# Prediction of football actions and identification of optimal sensor placements using a semi-supervised learning approach

Master Thesis

Aske Gye Larsen and Giovanni Papi

Aalborg University
Sports technology

**Sports Technology**
Aalborg University

**AALBORG UNIVERSITY**
STUDENT REPORT

**Title:**
Prediction of football actions and identification of optimal sensor placements using a semi-supervised learning approach

**Theme:**
Application of Scientific Methods in Sports Technology

**Project Period:**
Spring Semester 2023

**Project Group:**
10303

**Participants:**
Aske Gye Larsen
Giovanni Papi

**Supervisor:**
Anderson de Souza Castelo Oliviera

**Copies:** 1

**Page Numbers:** 28

**Date of Completion:**
May 31, 2023

**Abstract:**

The aim of this study was to combine the use of IMUs in football and the principles commonly used in Human Activity Recognition, to be able to predict some of the most common football actions, and herein also which placements of the sensors are the most important. This process was split into two parts, starting with predicting the football actions, i.e. pass, dribbling, first touch, and positioning, using a bidirectional Long Short Term Memory neural network (LSTM). The second part consisted of predicting head scans with a Deep Learning Artificial Neural Network (DNN) separately, as the head scan happened simultaneously with the other actions. 14 male and 3 female football players participated in the study. Prior to any predictions, the data was split 50/50 into labeled and unlabeled data, and the labeled data was further split 80/20 into training data and testing data. All data were normalized and balanced by using Adaptive Synthetic Sampling Approach (ADASYN). 5250 statistical time domain features were calculated over a sliding window of 200 ms, with 50 % overlap, but later reduced with a Principal Component Analysis retaining >95 % of the variance. A semi-supervised uncertainty-aware pseudo-labeling technique was used to decrease the time needed for labeling. The LSTM showed decent results for predicting football actions, with a cross-validation score of 0.74 and an F1-score of 0.74. The DNN prediction of head scans showed overall slightly better results, mainly due to the lower number of classes, with a cross-validation score of 0.79 and an F1-score of 0.78. The sensor placement that supplied the most relevant information to the LSTM was the one placed on the right calf with an F1-score of 0.65. For the DNN the most important sensor placement was the one placed on the head, which showed an F1-score of 0.69.

Aalborg University, May 31, 2023

Aske Gye Larsen
asklar21@student.aau.dk

Giovanni Papi
gpapi21@student.aau.dk

# Prediction of football actions and identification of optimal sensor placements using a semi-supervised learning approach

**Aske Gye Larsen & Giovanni Papi**

**Abstract**

The aim of this study was to combine the use of IMUs in football and the principles commonly used in Human Activity Recognition, to be able to predict some of the most common football actions, and herein also which placements of the sensors are the most important. This process was split into two parts, starting with predicting the football actions, i.e. pass, dribbling, first touch, and positioning, using a bidirectional Long Short Term Memory neural network (LSTM). The second part consisted of predicting head scans with a Deep Learning Artificial Neural Network (DNN) separately, as the head scan happened simultaneously with the other actions. 14 male and 3 female football players participated in the study. Prior to any predictions, the data was split 50/50 into labeled and unlabeled data, and the labeled data was further split 80/20 into training data and testing data. All data were normalized and balanced by using Adaptive Synthetic Sampling Approach (ADASYN). 5250 statistical time domain features were calculated over a sliding window of 200 ms, with 50 % overlap, but later reduced with a Principal Component Analysis retaining >95 % of the variance. A semi-supervised uncertainty-aware pseudo-labeling technique was used to decrease the time needed for labeling. The LSTM showed decent results for predicting football actions, with a cross-validation score of 0.74 and an F1-score of 0.74. The DNN prediction of head scans showed overall slightly better results, mainly due to the lower number of classes, with a cross-validation score of 0.79 and an F1-score of 0.78. The sensor placement that supplied the most relevant information to the LSTM was the one placed on the right calf with an F1-score of 0.65. For the DNN the most important sensor placement was the one placed on the head, which showed an F1-score of 0.69.

## Introduction

The use of technology has seen a steep and continuous rise within the world of sports, both for training and in competition. This has lately been boosted even further by the rise of Artificial Intelligence (AI), which is used with most sports tracking devices. More specifically in football (soccer), it is nowadays common for players to practice or play in competitions wearing vests with a tracking device below their shirts. The tracking devices normally include a GPS and an Inertial Measurement Unit (IMU), which is a sensor that typically utilizes the combination of an accelerometer, gyroscope, and magnetometer to measure the movements of the agents in 3D. The combination of tracking devices and AI have the ability to accurately predict future risk factors for injuries and have therefore been helpful in keeping athletes healthy by decreasing the risk of overtraining, or by detecting movement patterns that are similar to athletes who possess a greater risk of injury (Richter et al. 2019, Franklyn-Miller et al. 2017, Rommers et al. 2020). Moreover, in the competitive nature of sports, every involved agent is seeking to achieve advantages over competitors in one way or another, which opens the door for AI to be an important way for ath-

letes, clubs, trainers, etc. to gain that competitive edge by using it to increase performance. One way to increase performance is by supplying the clubs opportunities to perform spatial temporal analysis on individual players, or on a team-level basis (Gudmundsson & Horton 2017, Aughey 2011, Barron et al. 2014). However, most tracking devices are placed on the torso of the player, as this is an area that rarely interferes with the ball or other players, making these the only ones having been approved by FIFA to be used on a professional level. However, in recent years multiple companies have applied for FIFA's approval of their hardware, which often are located on the players' calves and just beneath the lateral ankles. Each sensor placement supplies different information, and sensors placed on the lower body are today used to predict the football actions done by the players. These actions include passing, first touches, dribbling, etc. This can give valuable information to the coaches and players, as it is a way to quantify workload, how fast a player can perform actions with the ball etc. Therefore, it would be beneficial for the players and coaches to have a system that recognizes these actions automatically, as the quality of feedback to the players would increase if it is backed up and helped by advanced data.

The recognition of movements within sports is a branch of human activity recognition (HAR), which is an area that has captivated researchers in recent years. The rise of the interest has especially been noteworthy since the breakthrough of wearables such as smartwatches and smartphones, where the extraction of the data from the accelerometers has been a key development since it enables tracking of the persons movements without invasive maneuvers (Bayat et al. 2014). However, most literature on HAR solely or primarily classify activities of daily living (ADL), which usually consist of continuous movements, such as walking, sitting, lying, etc. Some have tried to automatically recognize specific sports actions using video recordings (Tsunoda et al. 2017, Wu et al. 2022, Xing & Li 2022), however, within sports, video-based movement recognition has the disadvantage that they are costly and not easily transportable to new locations (Cuperman et al. 2022), principally making them unavailable for smaller clubs or individuals. Another issue with predicting sports movements is that they are often categorized as being discrete, which is opposite to the continuous movements seen in ADL. This means that an athlete can perform many actions within a short amount of time. This can be an issue, as usually classification algorithms are based on supervised learning, meaning all actions needs to be labeled. This is especially time-consuming when using temporal data, as each frame usually needs to be labeled (Berthelot et al. 2019). One way to avoid this problem is to move away from a supervised learning approach, to a more time and labor-beneficial semi-supervised learning (SSL) approach (Rizve et al. 2021, Zhou 2018, