using Deep Learning

In the publication “xLength: Predicting Expected Ski Jump Length shortly after Take-off using Deep Learning” [P2] presented in Appendix B, my co-authors and I proposed a novel, innovative performance prediction in ski jumping.

4.2.1. Abstract of Original Publication

With tracking systems becoming more widespread in sports research and regular training and competitions, more data are available for sports analytics and performance prediction. We analyzed 2523 ski jumps from 205 athletes on five venues. For every jump, the dataset includes the 3D trajectory, 3D velocity, skis' orientation, and metadata such as wind, starting gate, and ski jumping hill data. Using this dataset, we aimed to predict the expected jump length (xLength) inspired by the expected goals metric in soccer (xG). We evaluate the performance of a fully connected neural network, a convolutional neural network (CNN), a long short-term memory (LSTM), and a ResNet architecture to estimate the xLength. For the prediction of the jump length one second after take-off, we achieve a mean absolute error (MAE) of 5.3 m for the generalization to new athletes and an MAE of 5.9 m for the generalization to new ski jumping hills using ResNet architectures. Additionally, we investigated the influence of the input time after the take-off on the predictions’ accuracy. As expected, the MAE becomes smaller with longer inputs. Due to the real-time transmission of the sensor’s data, xLength can be updated during the flight phase and used in live TV broadcasting. xLength could also be used as an analysis tool for experts to quantify the quality of the take-off and flight phases.

4.2.2. Author Contributions

J.L. and T.K. conceptualized the study, and J.L., L.S., F.P., and T.K. developed the methodology. J.L. was responsible for the software development, and J.L., L.S., F.P., and T.K. performed validation. Formal analysis was conducted by J.L., while J.L. conducted the investigation and performed data curation. B.M.E. provided resources for the study, and J.L., L.S., and F.P. prepared the original draft of the manuscript. J.L., L.S., F.P., T.K., and B.M.E. contributed to the writing, reviewing, and editing of the manuscript, and J.L., L.S., and F.P. created visualizations. B.M.E. supervised the project and was responsible for project administration and funding acquisition.

4.2.3. Discussion

The presented contribution aimed to develop an easily interpretable performance prediction in ski jumping, and we achieved this goal by developing xLength. In previous competitions, different parameters were visualized during TV broadcasting, using the WRTTS. This included the take-off speed, speed at 30 m after take-off, change of speed, the speed at the landing, and the opening angle at 30 m after take-off. Even though the intuitive interpretation of the take-off speed is correct, that is, the faster the better [57], things like speed 30m after take-off or at landing are already hard to interpret for non-experts. This is especially true for the opening
4. Discussion of Contributions

angle, where no clear, intuitive interpretation exists. With the developed xLength, we wanted to enhance the spectator experience of ski jumping. Especially in combination with the distance to beat, which is calculated in real-time and shows the desired length to achieve first place, ski jumping could get even more exciting when both are in close proximity.

Another possible use case for xLength is the retrospective analysis to quantify the take-off quality objectively and compare the take-off between individual jumps and athletes.

As presented in Section 3.2.2, various studies exist predicting the trajectory of balls or predicting athletes’ movement based on pose estimation. However, to my knowledge, no study exists predicting the ski jump length. Additionally, the existing literature mainly used camera data rather than wearable sensor data. Therefore, the presented contribution extends the current SOTA of sports analytics.

Additionally, the presented publication includes the largest dataset acquired for a study using a tracking system in ski jumping, where data acquisition is costly. This is mainly due to the small number of athletes, the small number of ski jumping venues, and the high risks.

During the deployment preparation for xLength, we noticed that we used the wind compensation factor and tangential wind measurement used for the final score of the athletes for calculating xLength. However, this factor is only finally determined after the athlete’s landing. However, the wind compensation factor and the tangential wind are computed continuously before and during the athlete’s jump and can be accessed within the operating system. Therefore, this would likely not affect the prediction process only if there is very gusty wind. However, the final wind compensation factor also does not correctly represent gusty wind. The reason for this is that the wind compensation factor is calculated as the mean of the tangential wind measurements at seven locations across the landing hill, sampled with 4 Hz, and averaged over five seconds.

This also leads to one future improvement of xLength. Instead of using only the averaged tangential wind and wind compensation factor, the raw wind measurements of all sensor locations could be used as an input for the Deep Neural Network (DNN). This would describe the wind conditions on the ski jump hill much more accurately and cover gusty winds. Using the wind information as a time series data would probably improve the prediction accuracy of xLength.

xLength was presented to the official timing and data provider of Fédération Internationale de Ski (International Skiing Federation) (FIS) and FIS itself. After minor adjustments due to the structure of hardware systems during live competitions, xLength is in preparation for deployment. After the successful deployment, the system should be tested in competitions in real-time. Afterward, the official timing and data provider and FIS will discuss possible future use cases.

Apart from the future application and analysis of xLength, the dataset it is built on can be used for various new research ideas. Data is costly to acquire in ski jumping. Therefore, the large dataset could be used to investigate different ski jump styles or to understand the complex ski jumping process in more detail. This may include extracting the key features for long jumps using explainable Artificial Intelligence (AI) methods.
4.3. Wearable Sensors for Activity Recognition in Ultimate Frisbee Using Convolutional Neural Networks and Transfer Learning

As also stated in the paper, combining the WRTTS with a camera at the take-off might be beneficial for a more detailed analysis of the take-off, which is crucial in ski jumping, to improve the prediction accuracy. By applying a DNN for pose estimation, a sensor fusion approach similar to [115] could be used. Through the pose estimation, the joint angles of, for example, the knees and hips could be analyzed, which are essential to determine whether an athlete’s take-off is too early or too late. With just one additional camera, the additional effort is small, and the camera can be placed near the inrun, so near the athlete, making bad vision during bad weather no problem.

Additional future work could be extending xLength to ski flying. While being a challenging task, the extension could be used to analyze the differences between ski jumping and ski flying. Currently, the most giant ski jump hill included in the dataset has a hillsize of 140 m. With ski jumping hills starting at 185 m up to 240 m, additional data needs to be acquired, which is particularly cumbersome since ski flying competitions are rare and also training is done way less than on ski jumping hills.

More generally, developing new innovative performance metrics and performance prediction based on DL will be a growing field of research. Especially in the performance prediction of individual athletes, many open research questions need to be addressed in future work.

4.3. Wearable Sensors for Activity Recognition in Ultimate Frisbee Using Convolutional Neural Networks and Transfer Learning

In the contribution “Wearable Sensors for Activity Recognition in Ultimate Frisbee Using Convolutional Neural Networks and Transfer Learning” [P3] presented in Appendix C my co-authors and I proposed a Human Activity Recognition (HAR) system for ultimate frisbee.

4.3.1. Abstract of Original Publication

In human activity recognition (HAR), activities are automatically recognized and classified from a continuous stream of input sensor data. Although the scientific community has developed multiple approaches for various sports in recent years, marginal sports are rarely considered. These approaches cannot directly be applied to marginal sports, where available data are sparse and costly to acquire. Thus, we recorded and annotated inertial measurement unit (IMU) data containing different types of Ultimate Frisbee throws to investigate whether Convolutional Neural Networks (CNNs) and transfer learning can solve this. The relevant actions were automatically detected and were classified using a CNN. The proposed pipeline reaches an accuracy of 66.6%, distinguishing between nine different fine-grained classes. For the classification of the three basic throwing techniques, we achieve an accuracy of 89.9%. Furthermore, the results were compared to a transfer learning-based approach using a beach volleyball dataset as the source. Even if transfer learning could not improve the classification accuracy, the training time was significantly reduced. Finally, the effect of transfer learning on a reduced dataset, i.e., without data augmentations, is analyzed. While having the same number of training subjects, using the pre-trained weights improves the generalization capabilities of the network, i.e., increasing the accuracy and F1 score.
This shows that transfer learning can be beneficial, especially when dealing with small datasets, as in marginal sports, and, therefore, can improve the tracking of marginal sports.

4.3.2. Author Contributions

J.L., T.P., and M.S. conceptualized the study and developed the methodology. T.P. was responsible for software development, while J.L., T.P., and M.S. performed validation and formal analysis, as well as investigation. Data curation was conducted by J.L., T.P., and M.S., and the original draft of the manuscript was prepared by J.L., T.P., and M.S. All authors contributed to the writing, reviewing, and editing of the manuscript, and J.L., T.P., and M.S. created visualizations. B.M.E. provided resources for the study, supervised the project, and was responsible for project administration and funding acquisition.

4.3.3. Discussion

This publication aimed to develop the first HAR system for ultimate frisbee or frisbee throwing in general. This aim was successfully achieved using a wrist-worn Inertial Measurement Unit (IMU) and a DNN to distinguish seven throwing types, catches, and a null class. Additionally, we showed that transfer learning could help develop an HAR system with small-scale datasets, for example, in niche sports where large datasets are harder to acquire.

As shown in Section 3.1.1, to my knowledge, only one study existed for body-worn wearables in ultimate frisbee [125]. They detected cutting maneuvers and reconstructed lower-body kinematics. Apart from this publication, to my knowledge, no other HAR system for ultimate frisbee existed. Therefore, the presented contribution closes this gap in the literature and extends the SOTA.

As already stated in the paper, one limitation of the contribution is the sensor placement at the wrist of the dominant hand, which could have been more optimal for measuring the minor differences between the different fine-grained throwing techniques. This study design limitation could easily be fixed in future work by placing the sensor at the back of the hand instead. Many frisbee players wear gloves for higher friction so that the sensor could be mounted on the glove.

Although the placement of the sensor on the wrist is not ideal, it does have the benefit of being the same as that of smartwatches and fitness trackers. This means that incorporating our model into a smartwatch app for analyzing ultimate frisbee sports would be possible.

Another limitation was that no left-handed athlete was included in the study. This is because we acquired study participants from the local ultimate frisbee team. However, no left-handed athlete was a member of the team and willing to participate. Since ultimate frisbee is quite an unknown sport in Germany, no other left-handed ultimate frisbee athletes were available during the period of study execution. Therefore, future work should incorporate left-handed athletes and evaluate the network’s performance on left-handed athletes.

One limitation was the relatively controlled setting without an opposing player pressuring the athlete with the frisbee, as in a competitive match. This makes the throwing more controlled, and the throwing technique is probably easier to distinguish. Therefore, the system’s perfor-
formance should be investigated in a less controlled environment, for example, in a competitive match, in the future. Previous work has already investigated the performance of HAR systems in laboratory and real-world settings in soccer [5] and showed that DNNs are capable of generalizing to less controlled, real-world settings.

HAR in ultimate frisbee could be used to monitor the performance of athletes and to compare the execution of different throws between athletes and within one athlete. Additionally, it could be used to improve the throwing style and to determine player and match statistics automatically.

More generally speaking, future work could apply HAR to more sports so that less common sports can benefit from recent technical developments. Thereby, transfer learning could be used to train new DNNs with smaller datasets that need to be acquired. Further investigations should be done for feature transfer between different sports, sensor locations, and study settings, for example, from laboratory settings to competitive matches. This could help to quickly identify the best-suited dataset for transfer learning for a new HAR system.
5. Conclusion

This dissertation presented and discussed various contributions to the field of sports analytics. For the contributions, I combined two recent developments in sports analytics research. Firstly, the application of Deep Learning (DL) and, secondly, the use of wearables in sports. DL offers excellent potential to analyze large datasets and detect new patterns, while wearables can track athletes unobtrusively.

New, easy-to-interpret performance metrics are an emerging field in sports analytics, such as, for example, the Handball Performance Index, which was presented in the introduction. My contributions covered different steps in developing easy-to-interpret performance metrics for different sports, exemplarily shown for ski jumping and ultimate frisbee. I systematically validated a wearable tracking system for ski jumping in the field by comparing its measurements with multiple reference systems, used DL methods to detect and classify activities in ultimate frisbee based on wearables automatically, and developed a real-time performance prediction for ski jumping using DL and wearables.

5.1. Impact of this Thesis

This dissertation aimed to extend the current State of the Art (SOTA) in the field of sports analytics. My publication [P1] contributes to the field of validation of wearables in sports. With my publications [P2], and [P3], I contributed to the field of DL in sports analytics, which has just arisen in the last few years.

The Wearable Real-Time Tracking System (WRTTS), validated in [P1], is now used in all major ski jumping competitions, where athletes are required to wear it during official training and some events [192]. The collected data is used by coaches, sports scientists, and skiing federations for analysis and to investigate the safety of ski jumping. The system is also used in real-time during TV broadcasting to display athlete’s speed and ski opening angle to viewers.

The ski jump length prediction xLength, developed in [P2], was presented to Fédération Internationale de Ski (International Skiing Federation) (FIS) and its official timing and data provider. After minor adjustments, it is being prepared for deployment and real-time testing in competitions, with possible future use cases to be discussed with FIS and its official timing and data provider.

It should be emphasized that the contributions presented in this dissertation were published roughly one year or less prior to the dissertation’s composition. As a result, it is not currently feasible to conduct a comprehensive analysis of their impact.
5. Conclusion

5.2. Findings

I combined the application of wearables with SOTA DL methods resulting in new methods for sports analytics. The individual findings of this thesis can be summarized as follows:

- In [P1], I compared with my co-authors a WRTTS with camera measurements, total stations, and official video-based jump length measurements. We showed that the commercially available WRTTS validly estimated the skis’ 3D positions and skis’ 3D orientations during the flight and the ski jump length.

- In [P2], I developed with my co-authors a new, easily interpretable performance prediction in ski jumping. xLength is based on the measurements of the WRTTS and predicts the expected ski jump length directly after the take-off using a Deep Neural Network (DNN).

- In [P3], I showed with my co-authors that different frisbee throwing techniques and catches can be classified by a DNN using a wrist-worn Inertial Measurement Unit (IMU). Additionally, we demonstrated that transfer learning can be beneficial for classifying time-series data in sports. This might especially benefit other minor sports where no Human Activity Recognition (HAR) systems exist.

5.3. Outlook

Looking ahead, the findings of this study open up new avenues for exploration and present exciting opportunities for future research. Some of the limitations of my individual contributions could be overcome through follow-up studies. As discussed in Section 4.1.3, the WRTTS validated in [P1] will be used in all major ski jumping competitions, generating a larger dataset for further investigations. This includes analyzing ski jump features’ impact on the ski jump length, developing standardized flight phase definitions, and applying big data analysis and machine learning to identify athlete-specific patterns.

Based on the WRTTS, in [P2], we developed the Expected Ski Jump Length (xLength), where several follow-up investigations are very interesting. As discussed in Section 4.2.3, this includes the extension to ski flying, combining the wearable system with a camera for pose estimation at the take-off, or including more detailed time-series wind data. Additionally, of course, the deployment during competitions is a goal to work on. Besides the ski jump length prediction, additional sport science questions can be investigated by analyzing the acquired dataset.

Some limitations should be addressed for the activity recognition system in ultimate frisbee proposed in [P3]. As discussed in Section 4.3.3, the sensor placement at the back of the hand and a less controlled study setting should be investigated. In general, further investigations could be conducted regarding the application of transfer learning in sports analytics. This includes investigating the transfer between sensor positions, different sports and from lab to real-world settings.

Despite my contributions in this dissertation, the application of DL to wearables in sports or sports analytics, in general, remains a relatively new research area with lots of open topics. Future research might cover the development of performance prediction for multiple other
5.3. Outlook

sports or the application of explainable Artificial Intelligence (AI) methods to gain new insights into sports.

In general and from a wearable customer and user perspective, I believe an increasing use of wearable tracking systems during training and competitions can be expected. Additionally, more tracking systems and wearables will be developed and become commercially available. It is essential to validate them appropriately in real-world settings to give the athletes accurate and valuable feedback on their performance. Further, this will lead to more and more data available and make the application of DL an even more powerful tool to extract important learnings for the athletes from the data without overwhelming users with the raw data. Therefore, innovative and easily interpretable performance metrics need to be developed by sports scientists together with computer scientists. Then, elite athletes, recreational athletes, and spectators can benefit. Till now, these developments are mainly in the stage of research studies or only available for elite athletes. In the next few years, we will see the transformation of new DL methods in science into the everyday sporting life of athletes through commercially available products. Then not only do elite athletes supported by sports scientists benefit from the newest developments, but recreational athletes using smartwatches or other fitness trackers can significantly benefit.
Appendix

A. Experimental Validation of Real-Time Ski Jumping Tracking System Based on Wearable Sensors
Experimental Validation of Real-Time Ski Jumping Tracking System Based on Wearable Sensors

Johannes Link 1,*, Sébastien Guillaume 2 and Bjoern M. Eskofier 1

1 Machine Learning and Data Analytics Lab, Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91052 Erlangen, Germany; bjoeurn.eskofier@fau.de
2 Institute of Territorial Engineering, Haute Ecole d’Ingénierie et de Gestion du Canton de Vaud, 1400 Yverdon-les-Bains, Switzerland; sebastien.guillaume@heig-vd.ch

* Correspondence: johannes.link@fau.de

Abstract: For sports scientists and coaches, it is crucial to have reliable tracking systems to improve athletes. Therefore, this study aimed to examine the validity of a wearable real-time tracking system (WRRTS) for the quantification of ski jumping. The tracking system consists of wearable trackers attached to the ski bindings of the athletes and fixed antennas next to the jumping hill. To determine the accuracy and precision of the WRRTS, four athletes of the German A or B National Team performed 35 measured ski jumps. The WRRTS was used to measure the 3D positions and ski angles during the jump. The measurements are compared with camera measurements for the in-flight parameters and the official video distance for the jumping distance to assess their accuracy. We statistically evaluated the different methods using Bland–Altman plots. We thereby find a mean absolute error of 0.46 m for the jumping distance, 0.12 m for the in-flight positions, and 0.8°, and 3.4° for the camera projected pitch and V-style opening angle, respectively. We show the validity of the presented WRRTS to measure the investigated parameters. Thus, the system can be used as a tracking system during training and competitions for coaches and sports scientists. The real-time feature of the tracking system enables usage during live TV broadcasting.

Keywords: sports; wearable sensors; ultra-wideband; inertial measurement unit; ski jumping; validation study; tracking

1. Introduction

From the very beginning of ski jumping in the 19th century until now, the ski jumping technique has undergone various developments. This includes, amongst others, the Kongsberger Technique, in which the upper body is bent and the arms are extended over the head. Later, the arms were brought back next to the upper body. During all these changes, the skis remained parallel. The last major change was the introduction of the V-style in 1985, where the skis are not parallel but form a V [1,2]. Nowadays, this technique is used by almost all athletes. All these developments resulted in larger jumping distances. In the past, new jumping techniques were generally found by coincidence or trial and error. In contrast, today, researchers work hard to find the optimal flight technique [3–6] and study the biomechanics of ski jumping [7,8].

Therefore, sports scientists need reliable and automatic tracking systems. Various studies with different measurement systems have been conducted to investigate, amongst other things, the jumping distance [9], landing momentum [10], ground reaction forces [11], and ski position [12].

These studies incorporate the use of inertial measurement units (IMUs) [13,14], force insoles or differential Global Navigation Satellite System (dGNSS) [15]. The latter has the disadvantage of being obtrusive since antennas on the helmet and a backpack have to be attached to the skiers. Additionally, camera-based tracking methods are
used [5,7,16–21], which are unobtrusive but have the disadvantage of only covering a part of the jumping hill, or many cameras are needed and must be combined. Furthermore, the post-processing of video data has high computational costs and does not work during bad weather conditions.

The evaluated system uses ultra-wideband (UMB) radio communication and ranging in combination with IMUs to track the position of the athletes continuously during the flight. UWB has been used successfully in sport tracking applications in various sports. This includes indoor sports, such as ice hockey [22], and handball [23] and outdoor sports, such as soccer [24].

In previous studies, ski jumping parameters were obtained offline. For a later application during training sessions or TV broadcasting, a real-time system is beneficial and therefore required. Thus, e.g., the speed during the jump can be shown live during the TV broadcasting, or jumping trajectories can be compared directly using a 3D visualization. For a live application, wireless transmission of the obtained data is required. However, for the athletes and coaches, this also has several advantages. This includes easier handling, and the athletes are not interrupted in their focus, which is crucial in competitions and training due to the high risk associated with ski jumping.

Within this work, we investigate the accuracy of a wearable tracking system measuring several ski jumping parameters. The tracking system is unobtrusive, which is essential to be used in competitions so as not to distract the athletes during the crucial flight phase. The parameters include the jumping distance, in-flight positions, and in-flight orientation of the skis. All parameters are determined and transmitted in real-time. The wearable real-time tracking system (WRRTS) brings the lab to the field. In contrast to previously proposed systems, the evaluated system is real-time capable. Another advantage of the investigated system over previous ones is the wireless data transmission, which results in no interaction with the athlete during training or competition and does not disturb their focus.

The main objective of this study is the systematic validation of a wearable real-time tracking system for ski jumping. Prior studies have only focused on single aspects of tracking ski jumps. The investigated system provides multiple measured parameters obtained and transmitted in real-time, making it very applicable for further research of the kinetics and kinematics of ski jumping. Therefore, the accuracy of the tracking system needs to be investigated so that sports scientists can derive valuable insights.

The remainder of the paper is structured as follows. Section 2 introduces the study procedure with the WRRTS as well as the reference measurement systems. Section 3 shows the results of the comparison of the WRRTS with its respective reference systems. In Section 4, the previously presented results of the validation study are discussed. Finally, in Section 5 the conclusions of this work are presented.

2. Materials and Methods

In this section, we introduce the procedure of the data acquisition and the different measuring methods with their respective setup at the ski jumping venue. Furthermore, the evaluation of the study data is presented.

2.1. Measurement Systems

The data of the jumps was acquired using the WRRTS, a camera system operated by ccc software GmbH, the total station tracking system QDaedalus [25,26], and the official video-based jumping distance measurement. All measurement systems are described in more detail in the following.

2.1.1. Wearable Real-Time Tracking System (WRRTS)

We use the tracking system that was developed based on the previous work of Groh et al. [9,10] in combination with UWB-based tracking capabilities.
The WRRTS consists of two main components. The first main component of the tracking system comprises antennas with fixed positions along the jumping hill. Figure 1 shows the positions of the antennas along the ski jumping facility as well as the positions of the reference measurement devices.

Figure 1. Definition of the local coordinate system with respect to the ski jumping hill. The origin of the coordinate system is the middle of the edge of the jump-off platform. With respect to the jumping direction, the axes are defined as: X-axis: horizontal, forward, Y-axis: horizontal, left, Z-axis: vertical, upward. Additionally, the setup at the ski jumping hill is presented. This includes the positions of the antennas of the WRRTS, the total stations of the QDaedalus system, the cameras, and the 3D models of the ski jumping in-run and landing hill.

The other main component is the mobile trackers that are attached to the skis of the athletes. In cooperation with the ski binding manufacturer Slatnar, the trackers were designed for easy attachment to the bindings to facilitate their use. An example of the attachment is shown in Figure 2.

Figure 2. Attachment of the wearable tracker on top of the binding of the skis. The tracker is mounted on top of the regular ski jumping binding in front of the foot (tracker in white).
The tracking system uses ultra-wideband radio technology and microelectromechanical inertial measurement units. This combination allows the continuous measurement of ski orientation, acceleration, velocity, and position of both skis during the complete jump in real-time. The internal update rate of the inertial measurement units is 1000 Hz, and that of the ultra-wideband radio technology as well as the transmitted output data is 20 Hz. The measurements are acquired in a local Cartesian coordinate system centered at the middle of the edge of the jump-off platform, as shown in Figure 1.

Furthermore, a 3D scan of the landing hill is acquired using a total station. A total station is a theodolite with integrated distance measurement, thus measuring the vertical and horizontal angles and distance to an aimed point. This is used to obtain a mapping from the local Cartesian coordinate system to measurements with respect to the landing hill (e.g., height over ground, jumping distance).

2.1.2. Official Distance Measurements

The jumping distance of the recorded jumps was determined with the official video-based measurement system that has been used in Fédération Internationale de Ski (FIS) competitions for over 25 years. The system operator was certified by FIS for video distance measurements. To determine the jumping distance, the system operator determines the first camera frame where both skis are flat on the landing hill. Through the camera calibration, the respective jumping distance is determined and rounded down to a resolution of 0.5 m. An example of the determination of the official jumping distance is shown in Figure 3.

![Figure 3. Example for the manual determination of the official video-based jumping distance. The red lines are the calibrated projection of the official jumping distance (to the jump-off platform). The blue line corresponds to the determined jumping distance.](image)

2.1.3. Camera Measurements (ccc Software GmbH)

In the field of sports software, ccc develops video analysis systems for training optimization and competition control and platforms for the storage of training and competition data [27].

ccc provided camera-based measurements of the ski jumps and was used as a reference for the validation study. To this end, multiple fixed cameras were installed next to the jump-off platform and next to the landing hill at 8 m, 18 m, 30 m, 45 m, and 60 m after the jump-off platform. Additionally, a pan–tilt–zoom (PTZ) camera was mounted at the upper end of the in-run. Due to a defective calibration, the deviation of the WRRTS in the Y-coordinate could not be investigated using the PTZ camera. The position of the PTZ camera was measured using the total station, enabling the projection of the V-opening-angle to be determined on the PTZ camera image plane. For the camera at 30 m, the calibration was faulty and therefore could not be used. Figure 1 shows the positions of all installed cameras.
For assessing the WRRTS measurements using the ccc videos, two different approaches were taken.

The first one was the projection of the 3D positions of the WRRTS into the 2D image plane using intrinsic and extrinsic camera calibration parameters. This approach does not provide a meaningful quantification of the measurement errors (only in pixels, not in meters). However, it provides an intuitive means for visually assessing the data quality since the WRRTS measurements can be seen directly in the image.

The second approach is the comparison of the video data with the WRRTS data in 3D. Since the jumper is only visible in one camera image at a time during a jump, full stereoscopic measurements are not possible. However, it is possible to associate every pixel coordinate in the camera image with a 3D vector that is formed by connecting the camera position (i.e., its optical center) with the corresponding point in the image plane. The deviation of the WRRTS positions from this 3D vector can then be determined. Based on this geometric model, comparisons of the ski orientation can also be performed. It should be noted that due to the two-dimensional nature of the camera images, only deviations in the image plane are captured by this method.

To compare the WRRTS position in 3D with the camera vectors, the trackers and skis are manually labeled in the videos. An example of the manual labeling of the skis and trackers for one of the cameras next to the landing hill is shown in Figure 4. Additionally, Figure 5 shows the manual labeling of the V-angle in the PTZ camera.

To compare the camera measurements, we determine the point of the interpolated WRRTS trajectory with the shortest distance to the camera vector. We use the instant of time of this point of the WRRTS trajectory for the position and angle comparison. This procedure is described in more detail in Section 3.

![Figure 4. Example for the manual labeling of the tracker positions and skis for the cameras next to the jumping hill.](image)

2.1.4. Total Station Tracking (QDaedalus System)

Since the camera measurements did not provide full 3D positions of the skis, the 3D tracking system QDaedalus was additionally used for validating the three-dimensional position. QDaedalus is a measurement system developed at the Geodesy and Geodynamics Lab at the ETH Zurich and consists of a combination of total stations and CCD cameras that allow the accurate triangulation of objects in 3D. For this study, we used Leica TCA 1205 total stations in the QDaedalus system. The positions of the QDaedalus stations are shown in Figure 1. The raw position data measured with the QDaedalus system were filtered and interpolated using a least-squares collocation. The trajectories measured with QDaedalus were registered to the same coordinate system as the WRRTS measurements (see Figure 1) and compared to the WRRTS data regarding 3D positions.
2.2. Venue

The data acquisition took place at the hill size 100 ski jumping hill in Oberhof (Germany). In contrast to the second day, foggy weather conditions impeded data acquisition on the first day of the acquisition. This affected some of the camera measurements during this day since the manual labeling requires a clear vision of the wearable tracker and skis. Additionally, the QDaedalus measurements were affected during this day since the measurement principles require inter-visibility between the total stations and the athletes.

In total, data of 35 jumps were collected over two days. Four athletes participated in the data acquisition (three male, one female). One of the athletes was part of the German A National Team and three of the German B National Team. Measurements of the jumps were acquired using the WRRTS and the reference measurement systems described below.

2.3. Evaluation

In this subsection, the synchronization of the WRRTS and the camera measurements, the procedure for the position, and angle comparison is described. Additionally, the statistical analysis is presented.

2.3.1. Synchronization

The WRRTS and the camera system used for evaluation do not have a common synchronized time. Therefore, the measurements were synchronized in retrospect. Different synchronization procedures were used for the PTZ camera, and the cameras at the edge of the jumping hill.

To synchronize the PTZ camera with the WRRTS measurements, we applied a time offset correction. The internal time of the WRRTS was set to zero when passing the edge of the jump-off platform (origin of the local coordinate system). Therefore, by manually labeling the frame in the camera image where the tracker passed the edge of the jump-off platform, we determined the offset of the time of the PTZ camera. Using this offset, we synchronized both measurement systems for each jump individually.

To synchronize the lateral cameras at the side of the jumping hill with the WRRTS a different approach had to be used. Due to their limited field of view, the edge of the jump-off platform is not visible and, therefore, cannot be used as a reference point for synchronization. Instead, we take an alternative approach. First, we project the manually annotated pixel coordinates of the tracker into 3D using $y = 0$. Then we linearly interpolate the measurements of the WRRTS. Then, the time offset that minimizes the distance between the interpolated WRRTS measurements and the line of sight between the camera position...
and the projected annotation is determined. This time offset is then used to align the WRRTS and camera measurements.

For the comparison of the X- and Z-coordinate with the camera measurements, we project the manual annotation in 3D using the Y-coordinate of the synchronized WRRTS measurement.

Since the WRRTS and QDaedalus measurements do not have a synchronized time base either, their temporal relation also had to be determined in retrospect. Therefore the trajectories of both systems are up-sampled via linear interpolation. The up-sampled trajectories are shifted in time to find the best temporal agreement. As the measure for agreement, the mean absolute error of the overlapping trajectories is used.

2.3.2. Statistical Analysis
The results of the WRRTS and the respective reference system are statistically analyzed using different error measures. These are briefly explained in this subsection.

Bland Altman introduced a graphical approach to compare two measuring methods [28]. It consists of a scatter plot with the difference between paired measures against their mean. The mean of the two methods is used since both methods have a measurement error, and the true value is not known. Therefore the mean of the paired measures is the best estimate of the true value. Plotting the difference against the mean enables us to see if there is a dependency of the measurement error on the value of the measurement.

Additionally, the mean difference \( \bar{d} \), as well as the upper and lower limit of agreement (LoA) is plotted as a horizontal line. The limits show the range of the difference in which 95% of the data are located. The limits of agreement are calculated as \( \bar{d} \pm 1.96 \cdot \sigma \), where \( \sigma \) is the standard deviation.

The assumption for the Bland–Altman plot is that the difference between the two methods is normally distributed. To test whether the assumption of a normal distribution is valid, we use a Kolmogorov–Smirnov test [29]. For a \( p \)-value \( \geq 0.05 \), we accept the null hypothesis of a normal distribution. For a \( p \)-value < 0.05, the differences to a normal distribution are significant, and we test for other distributions to calculate the limits of agreement.

Apart from the mean, we also investigate the standard error of the mean (SEM). It is defined as

\[
SEM = \frac{\sigma}{\sqrt{N}} \tag{1}
\]

where \( N \) is the number of compared measurements and \( \sigma \) the standard deviation. The SEM gives an estimate of how far the mean of the sampled data differs from the mean of the whole population.

The deviation is summarized in terms of the mean absolute error (MAE), which is calculated as

\[
MAE = \frac{\sum_{i=1}^{N} |x_{i,WRTS} - x_{i,REF}|}{N} \tag{2}
\]

where \( N \) is the number of compared measurements \( x_{i,WRTS} \) is the \( i \)-th measurement obtained with the WRRTS (e.g., jumping distance), and \( x_{i,REF} \) is the \( i \)-th measurement obtained with the corresponding reference measurement system. The MAE is an easy-to-interpret measure combining the bias and precision of a distribution. Thus, we use it as the figure of merit for characterizing the accuracy of the WRRTS.

3. Results
3.1. Official Video Distance
The jumping distance determined with the WRRTS was compared to the official video distance.

The differences between measurement methods are visualized in Figure 6 using a Bland–Altman plot. The plot contains the jumping distance of 35 ski jumps. The jumping distances range from 70 m to 106 m. The bias of the WRRTS is 0.31 m with an SEM of
±0.08 m. The precision stands at 0.44 m which leads to upper and lower limits of agreement of −0.56 m and 1.17 m, respectively. Two data points exceed the limits of agreement. This corresponds to 5.7% of the data. The MAE amounts to 0.46 m. The data are distributed equally around the bias, with no dependency on the jumping distance.

Figure 6. Bland–Altman plot for the comparison jumping distance determined with the WRRTS and the official video distance.

3.2. Camera Measurements (ccc Software GmbH)

3.2.1. Projection of 3D WRRTS Measurements into the 2D Image Plane

An example of the three-dimensional WRRTS position measurements projected into the image plane of a video frame can be seen in Figure 7. There, the cyan curve shows the projected trajectory of the tracker placed on the left ski of the athlete. The magenta curve is the projection of the right ski. Additionally, the coordinates for some of the measured samples are shown in the figure.

3.2.2. Comparison of 3D WRRTS Position Measurements with 3D Camera Vectors

In this subsection, the in-flight positions, measured with the WRRTS, are compared to the camera measurements. Due to the missing calibrations, for the ccc camera placed at 30 m after the jump-off platform, these recorded videos could not be evaluated. Furthermore, for the PTZ camera, no calibration was available, so the Y-coordinate of the 3D positions could not be inspected with the camera measurements. Assuming a relatively small camera lens distortion, we nevertheless investigate the V-angles recorded by the PTZ camera.

Firstly, we have a closer look at the X-coordinate of the in-flight three-dimensional position. In Figure 8, the Bland–Altman plot visualizes the results. The data range from $x = -1.6$ m to $x = 58.4$ m and consist of 741 paired measurements. The bias of the WRRTS is $-0.039$ m with an SEM of 0.002 m. The precision of the difference is 0.05 m. The resulting upper and lower limit of agreement are at $-0.14$ m, $0.07$ m, respectively. The MAE condenses this to 0.04 m.

The difference does not follow a normal distribution or another basic distribution. The Bland–Altman plot shows that the bias of the difference varies with the X-coordinate. Additionally, as expected, the precision is worsening with larger X-coordinates. In more detail, the results for every camera individually are shown in Table 1.
Figure 7. Projection of the WRRTS position data onto the video of the camera at 18 m after take-off. The cyan curve shows the trajectory of the WRRTS tracker placed on the binding of the left foot. The magenta curve shows the trajectory of the right tracker. The numbers indicate the corresponding 3D coordinates in meters. Coordinates as depicted in Figure 1.

Figure 8. Bland–Altman plot for the comparison of the X-coordinate determined with the WRRTS and camera vectors.

Table 1. Bias, SEM, and precision of the X- and Z-coordinate determined with the WRRTS and camera vectors for every camera individually.

<table>
<thead>
<tr>
<th>Coordinate</th>
<th>Metric</th>
<th>Camera Position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0 m</td>
</tr>
<tr>
<td>X</td>
<td>bias [m]</td>
<td>−0.001</td>
</tr>
<tr>
<td></td>
<td>SEM [m]</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>precision [m]</td>
<td>0.01</td>
</tr>
<tr>
<td>Z</td>
<td>bias [m]</td>
<td>−0.001</td>
</tr>
<tr>
<td></td>
<td>SEM [m]</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>precision [m]</td>
<td>0.02</td>
</tr>
</tbody>
</table>
The Bland–Altman plot that corresponds to the Z-coordinate is shown in Figure 9. The data range from \( z = -25.7 \) m to \( z = 0.3 \) m and contain 741 paired measurements. The bias of the WRRTS is \(-0.071\) m with an SEM of \(0.003\) m. The precision of the difference is \(0.08\) m, which leads to an upper and lower limit of agreement of \(0.09\) m and \(-0.23\) m, respectively. The MAE of the difference stands at \(0.08\) m.

![Bland–Altman plot](image)

**Figure 9.** Bland–Altman plot for the comparison of the Z-coordinate determined with the WRRTS and camera vectors.

As the difference of the X-coordinate, the difference of the Z-coordinate does not follow a normal distribution or any other basic distribution. The precision of the difference increases for smaller Z-values, which means farther away from the jump-off platform. Furthermore, the bias of the WRRTS increases for decreasing Z-values. The influence of the camera position on the bias and precision of the Z-coordinate is shown in Table 1.

### 3.2.3. Comparison of 3D WRRTS Position Measurements with 3D QDaedalus Measurements

In this subsection, the three-dimensional position during the jump determined with the WRRTS is compared with the results of the QDaedalus system. Due to the foggy weather on the first study day for this investigation, only jumps from the second study day are usable. This is because the jumpers could not be reliably seen and therefore not be tracked with the QDaedalus system.

This results in eight jumps investigated with a total of 618 data points. The data are visualized in Figure 10. Because we investigate three-dimensional data, a conventional Bland–Altman plot is not appropriate. Therefore, we plot the distance between the WRRTS position and the position of the QDaedalus system against the mean distance of both systems to the jump-off platform.

Since the distance between the points of both measurement systems is a projection of the distance vector, it is not normally distributed. We assume a beta distribution and test it with a Kolmogorov–Smirnov test. We therefore get a \(p\)-value of 0.38, which confirms our assumption. The right part of the plot shows a histogram of the projection of the difference between the two methods. Additionally, the fitted beta distribution is plotted.

The mean difference between both systems is \(0.290\) m with an SEM of \(\pm 0.007\) m. Since the distance is a positive function, the mean is also the MAE. The precision of the difference is \(0.18\) m. The limit of agreement is not calculated as \(1.96 \times \sigma\) but with the percentage point function of the beta distribution. We thereby get a limit of agreement of \(0.64\) m.
Figure 10. The left part shows a scatter plot for the comparison of the three-dimensional position measured with the WRRTS and QDaedalus tracking. The distance on the X-axis is the mean distance of WRRTS and QDaedalus to the jump-off platform. The histogram of the projection of the distance between both measurement methods is visualized in the right part. The fitted beta distribution is also plotted in the projection.

Looking at the scatter plot, the distance between the two methods tends to increase with the distance from the jump-off platform slightly.

3.2.4. Comparison of 3D WRRTS Angle Measurements with 3D Camera Vectors

Lastly, we inspected the ski orientation during the flight phase. The difference between the angles in the ccc videos from the side of the landing hill and the WRRTS angles projected to the image plane are shown in Figure 11. The X-axis is thereby the mean of the angle to the unit vector in X-direction in three dimensions projected into the image plane. The data for the angle comparison consist of 662 paired measurements. The WRRTS has a bias of 0.26°, and the SEM stands at 0.04°. The precision of the difference is 1.1°. The MAE is 0.8°. The difference does not follow a normal distribution but a Student’s t-distribution. The Kolmogorov–Smirnov test for a Student's t-distribution results in a p-value of 0.57. Using the respective percentage point function, the limits of agreement are determined to be $-1.9^\circ$ and $2.4^\circ$.

Besides the cameras next to the landing hill, one PTZ camera is mounted at the start of the in-run. The videos recorded with this camera are evaluated and compared to the angle determined by the WRRTS. Due to the missing calibration of the PTZ camera, we can only determine the angles in the image plane but no positional information. The projected V-angle is compared at $t = 0.5, 1.0, 2.0$ s after crossing the edge of the jump-off platform.
Figure 11. The left part shows the Bland–Altman plot for the comparison of the angle measured with the WRRTS and the angle determined in the camera images from the side along the jumping hill. The right part shows the projected histogram of the difference as well as the fitted Student’s t-distribution. The gray arrows and numbers at the top of the plot indicate three measured angles out of the plot’s range.

Figure 12 visualizes the Bland–Altman plot for the V-angle determination. The data for the projected V-angle range from 41° to 76° and consist of 70 measurements. For some of the jumps, due to the fog on the first day of the study, no camera measurements could be taken. The WRRTS has a bias of 3.38° and an SEM of 0.24°. The standard deviation of the differences is 1.96°. Therefore, the limits of agreement stand at −0.47° and 7.23°.

The MAE of the differences of the projected V-angle amounts to 3.4°.

Figure 12. Bland–Altman plot for the comparison of the projected V-angle measured with the WRRTS and the PTZ camera.
4. Discussion

4.1. Jumping Distance

The WRRTS determines the jumping distance in good agreement with the official video distance. Even though the accuracy of the WRRTS might not be sufficient for competitions, its usage is helpful for training since no manual operator is needed to determine the jumping distance individually. It is important to mention that the true jumping distance is unknown due to the rounding during the determination of the official video distance. This rounding down of the jumping distance of the official video distance introduces an MAE of 0.25 m compared to the unknown true value. Nevertheless, this MAE introduced by rounding cannot be subtracted from the MAE of the difference of the WRRTS and the official video distance as one might intuitively think.

The main advantage of the investigated system over previous studies [9] is the much higher accuracy. Groh et al. proposed an accuracy of 0.8 ± 2.9 m compared to a camera system. Additionally, it has the advantage over the official jumping distance measurement of automating the process, so no FIS certified instructor is needed to label frames in the video manually. Therefore, the WRRTS is also faster than the official distance measurement.

4.2. Projection of 3D WRRTS Measurements into the 2D Image Plane

The projection of the 3D positions captured with the WRRTS showed good correspondence between the WRRTS measurements and the video material. The tracker positions of the jumpers in the videos are close to the projected WRRTS data, indicating that the WRRTS positions are plausible. Although this method does not allow a quantification of the measurement error in the local 3D coordinate system, it serves to provide a visual impression of the measurement quality in an intuitive way.

4.3. Comparison of 3D WRRTS Position Measurements with 3D Camera Vectors

The 3D in-flight positions of the WRRTS are compared to the QDaedalus and camera measurements. As seen in the corresponding Bland–Altman plots, the precision of the X- and Z-coordinate worsens during the jump, in comparison with the camera measurements. Since the bias also varies, a log transformation or investigation of the relative error does not lead to a difference following a normal distribution. Therefore, looking at the whole data range, the determined limits of agreement tend to be too far, which is also mentioned by Bland and Altman [28].

Calculating the MAE in 3D with only the camera measurements from the lateral cameras is not possible due to the missing calibration for the PTZ camera and thus no determined MAE of the Y-coordinate.

However, we calculate the MAE in 3D by combining the MAE of the X- and Z-coordinate determined with the camera measurements with the MAE of the Y-coordinate determined with the QDaedalus system (\(\text{MAE}_Y = 0.08\) m). We thereby find an MAE in 3D of 0.12 m, which is more than a factor of two better than the MAE observed with the QDaedalus system. Since the precision of the differences depends on the precision of both measurement methods, the discrepancy of the MAE shows that the QDaedalus tracking introduces this difference in the MAE.

In comparison with a recent study, which uses dGNSS and achieves a better accuracy of smaller than 0.05 m [15], the advantage of the investigated system is that it is unobtrusive in contrast to the backpack and antennas at the helmet needed for the dGNSS.

Based on the 3D in-flight positions, the WRRTS also measures and transmits the in-flight velocity in real-time. Due to the accurately measured positions, we can assume that the velocity is also measured accurately, with its precision worsening with the distance to the jump-off platform.

4.4. Comparison of 3D WRRTS Angle Measurements with 3D Camera Vectors

The determination of the V-angle has a worse precision compared to the angle determined from the cameras next to the landing hill. One reason for this is that the V-angle
calculated using the orientation of both skis, which both introduce a measurement error. Apart from this, due to the missing calibration of the PTZ camera, no 3D vectors could be used to measure the V-angles. Instead, they were taken directly from the image. The angles measured by the WRRTS are projected onto the image plane, but effects such as radial distortion through the camera lens are not corrected. This may introduce a bias in the comparison of both methods. Another reason for the bias and worse precision may be the less accurate labeling of the PTZ camera images. This is introduced through the bending of the skis in conjunction with the jumper appearing relatively small in the image (see Figure 5). In contrast to this, in the cameras at the side of the landing hill, the jumper appears large, and thus the unbent part of the skis is identified more precisely (see Figure 4).

The investigated system clearly outperforms previous studies also based on IMUs, which achieve an angular precision of $-0.2 \pm 4.8^\circ$ for the lateral angle [13].

5. Conclusions

In this paper, we investigated the accuracy of a WRRTS for ski jumping. The system is based on UWB and IMU and capable of real-time data transmission. Overall, we see good correspondence for the investigated parameters. Table 2 summarizes the findings of this study.

Table 2. MAE between the WRRTS and the respective reference system for all parameters investigated in this study.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Accuracy (MAE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>jumping distance</td>
<td>0.46 m</td>
</tr>
<tr>
<td>3D position</td>
<td>0.12 m</td>
</tr>
<tr>
<td>lateral angle</td>
<td>0.8°</td>
</tr>
<tr>
<td>V-angle</td>
<td>3.4°</td>
</tr>
</tbody>
</table>

The WRRTS and the respective reference system can be used interchangeably, accepting the determined bias and precision of their difference. The bias and precision of the difference is an estimation for the upper limit for the true unknown bias and precision of the WRRTS since the reference system also introduces a measurement error.

Looking at the possible applications of such a tracking system, we see great potential for the live application of the WRRTS during TV broadcasting, since information such as live-speed or trajectory comparison might be fascinating for the spectators. Furthermore, for the application during training, the WRRTS shows high potential to support coaches and sports scientists to improve further the technique of the athletes to jump even further.

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

Funding: This research was funded by German Research Foundation (DFG) grant number Heisenberg: 434/8-1.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of the Friedrich-Alexander-Universität Erlangen-Nürnberg (number 106_13B).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data is available from the authors on reasonable request.
Acknowledgments: Bjoern Eskofier gratefully acknowledges the support of the German Research Foundation (DFG) within the framework of the Heisenberg professorship program (grant number ES 434/8-1). Furthermore, we would like to thank Markus Streicher for their essential contributions regarding project conception, data acquisition, and evaluation.

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

Abbreviations

The following abbreviations are used in this manuscript:

dGNSS differential Global Navigation Satellite System
FIS Fédération Internationale de Ski
IMU inertial measurement unit
LoA limit of agreement
MAE mean absolute error
PTZ pan-tilt-zoom
SEM standard error of the mean
UWB ultra-wideband
WRTTS wearable real-time tracking system

References

12. Bessone, V.; Petrat, J.; Schwitz, A. Ski position during the flight and landing preparation phases in ski jumping detected with inertial sensors. *Sensors* 2019, 19, 2575. [CrossRef] [PubMed]
13. Chardonnens, J.; Favre, J.; Cuendet, F.; Gremion, G.; Aminian, K. A system to measure the kinematics during the entire ski jump sequence using inertial sensors. *J. Biomech.* 2013, 46, 56–62. [CrossRef]


Appendix

B. xLength: Predicting Expected Ski Jump Length shortly after Take-off using Deep Learning
**Abstract:** With tracking systems becoming more widespread in sports research and regular training and competitions, more data are available for sports analytics and performance prediction. We analyzed 2523 ski jumps from 205 athletes on five venues. For every jump, the dataset includes the 3D trajectory, 3D velocity, skis' orientation, and metadata such as wind, starting gate, and ski jumping hill data. Using this dataset, we aimed to predict the expected jump length (xLength) inspired by the expected goals metric in soccer (xG). We evaluate the performance of a fully connected neural network, a convolutional neural network (CNN), a long short-term memory (LSTM), and a ResNet architecture to estimate the xLength. For the prediction of the jump length one second after take-off, we achieve a mean absolute error (MAE) of 5.3 m for the generalization to new athletes and an MAE of 5.9 m for the generalization to new ski jumping hills using ResNet architectures. Additionally, we investigated the influence of the input time after the take-off on the predictions' accuracy. As expected, the MAE becomes smaller with longer inputs. Due to the real-time transmission of the sensor's data, xLength can be updated during the flight phase and used in live TV broadcasting. xLength could also be used as an analysis tool for experts to quantify the quality of the take-off and flight phases.

**Keywords:** wearable sensors; ultra-wideband; inertial measurement unit; performance prediction; sports analytics; performance analysis

1. Introduction

In recent years, the use of tracking systems in sports has become more widespread. As a result, a constantly increasing amount of data are becoming available for analysis. The use of sensors in sports applications was first adopted in a research context, which included studies with small-scale datasets. Over time, wearable devices and sensor systems spread to everyday use and were adapted in professional sports to monitor athletes during training and competition. Thus, data availability increased further, from single runs and actions being analyzed to big data.

Initially, works in the area of sports analytics are mainly based on statistics that are manually extracted during or after the match. MacDonald [1] proposed to train a ridge regression model on variables such as hits, faceoffs, shots, missed shots, blocked shots, and other statistics to predict each player’s contribution to the number of expected goals (xG) in hockey matches. Subsequent work by Rathke [2] translated the approach to soccer and observed that the distance to the goal and the shot angle are the most important variables to predict the xG in soccer. The expected goals metric is a very active field in research [3–8] and has found its way into the application. For example, in the Bundesliga, the highest German soccer league, the expected goals are shown in live TV broadcasting for individual goals and scoring moments and their sum as a conclusion of a match [9].

Instead of manually extracting statistics, later work utilized cameras and inertial measurement units (IMUs) to detect and classify activities in sports automatically. Camera-based
tracking systems are used for example in ball tracking \cite{10-12} and pose estimation \cite{13,14}. However, camera tracking systems are hindered by their expensive cost and are difficult to use under bad weather conditions or in large areas. The other option is sensor-based solutions, which circumvent some of these problems. One main advantage of sensor-based solutions is that they can easily cover large volumes as the sensor is not restricted to a specific area, depending on the technology used. For example, sensors based on IMUs or Global Positioning System (GPS) work independently of reference stations. Sensors based on UWB need a radio connection to reference stations in contrast to a camera-based solution where an unobstructed view between the athlete and the cameras is needed.

Moreover, sensors perform independently from the current weather conditions. Blank et al. \cite{15} used IMUs to detect and classify strokes in table tennis. Kautz et al. \cite{16} explored Deep Neural Networks (DNNs) to classify activities in beach volleyball. Stöve et al. \cite{17} used IMUs to detect individual shots and passes of soccer players with machine learning. Cust et al. \cite{18} provide an overview of model development and performance for machine and deep learning for movement recognition in sports. Best practices have been established in the literature regarding data aggregation, data cleaning, model selection, and other components \cite{16}. Claudino et al. \cite{19} find that out of the 58 studies that they analyzed in their review about injury risk and performance prediction, 26% were about soccer, 22% considered basketball, 10% considered handball, 9% considered Australian football, 9% considered baseball, 9% considered volleyball, 7% considered American football, 5% considered ice hockey, 3% considered rugby, and the remaining 2% addressed field hockey, cricket, and beach volleyball. While a considerable amount of studies have been conducted in the area of performance prediction in sports, some disciplines, such as ski jumping, remain understudied. To the best of our knowledge, no performance prediction study has been conducted in the area of ski jumping so far.

Ski jumping is unique, in contrast to almost all other sports, because there are nearly no amateur athletes. This makes data acquisition generally much more complicated and costly. Therefore, all existing studies on ski jumping tracking are based on relatively small datasets applying various tracking techniques. Elfmark et al. \cite{20} use a differential Global Navigation Satellite System (dGNSS) and video-based pose estimation for performance analysis in ski jumping. They also use the data collected with the dGNSS to determine the aerial phase in ski jumping \cite{21} and assess the steady glide phase \cite{22}. Camera-based tracking was also widely used in ski jumping studies to analyze the take-off \cite{23,24}, flight styles \cite{25}, ski jumping phases \cite{26-28}, dynamics \cite{29}, and aerodynamic forces \cite{30}.

Our research contributes in the following ways. Firstly, we acquired the first large-scale study of ski jumping athletes using wearable sensors, including IMUs and ultra-wideband technology. Secondly and mainly, we contribute the first ski jump length prediction benchmark. Our dataset consists of position, velocity, and skis’ orientation measured during the jumps and additional metadata, including the height of the jumping hill and the weather conditions. We investigate the ability of different deep learning algorithms to predict the jump length of 205 individual athletes on five different ski jumping hills. Specifically, we use a fully connected neural network, two different convolutional neural networks (CNNs), and an LSTM model. Our experiments demonstrate the feasibility of predicting the jumping distances of athletes using deep learning. Here, we investigate the ability of the different models to generalize to unseen athletes and jumping hills. Our results indicate that a pretrained model could be used for new athletes and jumping hills without requiring retraining of the model. Moreover, we explore the behavior of the prediction error with respect to the duration of the time series input and observe a consistent increase in the predictive power of the respective models.

A graphical summary of the proposed contribution is shown in Figure 1.
Figure 1. Two wearable trackers are mounted on the athletes’ skis. They measure the 3D position, 3D velocity, and 3D orientations of the skis. In a later application, the data are transmitted in real time and are used to predict the expected jump length (xLength) using deep learning. Additionally, the pipeline of the study methodology is summarized.

2. Materials and Methods

In the following section, the proposed methods are presented. Firstly, the acquired dataset and the respective tracking system are introduced. Consecutively, the deep-learning architectures and their training and hyperparameter tuning are described. Lastly, we introduce the different experiments run in this study.

2.1. Dataset

The dataset was acquired using a wearable real-time tracking system (WRTTS). It consists of trackers on top of the athlete’s ski bindings and mobile antennas next to the jumping hill. The WRTTS combines an inertial measurement unit (IMU) with ultra-wideband positioning. Information about the working principle can be found in [31–33]. The accuracy of the tracking system has been validated in a previous study [33].

The dataset consists of 2523 jumps acquired during competitions and training of the world’s leading athletes. Every jump includes the 3D trajectory, 3D velocity, and 3D orientation of both skis as well as the wind, wind compensation parameter, gate, gate compensation, and gender. The wind in the dataset is the mean of the tangential wind along the landing hill. This is used in the competitions of the Fédération Internationale de Ski (FIS) to calculate the wind compensation, which is also included in our dataset. The jump length is transformed into points to make competitions fairer, and the wind compensation value is added to compensate for changes in the wind conditions during the competition.

The gate corresponds to the location on the in-run where the athletes start. Since this may change during a competition, this information is crucial since this affects the speed of the athlete at the take-off. Like wind compensation, gate compensation is added to the points of the reached jump length to compensate for changes in the starting gate during a competition.

Additionally, parameters of the ski jumping venues are included and used for the prediction. The venue geometry is described and named according to the ski jumping hill construction standards of the FIS [34]. This involves the height difference (h) between the
take-off table edge and the construction point (K), the hill size (HS), the end of landing area (L), the horizontal distance between the edge of the take-off table and K (n), the start of the landing area (P), the height of the take-off table (s), the height difference between the edge of the take-off table and K (U), the inclination of the tangent at P (β_P), the inclination of the tangent at K (β), the inclination of the tangent at L (β_L). Table 1 shows a summary of the acquired dataset.

Table 1. Summary of the dataset used in this study. Data were acquired during competition and the training of world-leading athletes. The venue data are named according to the ski jumping hill construction standards of the FIS [34].

<table>
<thead>
<tr>
<th>Subjects (female/male)</th>
<th>205 (50/155)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of jumps</td>
<td>2523</td>
</tr>
<tr>
<td>Number of venues</td>
<td>5</td>
</tr>
<tr>
<td>Hillsizes (m)</td>
<td>106, 134, 137, 140</td>
</tr>
<tr>
<td>Jump length range (m)</td>
<td>50.5 to 138.5</td>
</tr>
<tr>
<td>Skill level</td>
<td>Professional</td>
</tr>
<tr>
<td>Sampling rate</td>
<td>20 Hz</td>
</tr>
<tr>
<td>Time series data</td>
<td>3D position, 3D velocity, 3D orientation of both skis</td>
</tr>
<tr>
<td>Athlete data</td>
<td>Athlete ID, gender</td>
</tr>
<tr>
<td>Venue data</td>
<td>P, K, HS, L, U, n, h, β_P, β, β_L, s</td>
</tr>
<tr>
<td>Jump metadata</td>
<td>Wind, wind compensation, gate, gate compensation</td>
</tr>
</tbody>
</table>

The jump length within our dataset, which is the regression goal, is determined with the WRTTS and not the manually labeled video distance as in official competitions. Nevertheless, in a previous study, we showed that the jump length determined by the WRTTS and the official video distance differ by 0.31 ± 0.44 m [33]. Figure 2a shows the distribution of the jump length within the dataset. It ranges from 50.5 to 138.5 m and has two peaks at roughly 93 and 120 m. The median jump length is 97.3 m. In red, the kernel density estimate using Gaussian kernels is depicted.

The data were acquired on five different ski jumping venues ranging from a hill size (HS) of 106 m to 140 m. Figure 2b shows a histogram of the number of jumps per venue with the corresponding HS. The venue with the most jumps is Zhangjiakou, with an HS of 106 m. The ski jumping hill in Zhangjiakou, with an HS of 140 m, has the second most jumps recorded. Both have individually more jumps than the remaining three combined.

Figure 2c shows the distribution of the number of jumps per athlete in the dataset. The median number of jumps per athlete is eleven, and the maximum number is 36 jumps.

![Figure 2a](image1.png)  
(a) Distribution of the jump length.  
![Figure 2b](image2.png)  
(b) Hillsizes.  
![Figure 2c](image3.png)  
(c) Jumps per athlete.  

Figure 2. Distribution of the (a) jump length, (b) jumps per venue, and (c) the number of jumps per athlete present in the dataset of this study. The kernel density estimate using Gaussian kernels is also shown in red for the jump length and the jumps per athlete.
Figure 3 shows the exemplary position, velocity, and orientation of 20 ski jumps. The data are shown for the time range from two seconds before to one second after the take-off. This time range is also the input for the deep learning architectures to predict the jump length.

Figure 3. Example data for 20 ski jumps. The figure shows the 3D position, 3D velocity, and the skis' orientation. The x-axis of the subplots is shared along the columns and the y-axis along the rows.
The origin of the coordinate system is the edge of the take-off table. With respect to the jumping direction, the x-axis is defined as horizontal, forward, the y-axis as horizontal, left, and the z-axis as vertical, upward.

Before the take-off, the position data are similar for the different jumps due to the tracks along the in-run. The differences only occur due to different venues and varying speeds. Due to the in-run, the \( v_y \) is 0 before the take-off since the athlete cannot move in the right–left direction.

2.2. Deep Learning Architectures

To predict xLength, we tested several deep learning architectures. This includes a fully connected network (FCN), CNN, ResNet [35], and LSTM [36]. With the acquired sensor data, we used the FCN as a baseline to explore the feasibility of predicting the jump length. Furthermore, we explored two CNN-based architectures, a standard CNN and a ResNet, to investigate whether the temporal information in the sensor data improves the generalization of the models to different athletes and venues. Lastly, we used an LSTM to model the temporal aspect of the sensor data more directly and assess if this further improves the regression performance.

As a network input, we used the 3D position, 3D velocity, the orientation of both skis, and the following meta information. As meta information, we used the gender/sex, wind, wind compensation, gate, gate compensation, and the venue-related parameters stated in Table 1. For the FCN, we concatenated all features, including metadata, into a single vector and used it as the input for the neural network. In the case of the CNN and ResNet architecture, we created a two-dimensional vector \( x \in \mathbb{R}^{N \times C} \), where \( N \) is the number of sampled sensor values over time, and \( C \) is the number of different sensor signals (i.e., velocity in the x direction or position in the y direction). Additionally, we concatenated the metadata information to the flattened feature vector after the last convolutional layer of the networks, which is followed by fully connected layers. For the LSTM, we used all the measurements for a single time point as an input and obtained the final prediction by performing a sequential prediction over the whole time series. Here, we also included the metadata information after the last LSTM layer, which is followed by fully connected layers.

The athlete or any past-performance-related features such as ranking in the world cup or previous jumps were not included.

2.3. Training and Hyperparameter Tuning

For every deep learning architecture, we performed hyperparameter tuning. We applied a nested 5-fold cross-validation for the hyperparameter tuning using the Bayesian optimization tuning with the Gaussian process implemented in Keras [37]. The nested cross-validation is chosen not to overfit the dataset while performing a hyperparameter tuning and model selection and obtain optimistically biased performance estimations [38]. The search spaces for the different deep learning architectures are depicted in Table 2.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>FCN</th>
<th>CNN</th>
<th>ResNet</th>
<th>LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of fully connected layers</td>
<td>( \in [2, 6] )</td>
<td>( \in [2, 6] )</td>
<td>( \in [2, 6] )</td>
<td>( \in [2, 6] )</td>
</tr>
<tr>
<td>Nodes per fully connected layer</td>
<td>( \in [16, 272] )</td>
<td>( \in [16, 272] )</td>
<td>( \in [16, 272] )</td>
<td>( \in [16, 272] )</td>
</tr>
<tr>
<td>Dropout rate</td>
<td>( \in [0.0, 0.3] )</td>
<td>( \in [0.0, 0.3] )</td>
<td>( \in [0.0, 0.3] )</td>
<td>( \in [0.0, 0.3] )</td>
</tr>
<tr>
<td>Noise</td>
<td>( \in [0.0, 0.1] )</td>
<td>( \in [0.0, 0.1] )</td>
<td>( \in [0.0, 0.1] )</td>
<td>( \in [0.0, 0.1] )</td>
</tr>
<tr>
<td>Batch size</td>
<td>( \in [16, 272] )</td>
<td>( \in [16, 272] )</td>
<td>( \in [16, 272] )</td>
<td>( \in [16, 272] )</td>
</tr>
<tr>
<td>Number of specific layers</td>
<td>( \in [1, 6] )</td>
<td>( \in [1, 6] )</td>
<td>( \in [1, 6] )</td>
<td>( \in [1, 6] )</td>
</tr>
<tr>
<td>Filters/LSTM units per layer</td>
<td>( \in [16, 272] )</td>
<td>( \in [16, 272] )</td>
<td>( \in [16, 272] )</td>
<td>( \in [16, 272] )</td>
</tr>
</tbody>
</table>
Since the dataset contains strongly dependent data and physical processes, data augmentation techniques are limited and must be chosen carefully. For example, the skis’ orientation affects the aerodynamic drag, which affects the speed and uplift. This again influences the trajectory and the jump length. Therefore, standard augmentation techniques such as random rotation are not applicable.

We, firstly, doubled the number of jumps in the training dataset by mirroring them at the x–z plane. This is equivalent to swapping left and right from the perspective of the ski jumper. Secondly, we added Gaussian noise to all input variables as data augmentation. Apart from these data augmentations, we used standardization as another preprocessing step, i.e., scaling all data to a mean of 0 and a standard deviation of 1.

For the training process, we use the Adam optimizer [39] in combination with a reduction of the learning rate when the loss reaches a plateau. We use the mean squared error as a loss function and terminate the training process via early stopping.

2.4. Experiments

Using the previously presented dataset, we perform several experiments. Firstly, we compare the performance of the different deep learning architectures to predict the jump length one second after the take-off. For the comparison, we evaluate the model performances by investigating the following metrics. The residual \( r \) is, within this work, defined as

\[
 r = x_{\text{true}} - x_{\text{prediction}},
\]

where \( x_{\text{prediction}} \) is the ski jump length predicted by a DNN and \( x_{\text{true}} \) is the measured ski jump length. To compare the networks’ performance on the whole dataset, we investigate the mean of the residual also called bias

\[
\mu = \frac{\sum_{i=1}^{N} r^{(i)}_{\text{true}} - r^{(i)}_{\text{prediction}}}{N},
\]

where \( N \) is the number of investigated jumps. Additionally, we analyze the standard deviation of the residual

\[
\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (r^{(i)} - \mu)^2}.
\]

The performance of the DNNs is summarized in terms of the mean absolute error (MAE), which is calculated as

\[
\text{MAE} = \frac{\sum_{i=1}^{N} |x^{(i)}_{\text{true}} - x^{(i)}_{\text{prediction}}|}{N}.
\]

Secondly, we test the generalization capabilities of the different deep learning architectures on unseen athletes and venues. Therefore, we perform the outer split of the nested cross-validation by athletes or venues, respectively. For example, we perform the hyperparameter tuning and model selection on four different ski jumping venues and evaluate this model on the remaining venue.

In addition, we investigate the influence of the input length after take-off on the prediction performance. Since, for longer inputs, the number of weights for the fully connected network would drastically increase; we use a CNN for this experiment. Therefore, we run a hyperparameter tuning for the ResNet for input lengths from 0.5 to 4.0 s after the take-off.

3. Results

The following section presents the results obtained from the experiments described in the previous section. We start with comparing the results of the different deep learning architectures for the input interval of two seconds before to one second after the take-off.
This also includes investigating the generalization to new athletes and venues. After that, we present the results for the accuracy investigation as a function of the input length.

3.1. Prediction of $x_{Length}$

Table 3 shows the MAE, mean and standard deviation of the prediction error. This includes the folds split by athletes and venues and all tested deep learning architectures. We can see that for the data split by athletes, the ResNet has the best prediction of the jump length with an MAE of 5.3 m, a mean error of 0.1 m and a standard deviation of 6.8 m. The LSTM has the same absolute value for the mean residual with −0.1 m, and the FCN has the same standard deviation of the residual as the ResNet. In general, all architectures achieve similar performance.

Looking at the performances when we split the data by venues, the performances of all architectures decrease in terms of MAE. In addition, the performance of the ResNet decreases in all three metrics. However, the ResNet still has the smallest MAE with 5.9 m compared to the other deep learning architectures. The ResNet reaches a mean residual of 0.7 m, which is also the smallest value among the different architectures. For the standard deviation of the residual, in contrast, the FCN has the smallest value 7.4 m. The ResNet reaches a standard deviation of 7.6 m. Especially the LSTM has problems with the generalization over different venues, which results in the largest MAE for the venue split 9.2 m.

Table 3. Mean absolute error, mean, and standard deviation of the residual of different deep learning architectures. The prediction of an FCN, CNN, ResNet, and LSTM is tested for nested cross-validation split by athletes and venues.

<table>
<thead>
<tr>
<th></th>
<th>Athlete Split</th>
<th>Venue Split</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE (m)</td>
<td>Mean (m)</td>
</tr>
<tr>
<td>FCN</td>
<td>5.4</td>
<td>0.5</td>
</tr>
<tr>
<td>CNN</td>
<td>5.5</td>
<td>0.8</td>
</tr>
<tr>
<td>ResNet</td>
<td>5.3</td>
<td>0.1</td>
</tr>
<tr>
<td>LSTM</td>
<td>5.5</td>
<td>−0.1</td>
</tr>
</tbody>
</table>

Figure 4a shows the distribution of the true jump lengths, and Figure 4c shows the distribution of the predicted jump lengths. Qualitatively, we can see a good agreement between the two distributions.

Figure 4d shows the residual versus the true jump length. The absolute error is relatively constant over the whole span of true jump length, and we do not observe a relative error; i.e., an absolute error increases with increasing jump length. As the left middle plot shows, the predicted jump length at both ends tends toward the mean prediction. The projection of the residual is shown in Figure 4e. The residual follows a Gaussian distribution.

3.2. Dependency of Prediction Accuracy on Input Length

In the previous subsection, we investigated the performance of different deep learning architectures in predicting the expected jump length one second after the take-off. In this section, we analyze how the accuracy of the prediction changes with the input length, i.e., the time after the take-off.

Figure 5 shows the residual for the ResNets trained on different input lengths. The mean residual is depicted as a function of the input length of the time series data. The error bars show the standard deviation of the residual.
The standard deviation of the prediction error decreases from 0.5 to 4.0 s. The mean prediction error is approximately constant for all input lengths.

**Figure 4.** Subplot (a) shows the distribution of the true ski jump length in the dataset. Subplot (c) shows the distribution of the jump length predicted by the ResNet with the folds split by athletes. In subplot (b), the prediction is plotted versus the true jump length. The color represents the density of points calculated using a kernel-density estimate using Gaussian kernels. The brighter the color, the higher the density of the points. Subplot (d) shows the residual as a function of the true jump length. The color again represents the density of the points. Subplot (e) shows the distribution of the residual.

Additionally, on the right y-axis, the number of jumps in the dataset with a duration longer than the input length of the ResNet is shown. The number of samples is constant until 2.5 s after the take-off. For 3 s after the take-off, the number of samples slightly
decreases. For 3.5 s, the number is smaller, and for an input length of 4.0 s after the take-off, there is a factor of more than two fewer jumps than directly after the take-off.

![Figure 5](image-url)

**Figure 5.** The left y-axis corresponds to the prediction accuracy for ResNets trained on various input lengths after the athlete’s take-off. In blue, the mean of the residual is plotted with the standard deviation of the residual as the y-error bar. Additionally, the number of jumps with a minimum duration, as shown on the x-axis, is plotted in red.

4. Discussion

This work aimed to develop the first ski jump length prediction. The dataset and the prediction results are discussed in the following section.

4.1. Dataset

The dataset covers a wide range of jump lengths and athletes, which leads to good prediction accuracy over jump lengths and generalization over athletes. Even though the dataset has the largest number of different ski jumping venues and athletes used in a tracking study [21,31,33,40–44], the number of venues is relatively small and unevenly distributed compared to the number of athletes. This could be improved in the future.

Having only sensors on the skis is, on the one hand, unobtrusive and, therefore, perfect for application in competition, especially for world-leading athletes in a dangerous sport such as ski jumping. On the other hand, the tracking system does not cover much information about the athlete’s movement. This would especially be important during the take-off. Therefore, an additional camera next to the take-off table would be beneficial to understand the take-off better and whether an athlete is jumping off too early or too late. For example, pose estimation could be applied to measure knee and hip angles to improve the prediction.

The time series data used in this study were sampled with 20 Hz, which is relatively low considering the high speeds of the athletes at the take-off. The UWB measurements are sampled with 20 Hz, but the internal sampling rate of the IMU sensor is much higher at 1000 Hz. Using this raw and highly sampled IMU data would probably be beneficial for predicting xLength, since the take-off could be analyzed in much more detail. These raw data, unfortunately, were not available in this study.

Future work could consider extending the dataset to ski flying and calculating xLength for ski flying. One challenge is that the data acquisition is even more cumbersome since fewer ski jumping competitions exist.

4.2. Prediction of xLength

First of all, we have to emphasize that having a perfect prediction of the ski jump length shortly after the take-off would assume that the remaining flight and landing...
phase do not influence the jump length, which is, of course, not the case. Therefore, the prediction accuracy of the jump length has a fixed limit. To get as close as possible to this unknown limit is, therefore, the goal of such a prediction. Additionally, since this is the first performance prediction in ski jumping, we cannot compare the prediction results to previous studies.

Checking the generalization to new athletes, all deep learning architectures have a similar prediction performance. The ResNet, however, has the most accurate prediction of xLength in terms of MAE, mean and standard deviation of the prediction error. It could be expected that the generalization to new athletes is no problem for the deep learning architectures since the trajectories differ not much between individual athletes but rather between individual jumps. In addition, the number of athletes is high in the dataset, which benefits this generalization.

All deep learning architectures have worse performance in the MAE for the generalization of new ski jumping venues. It could be expected that the generalization to new venues is worse than to new athletes since the distribution over the venues is not uniform, and the vast majority of jumps are from only two venues. Considering this, the generalization to new venues is better than expected. Additionally, for a possible application, the generalization to new venues should not be a problem since only a few professional ski jumping venues exist, which are thus repeatedly used for competitions. Additionally, acquiring data would be possible during the training runs, which are completed at every venue before the competitions.

Looking at the residual of the ResNet with the data split by athletes (Figure 4), we see that the predictions tend toward the mean jump length. This is not surprising, since the number of samples is much smaller at both ends of the jump length range.

In future work, we will investigate if methods from the area of out-of-distribution generalization can further improve the algorithm’s precision to unseen venues and athletes [45]. Other possibilities include using transfer learning to utilize IMU data from other application areas, where data are more readily available, as completed in earlier work [46].

Another point to mention is that the WR/TTS measuring the jump length accuracy is $0.31 \pm 0.44$ m compared to the official video-based measurement, which is manually labeled and rounded to 0.5 m [33]. Since the deep learning architectures were trained on the jump length, determined with the WR/TTS, this would affect a possible application in a competition where the official video-based measurement is used.

In addition, to the prediction of xLength, we also tested interpreting the deep learning models using SHAP (SHapley Additive exPlanations) [47] but did not obtain consistent results over the different folds and deep learning architectures. Future work could further address explainable AI approaches or extracting features to calculate feature importance. This would benefit the athletes, coaches, and sports scientists to better understand the complex sport of ski jumping.

Coaches and sports scientists could use xLength directly after take-off to quantify the quality of the take-off. This would make the evaluation of the take-off objective and, therefore, comparable between athletes.

4.3. Dependency of Prediction Accuracy on Input Length

The expected jump length is predicted more precisely with longer inputs, as one would expect, since more flight trajectory data are available and the time to the predicted location is shorter.

Naively, one would expect that the prediction becomes more and more precise. However, the prediction accuracy is almost constant for inputs larger than three seconds. A reason for the performance might be that the jump length is harder to predict due to far fewer data. Additionally, for large jump lengths, which correspond to high jumping durations, the athletes land in the curvature at the bottom of the landing hill, which is not in detail described in the ski jumping hill data input into the network.

Coaches and sports scientists could use xLength of different input lengths to analyze and compare individual jumps. For example, if xLength becomes smaller at a specific time
during the jump, this indicates that the athlete made an error during the flight phase. In contrast, if xLength enlarges during the ski jump, this indicates that the athlete performs well in the flight phase. In addition, it might be possible to differentiate between athletes with a better take-off versus athletes with better flight phases.

Additionally, due to the live data transmission of the sensors, the predicted jump length could be updated during the whole jump. In combination with the calculated distance to beat, which is calculated before the athlete starts, this could be used to determine a live probability of reaching a specific position in the ranking or a determination of jump length given away compared to the take-off.

Another possible application could be to use it in a live visualization showing the expected landing corridor, which with increasing prediction accuracy, becomes smaller. So, this feature might, on the one hand, help experts analyze jumps and, on the other hand, make TV broadcasting more interesting for laypersons.

5. Conclusions

This work aimed to develop the first ski jump length prediction. Therefore, we analyzed the first large-scale ski jumping dataset included in a research study. The data are measured by wearable trackers on the athletes’ skis, measuring the 3D position, 3D velocity, and the skis’ orientation. Using the ski jump length also determined with the tracking system, we performed a hyperparameter tuning for different deep learning architectures.

Firstly, we compared the performance of the deep learning architectures for a prediction one second after the take-off. Here, we obtain the best results for the generalization to new athletes using a ResNet. This achieves an MAE of 5.3 m, mean residual 0.1 m, and a standard deviation of the residual of 6.8 m. For the generalization to new venues, we obtain slightly different results. Therefore, the ResNet has the smallest MAE with 5.9 m and a standard deviation of the residual of 7.6 m.

Another question was how the prediction accuracy changes with the input time after the take-off. Therefore, we investigated different ResNets trained on various input lengths from 0.5 to 4 s after the take-off. Thereby, the standard deviation of the residual becomes smaller with increasing input lengths.

We think that the proposed xLength can be used for live broadcasting due to the live data transmission and for the retrospective analysis of jumps by experts. This includes the quantification of the take-off, thus comparability between jumps and athletes, as well as the analysis during the flight to determine errors when xLength decreases.

Future work should also consider more explainable approaches than neural networks to obtain a better understanding of ski jumping. Additionally, a camera at the take-off might be beneficial to obtain more information about the take-off process, since the sensors on the skis do not provide any information about the body’s movement.

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

Funding: This research was funded by German Research Foundation (DFG) grant number Heisenberg 434/8-1.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of the Friedrich-Alexander-Universität Erlangen-Nürnberg (number 106_13B).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Due to confidentiality agreements, neither the data nor the source of the data can be made available.
Acknowledgments: Bjoern Eskofier gratefully acknowledges the support of the German Research Foundation (DFG) within the framework of the Heisenberg professorship programme (grant number ES 434/8-1). Furthermore, the authors would like to thank Markus Streicher and Lars Schuricht for their help regarding the dataset and valuable input.

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

Abbreviations

The following abbreviations are used in this manuscript:

CNN Convolutional neural network
FCN Fully connected network
FIS Fédération Internationale de Ski
h Height difference between take-off table edge and K
HS Hill size (distance between edge of take-off table and L)
IMU Inertial measurement unit
K Construction point
L End of landing area
LSTM Long short-term memory
MAE Mean absolute error
n Horizontal distance between take-off table edge and K
P Start of landing area
s Height of the take-off table
Uz Height difference between take-off table edge and the lowest point
WRTTS Wearable real-time tracking system
\( \beta_p \) Inclination of the tangent at P
\( \beta \) Inclination of the tangent at K
\( \beta_L \) Inclination of the tangent at L

References

13. Badiola-Bengoa, A.; Mendez-Zorrilla, A. A Systematic Review of the Application of Camera-Based Human Pose Estimation in the Field of Sport and Physical Exercise. Sensors 2021, 21, 3996. [CrossRef]


Appendix

C. Wearable Sensors for Activity Recognition in Ultimate Frisbee Using Convolutional Neural Networks and Transfer Learning
Wearable Sensors for Activity Recognition in Ultimate Frisbee Using Convolutional Neural Networks and Transfer Learning

Johannes Link *, Timur Perst, Maike Stoeve © and Bjoern M. Eskofier ©

Machine Learning and Data Analytics Lab, Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91052 Erlangen, Germany; timur.perst@fau.de (T.P.); maike.stoeve@fau.de (M.S.); bjoern.eskofier@fau.de (B.M.E.)

* Correspondence: johannes.link@fau.de

Abstract: In human activity recognition (HAR), activities are automatically recognized and classified from a continuous stream of input sensor data. Although the scientific community has developed multiple approaches for various sports in recent years, marginal sports are rarely considered. These approaches cannot directly be applied to marginal sports, where available data are sparse and costly to acquire. Thus, we recorded and annotated inertial measurement unit (IMU) data containing different types of Ultimate Frisbee throws to investigate whether Convolutional Neural Networks (CNNs) and transfer learning can solve this. The relevant actions were automatically detected and were classified using a CNN. The proposed pipeline reaches an accuracy of 66.6%, distinguishing between nine different fine-grained classes. For the classification of the three basic throwing techniques, we achieve an accuracy of 89.9%. Furthermore, the results were compared to a transfer learning-based approach using a beach volleyball dataset as the source. Even if transfer learning could not improve the classification accuracy, the training time was significantly reduced. Finally, the effect of transfer learning on a reduced dataset, i.e., without data augmentations, is analyzed. While having the same number of training subjects, using the pre-trained weights improves the generalization capabilities of the network, i.e., increasing the accuracy and F1 score. This shows that transfer learning can be beneficial, especially when dealing with small datasets, as in marginal sports, and therefore, can improve the tracking of marginal sports.

Keywords: inertial measurement unit; activity recognition; sensor-signal-based machine learning; convolutional neural network; deep learning; wearable sensors; marginal sports; transfer learning

1. Introduction

Monitoring athletes has been in the interest of players and coaches for a very long time. As early as the 1930s, two Germans used heart rate observations of middle-distance runners to improve the training procedure of their athletes. The proposed interval training has been adopted worldwide and has enabled athletes to set new records. The hunt for new training strategies based on scientific data of athletes has continued ever since [1]. Traditionally, monitoring was done by manually observing the athletes, which was time-consuming and needed many experts in the field. With the increasing computational capacities of computers, smaller and more affordable hardware, and the development of powerful algorithms, monitoring athletes has become feasible for a broad spectrum of protagonists [1]. A highly objective and reliable qualitative analysis using wearable sensors is an accepted alternative to traditional lab-based assessment. Wearable sensors are portable, low-cost, easy to use, and usually do not limit athletes in their movement [2]. Their success has opened the path for commercial use, for example, at pro soccer clubs, whose players wear sensors even during competitive matches [3]. However, the development of more advanced systems classifying distinctive movement patterns is still an active research topic [2].
Another important goal of monitoring athletes is the prevention of injuries. Even though most acute injuries occur in contact sports like American football, injuries resulting from overuse often happen in non-contact sports. A study amongst collegiate athletes revealed that over 29% of the injuries result from overuse, with a much higher estimated number of unreported cases [4]. Fine-grained monitoring of players’ actions can help to detect high training loads for specific body parts, which have been linked to injuries, thus allowing the coach to adjust the training individually [5].

In the literature, a great variety of sport-specific activity recognition systems exists, which can be used for fine-grained, athlete-specific monitoring [6]. To detect actions reliably, a sufficient amount of training data needs to be recorded, which is not only cumbersome but sometimes not possible at all, especially in marginal sports. To combat the lack of data, transfer learning approaches use data from another domain for pre-training. In a second step, model parameters are fine-tuned on the target dataset [7].

Our research contributes in the following ways. Firstly, we developed the first activity recognition system for Ultimate Frisbee. Therefore we trained a Convolutional Neural Network CNN to distinguish seven different throwing techniques plus catches. Secondly, we pre-trained the same architecture on an existing volleyball dataset [8] to investigate the potential of transfer learning for marginal sports. Thirdly, we explored the generalization capabilities of transfer learning when dealing with small-scale datasets.

A graphical summary of the procedure in this work is shown in Figure 1.

Figure 1. A visual summary of the workflow of our paper. We developed the first human activity recognition system for Ultimate Frisbee. Therefore, we use a convolutional neural network. Additionally, we investigated the possible improvement of the classification using a network pre-trained on volleyball activities.
2. Related Work

2.1. Sensor Based Human Activity Recognition in Sports

The scientific and commercial interest in HAR has been growing in recent years, leading to extensive research [9]. Traditionally, vision-based solutions have been used for activity recognition. However, those come with the disadvantages of being expensive, causing privacy issues, and having to be mounted at specific positions, which may not be feasible for minor sports outside of huge stadiums [10]. On the other hand, sensor-based systems using IMUs like accelerometers, gyroscopes, and magnetometers have been adopted in the research community [6]. Historically, accurate sensors had to be custom-built for each application. Smaller, cheaper, and better availability of sensors have made it possible to analyze numerous different sports and movements in recent years [9]. Cust et al. [6] provide a broad overview of machine and deep learning for sport-specific movement recognition. Their review includes a comparison of several sensor- and vision-based solutions for HAR in sports. Even though the respective authors used many different methods for their HAR systems, the general workflow included: preprocessing, segmentation, feature extraction, dimensionality reduction, and classification. During the preprocessing step, many authors propose low pass filters to remove unwanted noise [10]. There has been successful work with transforming the signal into the frequency domain, such as applying a wavelet transformation [11] or a fast Fourier transform [12]. The trend for the classification task is more and more towards neural networks, specifically to CNNs and long short-term memory (LSTM) [13], due to their good performance. CNNs have been initially used for image detection tasks because they preserve the spatial information of neighboring pixels [14]. Time series data also include spatial information between subsequent measurements [15]. In addition, deep learning-based approaches have the advantage of skipping the feature extraction step. Chen et al. [16] proved that CNNs can compete with other algorithms, reaching an accuracy of 93.8% when classifying eight tasks of daily living.

By now, the research community has developed a variety of classification pipelines that specifically target one discipline of sports. Anand et al. make use of the better availability of sensors and use smartwatches to classify strokes in swing sports, like tennis or badminton, and give the players feedback on their technique [17]. Using a bi-directional long short-term memory network, they achieve accuracies between 78.9% and 94.6% depending on the swing sport. Brock et al. classify common errors of ski jumpers and chose IMUs over a vision-based solution to account for bad weather conditions and to get numerically comparable data [18]. They achieve error recognition rates between 60% and 75%. Kautz et al. classified actions from beach volleyball players using data from a wrist-worn accelerometer. Amongst others, they programmed a deep convolutional neural network, which reached an accuracy of 83% [8].

2.2. Transfer Learning

Transfer learning refers to the technique of using knowledge gained in one domain (source domain) and applying it in another domain (target domain). Its advantages over training a classifier without prior knowledge include shorter training durations and the ability to solve previously unsolvable tasks.

Transfer learning has been applied to CNNs in various domains. This includes, among others, image classification [19–23] and image segmentation [24,25]. In order to improve their accuracy, some of the studies used several data pre-processing and augmentation techniques.

Apart from this, the review by Cook et al. [26] provides a broad general overview of transfer learning for activity recognition. Instance-based transfer learning can be used when the source and target domain are the same or very similar, but the task differs. Training samples are reweighted and directly fed into the classifier for the target task [27]. Tommasio et al. [28] use a form of parameter transfer where they split the parameters of a Support Vector Machine (SVM) w into two parts, where one is shared by the two tasks. Fawaz et al. [29] investigated transfer learning for time series classification and came to the
conclusion that the choice of the source dataset heavily influences the success or failure of a transfer learning-based approach. Better generalization capabilities can be reached with a suitable dataset for the target task. There has also been work on transfer learning applied to HAR. Morales et al. [30] were the first with the idea of transferring low-level features, learned in the lower layers of a CNN for an HAR task.

3. Materials and Methods

The following chapter describes the setup and methods used in this work. First, the activity detection pipeline is explained, which is used to extract relevant samples from the continuous sensor data stream. Then, the labeling procedure and the utilized augmentation techniques are presented. Next, the classification pipeline is described. The classifier was trained from scratch and compared to a network pre-trained on a beach volleyball dataset as a source domain. This transfer learning approach is explained, followed by the evaluation techniques.

3.1. Datasets

This section describes the two datasets used as a source and target dataset. Furthermore, we will explain how the studies to record the data were conducted and how actions were detected. The description of the volleyball dataset by Kautz et al. [8] is followed by a short introduction to Ultimate Frisbee. Afterward, we present the frisbee dataset recorded for this work.

3.1.1. Volleyball Dataset

This dataset consists of three-dimensional acceleration data of common actions in beach volleyball. It was acquired by Kautz et al. for their paper on “Activity recognition in beach volleyball using a Deep Convolutional Neural Network” [8]. Each of the 30 participants was outfitted with a wrist-worn accelerometer, which sampled at 39 Hz. The signal was recorded with 14 bits per axis and truncated at ±16 g. Note that the axis x, y, and z refer to the sensor’s coordinate system. There was no transfer to real-world coordinates. The players’ experience ranged from beginners to professional level athletes. The overall goal was to develop a recognition and classification system which extracts relevant sections from the continuous input and assigns them to a class. The classification incorporates ten different volleyball activities, including, among others, the underhand serve, block, and dig. Table 1 shows a summary of the two datasets.

Table 1. Summary of the volleyball and frisbee dataset used in this study.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Volleyball</th>
<th>Frisbee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actions of interest (without null class)</td>
<td>4284</td>
<td>3695</td>
</tr>
<tr>
<td>Subjects (female/male)</td>
<td>30 (11/19)</td>
<td>14 (2/12)</td>
</tr>
<tr>
<td>Skill Level</td>
<td>Beginner to professional</td>
<td>Experienced amateurs</td>
</tr>
<tr>
<td>Sensor</td>
<td>BMA280</td>
<td>Portables NilsPod</td>
</tr>
<tr>
<td>Sampling rate</td>
<td>39 Hz</td>
<td>512 Hz downsampled to 39 Hz</td>
</tr>
<tr>
<td>Sensor placement</td>
<td>Wrist of dominant hand</td>
<td>Manually labeled video</td>
</tr>
</tbody>
</table>

The first step of the processing pipeline, the detection of potentially relevant signal segments, was done directly on the worn microcontroller. Since relevant actions include a ball contact, the idea was to detect the high-frequency spikes in the acceleration data. This was achieved by passing the signal through a high-pass filter, computing the $L_1$ norm, and smoothing the signal using a low-pass filter. If the signal does not exceed a certain threshold, it is rejected. However, this approach was not sufficient to detect actions accurately. To further refine the process, the authors calculated the swing movement by averaging the absolute acceleration in all three directions for 200 ms before the impact. With the swing...
movement and the amplitude at the peak, Kautz et al. trained a decision tree to discard irrelevant actions. For a more detailed description of the action detection pipeline and the recorded dataset, see the original publication [8]. We use the extracted actions of interest from the dataset for our work.

The classification task used multiple feature-based classifiers: support vector machine, k-nearest-neighbor, Gaussian naive Bayes, decision tree, random forest, and VOTE as a meta classifier. A deep convolutional neural network (DCNN) outperformed all classifiers with hand-crafted features.

3.1.2. Ultimate Frisbee Dataset

In recent years, Ultimate Frisbee has been one of the fastest-growing sports globally. It has its origins in the 1950s in the United States and has since spread worldwide [31].

Ultimate Frisbee is played seven vs. seven on a field with the length of a soccer field and half its width. There are 15 to 18 m deep endzones comparable to American Football at both ends. A player can score a point for their team by catching the frisbee in the endzone. Teammates can pass the frisbee to each other as they please. After a catch, there is a ten-second time limit to get rid of the frisbee. A failure to do so, or an incomplete pass, results in a turnover. Travelling is—except for a basketball-style pivot step—forbidden. As a non-contact sport, every kind of physicality is illegal.

This work focuses on three basic throws: backhand throws, forehand throws, and overheads—sometimes called a hammer. Figure 2 shows a sketch of the basic throwing techniques. Depending on the release angle, the frisbee moves on a different trajectory. Therefore, forehand and backhand throws are further separated by the frisbee’s path: flat, outside-in, and inside-out. An example of the variety of backhand throws by a right-handed player is depicted in Figure 3. Table 2 shows a description of all throwing techniques together with the number of samples per class contained in our frisbee dataset.

Figure 2. The three main throwing techniques in Ultimate Frisbee are, from left to right, forehand, overhead and backhand. The sensor is placed at the wrist of the dominant hand. The sensor (white) is fixed with a wristband (black).
Figure 3. Depending on the angle of the release, the frisbee moves on a different trajectory. The different throwing types are shown for backhand throws.

Table 2. Description of the Ultimate Frisbee action classes used in this study for throwing with the right hand. The different angles of the frisbee during the release affect the trajectory of the frisbee, which are the namesakes for the throwing technique. Additionally, the number of samples of the respective class is specified.

<table>
<thead>
<tr>
<th>Type of Action</th>
<th>Description</th>
<th>Number of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backhand flat</td>
<td>The athlete is standing with the shoulder axis pointing to the throwing target. The frisbee is moved from the back shoulder towards the target in front of the body. The frisbee is flat during the release. It is the most common frisbee throw.</td>
<td>346</td>
</tr>
<tr>
<td>Backhand outside-in</td>
<td>The basic movement is the same as the backhand flat but with the top of the disc facing towards the thrower during the release.</td>
<td>295</td>
</tr>
<tr>
<td>Backhand inside-out</td>
<td>The basic movement is the same as the backhand flat but with the top of the disc facing away from the thrower during the release.</td>
<td>305</td>
</tr>
<tr>
<td>Forehand flat</td>
<td>The athlete is standing with the shoulder axis perpendicular to the target. The frisbee is moved on the side of the throwing hand next to the torso. The frisbee is flat during the release.</td>
<td>324</td>
</tr>
<tr>
<td>Forehand outside-in</td>
<td>The basic movement is the same as the forehand flat but with the top of the disc facing towards the thrower during the release.</td>
<td>313</td>
</tr>
<tr>
<td>Forehand inside-out</td>
<td>The basic movement is the same as the forehand flat but with the top of the disc facing away from the thrower during the release.</td>
<td>307</td>
</tr>
<tr>
<td>Overhead</td>
<td>The athlete stands with the shoulder axis perpendicular to the target. The frisbee is moved over the head, similar to a slap shot.</td>
<td>341</td>
</tr>
<tr>
<td>Catch</td>
<td>The athlete catches the frisbee with the dominant hand or both hands.</td>
<td>1556</td>
</tr>
<tr>
<td>Attempted catch</td>
<td>The athlete tries to catch the frisbee with the dominant hand or both hands, but the frisbee bounces off or falls to the ground.</td>
<td>148</td>
</tr>
<tr>
<td>Null class</td>
<td>Non-frisbee actions during the data acquisition like running, clapping, and undefined motions.</td>
<td>631</td>
</tr>
</tbody>
</table>

Since we wanted our classifier to distinguish between flat, outside-in, and inside-out throws, which require players to have an advanced level of skill, we recruited most of the players from a local Ultimate Frisbee team. We recorded 14 participants (12 male, 2 female) who throw mainly with their right hand. The mean age of the players was 31 years (standard deviation: 7 years). Everyone except one participant reports at least four years of experience in Ultimate Frisbee (mean: 11 years). Except for one player, everyone has played Ultimate Frisbee competitively in any form of club or association. Two players have participated in the world championships.

Each player was outfitted with an IMU (Portabiles NilsPod) at the wrist of their dominant hand (see Figure 3). The sensor initially sampled at 512 Hz and recorded
acceleration and gyroscope data in all three dimensions. The sensors write the recorded data to the built-in memory. The measurement range for the accelerometer is $\pm 16 \text{ g}$ and $\pm 2000^\circ \text{ per second}$ for the gyroscope.

One recording session featured two players at once. After a couple of throws to warm up, the players positioned themselves at a predefined distance. The distances visually marked by pylons were ten meters in round one, 15 m in round two, and 25 m in round three. Some of the participants also did a fourth-round with 30 m distance. Each player threw each kind of throw at least five times in each round. Table 2 lists the types of actions and the number of occurrences in the dataset. For labeling the activities, the data acquisition was filmed using a tripod-mounted wide-angle action camera.

3.2. Activity Detection

In order to classify the different throws, potentially relevant actions had to be detected first. Comparing the acceleration and gyroscope plots with the reference video reveals that actions involve high-frequency peaks in the plots. Furthermore, it was essential to exclude another ordinary and somewhat similar-looking but uninteresting activity: Running to retrieve a not caught frisbee. For the action detection, the gyroscope data was disregarded. For this purpose, we first calculated the acceleration norm and used the z-score algorithm to detect peaks [32]. We used a window size of 1.6 s, a threshold of 3.5, and an influence of 0.5.

During continuous actions (like running), the moving standard deviation of the signal is higher than when the subject is in an idle state. This property helps to more accurately detect throws because they usually follow a situation where the player is, by rule, not moving. It can also detect catches because even though the participant might run to make a catch, the frisbee’s abrupt deceleration transfers much energy to the player’s hand and thus creates very high-frequency peaks.

Since this first step has the issue of recognizing multiple peaks for only one action, peaks were forced to be at least 400 samples (0.78 s) apart. If this condition was not met, only the first peak gets included due to the assumption that the first high-frequency peak indicates the initial action, whereas the latter ones are only post-pulse oscillations. The gyroscope data were disregarded for the activity detection and only used during the classification. After the peak detection, we resampled the IMU data to 39 Hz to match the volleyball data.

3.3. Augmentation Techniques

In order to get a more robust classifier and achieve better generalization results, we employed three random-based augmentation techniques. Each peak corresponds to one detected action, and the classifier works with samples of a fixed length. The continuous stream of acceleration and gyroscope data is cut into windows of a fixed length which serve as an input feature vector $x \in \mathbb{R}^{6 \times l}$ with $l$ being the number of samples in the window. The window starts 0.8 seconds before the peak and lasts until 1.0 s after the peak ($l = (0.8 \text{ s} + 1.0 \text{ s}) \cdot 39 \text{ Hz}$). Augmented samples are then generated by adding three randomly moved windows per detected action. The maximum shift of the peak thereby is 0.2 s.

Furthermore, we tripled the number of samples by rotating existing ones around a random angle in 3D space. During the training phase, a normally distributed noise is added to each channel as the third augmentation technique. All these techniques are applied to both the source and the target dataset.

Additionally, we rebalanced the datasets by a combination of the synthetic minority over-sampling technique (SMOTE) and a clean-up by the edited nearest neighbor (ENN) algorithm. Since SMOTE can use marginal outliers, which generates unsatisfactory training samples, ENN is used to remove the generated samples, which are in the feature space far away from the majority of the class [33,34].
3.4. Classification

For classification, a CNN was developed. The architecture was partly adapted from Fawaz et al. [29]. It consists of two input layers: one for the acceleration data and one for the gyroscope data. Three convolutional layers follow each input layer. Batch normalization is used after each convolutional layer in order to reduce the training times and achieve better results [35].

The widths of the kernels of the convolutional layers are from the first to the last layer: eight, five, and three. After the convolutional layers, a global maximum pooling layer is used. Global average pooling has the same effect on the accuracy but has severe disadvantages concerning the training times [36]. In contrast to the findings of Boureau et al. [36], no improvement in accuracy could be observed when using maximum pooling. The results of the pooling layers are concatenated and passed via one fully connected layer with 64 neurons to the output layer. The convolutional and fully connected layers use the rectified linear unit (ReLU) as an activation function. The last layer uses the softmax function to assign probabilities to each class. The architecture of the network is depicted in Figure 4. Additionally, in Table 3 the network architecture is described in detail.

![Image of the architecture of the convolutional neural network](image)

**Figure 4.** The architecture of the convolutional neural network, which was used in this work. The output dimensions of the layers are given at the top right of each row. The connections of the last two fully connected layers were omitted for better readability. The architecture was partly adapted from Fawaz et al. [29].

During the training procedure, an optimizer implementing Adam’s algorithm is used [37]. Even though Adam computes individual adaptive learning rates for different weights, the implementation has shown more stable results when gradually decreasing the initial learning rate $\eta$ over the epochs. Therefore, we used early stopping with a patience of 10, restoring the best weights and reducing the learning rate via the inverse decay function. The training was performed on an NVIDIA GeForce RTX 2080 Ti.

We use a leave-one-subject-out cross-validation for network performance evaluation investigating the accuracy, and the macro averaged F1 score, which combines the class-wise F1 scores. Therefore, we investigate the accuracy, F1 score, and training time per fold averaged over 14 folds.
Table 3. Detailed description of the convolutional neural network used in this study. The first convolutional together with the batch normalization layers are used once for the acceleration data and once for the gyroscope data.

<table>
<thead>
<tr>
<th>Layer Type</th>
<th>Hyperparameter</th>
<th>Output Shape</th>
<th># of Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D Convolution</td>
<td>Filter: 128, kernelsize: 8</td>
<td>(64, 128)</td>
<td>3200</td>
</tr>
<tr>
<td>Batch Normalization</td>
<td>Momentum: 0.9, epsilon: 0.001</td>
<td>(64, 128)</td>
<td>256</td>
</tr>
<tr>
<td>1D Convolution</td>
<td>Filter: 256, kernelsize: 5</td>
<td>(60, 128)</td>
<td>164,096</td>
</tr>
<tr>
<td>Batch Normalization</td>
<td>Momentum: 0.9, epsilon: 0.001</td>
<td>(60, 128)</td>
<td>240</td>
</tr>
<tr>
<td>1D Convolution</td>
<td>Filter: 128, kernelsize: 3</td>
<td>(58, 128)</td>
<td>98,432</td>
</tr>
<tr>
<td>Batch Normalization</td>
<td>Momentum: 0.9, epsilon: 0.001</td>
<td>(58, 128)</td>
<td>232</td>
</tr>
<tr>
<td>Global max-pooling</td>
<td></td>
<td>(128)</td>
<td>0</td>
</tr>
<tr>
<td>Dropout</td>
<td>dropout: 0.2</td>
<td>(64)</td>
<td>16,448</td>
</tr>
<tr>
<td>Fully Connected</td>
<td></td>
<td>(64)</td>
<td>0</td>
</tr>
<tr>
<td>Dense</td>
<td></td>
<td>(5)</td>
<td>325</td>
</tr>
</tbody>
</table>

3.5. Transfer Learning

The volleyball dataset obtained by Kautz et al. [8] served as the source domain in our transfer learning approach. Participants of both studies were equipped with a comparable sensor at the same position at their wrist. Kautz et al. [8] have already done the activity detection for the volleyball dataset, which means that our work is based on annotated samples of a fixed length that contain relevant actions from the domain. At first, the network (compare Figure 4) is trained on the volleyball dataset. Since this dataset lacks gyroscope data, only one input layer and subsequent convolutional, normalization, and pooling layers are used. During each run, the samples of 25 participants serve as training data, and the samples of two participants as validation data. The samples of three participants are used as test data. The weights of the convolutional layers from the best fold are saved and will be reused for subsequent transfer learning approaches.

The model is evaluated using a 3-fold-cross validation following the work of the original authors, which ensures good comparability. Note that Kautz et al. [8] used a slightly different architecture than what is used in this work.

The second step of the transfer learning approach is to train the actual network on the frisbee dataset. Therefore, the previously saved weights of the convolutional layers are loaded into the network. Depending on the exact configuration, those layers may be frozen, which means their weights will not be updated during the backpropagation phase. Due to the sharp peaks in gyroscope and acceleration data, we assume that the kernels for both network inputs are similar in the first convolutional layers. Therefore, we use the pre-trained weights for both acceleration and gyroscope data even though only the acceleration data was used for the training.

The remaining layers are initialized using Glorot’s initializer to combat the problem of vanishing gradients. Glorot et al. [38] showed that when going through the layers of a neural network, the variance in the layer’s output increases, and thus the gradients for the lower layers become very small when initializing the weights with a Gaussian normal distribution. Thus, the lower layers are only receiving minimal to almost no updates [39].

3.6. Experiments

The classification task on the frisbee dataset was conducted using both the detailed and aggregated labels. The first task features nine distinct classes (see Table 2). The latter one features five classes aggregating the different forehand and backhand throws. We then look at the performance of transfer learning, varying the number of pre-trained and frozen layers. Finally, we will resolve whether transfer learning can improve generalization capabilities when fewer training samples are available. This represents a smaller dataset due to a smaller study design. It might be cumbersome to acquire study participants in marginal sports, especially for individual sports. For comparability to the previous
experiments, we did not decrease the total number of participants used for training but instead omitted the data augmentation steps. In general, a decrease in performance is expected for smaller datasets.

4. Results

The following paragraphs present the results obtained from the experiments described in the previous chapter. We start with the results of our architecture on the volleyball dataset, since its performance influences the performance of the transfer learning experiments. We then present the results of the frisbee dataset and the transfer learning results on the network’s performance.

4.1. Volleyball Dataset

The leave-three-out cross-validation on the volleyball dataset resulted in a mean accuracy of 86.6% over all folds surpassing the accuracy of Kautz et al. [8] by 3.45 percentage points. The average F1 score is 68.7%.

4.2. Frisbee Dataset

The training from scratch on the frisbee dataset yields an overall accuracy of 66.6% and a macro averaged F1 score of 52.3%. Figure 5 shows the confusion matrix of the classification performance. The values are normalized along the true classes. The confusion matrix clearly shows block matrices for the different forehand (classes 0 to 2) and backhand (classes 3 to 5) throws.

Furthermore, some overheads (class 6) have been misclassified as one of the forehead throws (3 to 5).

If we drop the distinction between flat, outside-in, and inside-out throws and focus on the aggregated classes backhand and forehead, the leave-one-subject-out cross-validation results in a mean accuracy of 89.9% and a macro averaged F1 score of 88.4%. The corresponding confusion matrix is shown in Figure 6.

Figure 5. Confusion matrix of classification performance. The backhand throws are abbreviated as bh and the forehead throws as fh. The true class is given on the left and the predicted class on the bottom. Each entry is normalized according to its true class and denoted in percentage. Thus, the highest possible value is 100, and the smallest value is zero.
4.3. Transfer Learning

As described previously, transfer learning is realized by freezing or fine-tuning the weights of the first few layers after the initial training of the network on a source domain. Figure 7 shows the mean training time per fold, F1 score, and accuracy for different training configurations. The training time is measured, including balancing operations and the patience of early stopping. Using the pre-trained weights for all three convolutional layers, but without freezing these layers, the accuracy drops from 66.6% with randomly initialized weights to 66.2%. In contrast, the F1 score slightly improves to 52.9%. The training time per fold decreases from 217.6 s to 198.6 s.

An increased number of frozen layers leads to a decrease in both accuracy and F1 score. While the training time per fold is only slightly shorter for one layer frozen (205.2 s) for the two layers initialized and frozen, the training time is considerably shorter (119.7 s).

4.4. Transfer Learning with Smaller Dataset

Looking at the transfer learning performance using the reduced training dataset, we generally see a drop in performance compared to the network trained on the entire dataset, i.e., with all data augmentations. Figure 8 shows the mean training time per fold, F1 score, and accuracy for this training scenario. With randomly initialized weights, the network achieves an accuracy of 52.0% and an F1 score of 31.8%, which takes 8.6 s per fold. Using the pre-trained weights for all three convolutional layers and not freezing them, the accuracy increases to 60.5% and an F1 score of 36.8% while also increasing the training time to 9.8 s. Having one or two layers initialized and frozen increases accuracy and F1 score compared to the random initialization. The performance with random initialization of weights is slightly worse than without freezing the layers. However, it leads to shorter training times.
Figure 7. Comparison of the accuracy, F1 score, and training time for different transfer learning configurations. The scale for the accuracy and F1 score is shown on the left y-axis, and for the training time on the right y-axis. In the configuration of the pre-trained weight, all three convolutional layers were initialized with the weights of the network trained on volleyball data but not frozen. For the random weights configuration, all layers were randomly initialized. For the configurations with one or two frozen layers, the model is initialized with pre-trained weights, and the weights are frozen in the later training procedure.

Figure 8. Comparison of the accuracy, F1 score, and training time for different transfer learning configurations for a reduced dataset, i.e., without data augmentation. The scale for the accuracy and F1 score is shown on the left y-axis and for the training time on the right y-axis. In the configuration of the pre-trained weight, all three convolutional layers were initialized, with the weights of the network trained on volleyball data but not frozen. For the random weights configuration, all layers were randomly initialized. For the configurations with one or two frozen layers, the model is initialized with pre-trained weights, and the weights are frozen in the later training procedure.
5. Discussion

The goal of this work was to create the first HAR system for different throws in Ultimate Frisbee using a convolutional neural network and study the effects of transfer learning from a beach volleyball action recognition task.

5.1. Study

The data recording took place at the end of March 2021, where the participants who practiced regularly multiple times a week were not allowed to do so for roughly five months due to COVID-19 restrictions in Germany. Even though most of the players met privately to practice throws, the intensity and complexity of the throws increases during competitive matches. The players were aware of this issue since they reported that the lack of practice negatively influenced their throwing capabilities. This decline in capabilities is not only a perceived effect as Korkmaz et al. [40] have shown. Their study conducted with amateur football players found that the months-long detraining process during the pandemic led to a significant deterioration in players’ physical and motoric abilities.

The data have been recorded with only two players at a time in a controlled environment. Future work has to investigate whether the proposed pipeline works under competitive conditions during a match. Stoeve et al. [41] investigated the transferability of a football activity recognition pipeline from controlled conditions to real-world scenarios. They observed a decrease in performance for a feature-based approach, whereas performance was comparable for the proposed CNN, thus indicating good generalization to complex scenarios [41].

5.2. Activity Recognition

The network achieved an accuracy of 89.9% in the 5 class problem without favoring the majority class, which supports the evidence that CNNs work well for time series classification. To put these numbers into perspective, Anand et al.’s CNN for stroke classification reached accuracies between 77.2% (badminton) and 93.8% (tennis) [17].

The decreased performance for the 9 class problem shows that separating flat, outside-in and inside-out throws is a tough challenge. The confusion matrix (Figure 5) reveals block matrices for the different forehand and backhand classes. Inside these sub-matrices, the error rate of the classifier was very high, the network did not predict reliable results, and one could argue that the classifier can not distinguish the detailed classes. Due to better applicability in a real-world scenario using smartwatches or fitness trackers, we fixed the IMUs at the wrist of the dominant hand. However, future work should consider placing the sensor at the back of the hand since the subtle differences in movement between performing flat, outside-in, and inside-out throws originate mainly from the wrist’s movement. A sensor placed above the wrist has disadvantages in recording these small changes and, therefore, worsens the classifier’s performance. Since many frisbee players wear gloves for better friction, a sensor could be fixed on the glove.

We also tested the classification using other architectures, including a ResNet architecture [42]. The tested networks achieved lower or, at most, comparable accuracies. The disadvantage of the architectures achieving the same accuracy as, e.g., the ResNet architecture is the much longer training time. In our case, using only three ResNet blocks already doubles the training time.

5.3. Transfer Learning

With the transfer learning approach, the accuracy and F1 score could not be improved. The reason for this is probably, as already mentioned, that the minor differences between the detailed forehand and backhand throws are not represented in the IMU data. Nevertheless, transfer learning decreases the training time while maintaining similar performance results. Therefore, depending on the individual application and focus, it might be beneficial to use transfer learning to achieve faster training times.
One drawback of the employed source dataset is the missing gyroscope data. We used the weights of the convolutional layers pre-trained on the acceleration of the volleyball data as initialization for the layers of the frisbee gyroscope data. Since we achieve comparable results, our assumption that both acceleration and gyroscope have similar low-level features stands. While the focus on only acceleration data may have been sufficient for classifying beach volleyball actions, investigations including multiple architectures revealed the necessity of including gyroscope data for frisbee activity recognition. Therefore, disregarding the gyroscope data has never been an option in our research.

5.4. Transfer Learning with Smaller Dataset

As expected, reducing the dataset leads to a general drop in performance in all training configurations. However, transfer learning improves both accuracy and F1 score in all tested training configurations. The training time slightly increases or decreases depending on the exact transfer learning setup. Overall, the results show that transfer learning can improve the network’s generalization capabilities when using small datasets.

The performance improvement through transfer learning was not the case for the entire dataset. In general, transfer learning may improve the performance depending on the specific dataset and training configuration. Therefore, it is beneficial to test transfer learning and data augmentation techniques to improve a network’s generalization capabilities.

Marginal sports in particular can benefit from a transfer learning approach, since acquiring study participants is cumbersome. This might also make activity recognition in marginal sports more interesting for smartwatches or fitness trackers, since the amount of data to build a solid activity recognition system is relatively small. Using activity recognition in marginal sports also increases professionalism and may thus lead to better performances in marginal sports.

6. Conclusions

The goal of this work was to create the first HAR system, which can classify different throwing techniques of Ultimate Frisbee players. Therefore we recorded 14 participants who performed seven different throws using a wrist-worn accelerometer and gyroscope. In order to establish a ground truth, the signal was manually annotated using a video reference. Following the action detection, a CNN was used to classify the throws and catches, and assign the false positives of the action detection to a null class. A leave-one-subject-out cross-validation resulted in a mean accuracy of 66.6%. However, the macro averaged F1 score of 52.3% is more meaningful for the imbalanced dataset. We found that separating flat, outside-in and inside-out throws is a tough challenge, and the network could not solve this task reliably. If we drop the distinction between these detailed classes and focus on the main throwing techniques and catches, the CNN reaches a macro averaged F1 score of 88.4%.

Furthermore, we studied the effects of transfer learning, using the beach volleyball dataset by Kautz et al. [8] as a source domain. We used the weights of the lower convolutional layers of a network pre-trained on the volleyball data. With all of the three convolutional layers pre-trained but not frozen, the CNN reached a comparable performance concerning the accuracy and F1 score, but achieved faster training times. Freezing more layers resulted in even shorter training times but had a slight negative impact on the prediction capabilities of the network.

Another question was whether transfer learning could improve classification tasks if only very few data samples were available. Therefore, we dropped the data augmentation steps to simulate a smaller dataset. While the smaller training dataset led to a general decrease in accuracy, the transfer learning approach improved both accuracy and F1 score. The training time slightly increases or decreases with transfer learning depending on the training configuration.

To conclude, transfer learning can be a powerful tool to combat long training times for HAR. Despite this, it should be considered in addition to data augmentation techniques to
improve the generalization capabilities, especially when dealing with small datasets, which is often the case in marginal sports.

**Author Contributions:** Conceptualization, J.L., T.P. and M.S.; methodology J.L., T.P. and M.S.; software, T.P.; validation, J.L., T.P. and M.S.; formal analysis J.L., T.P. and M.S.; investigation J.L., T.P. and M.S.; resources B.M.E.; data curation, J.L., T.P. and M.S.; writing—original draft preparation, J.L., T.P. and M.S.; writing—review and editing, J.L., T.P., M.S. and B.M.E.; visualization, J.L., T.P. and M.S.; supervision, B.M.E.; project administration, B.M.E.; funding acquisition, B.M.E. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by German Research Foundation (DFG) grant number Heisenberg: 434/8-1.

**Institutional Review Board Statement:** The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of the Friedrich-Alexander-Universität Erlangen-Nürnberg (number 106_13B).

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The data is available from the authors on reasonable request.

**Acknowledgments:** Bjoern Eskofier gratefully acknowledges the support of the German Research Foundation (DFG) within the framework of the Heisenberg professorship programme (grant number ES 434/8-1). Furthermore, we would like to thank Unwucht Erlangen for the help acquiring the participants of the study and Jonas Muschiol (www.jonas-muschiol.de/, accessed on 28 February 2022) for the graphic of frisbee throwing techniques.

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

**Abbreviations**

The following abbreviations are used in this manuscript:

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>Convolutional Neural Network</td>
</tr>
<tr>
<td>ENN</td>
<td>edited nearest neighbor</td>
</tr>
<tr>
<td>HAR</td>
<td>human activity recognition</td>
</tr>
<tr>
<td>IMU</td>
<td>inertial measurement unit</td>
</tr>
<tr>
<td>LSTM</td>
<td>long short-term memory</td>
</tr>
<tr>
<td>ReLU</td>
<td>rectified linear unit</td>
</tr>
<tr>
<td>SMOTE</td>
<td>synthetic minority over-sampling technique</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
</tbody>
</table>

**References**

Bibliography


Bibliography


**Own publications referring to this work**

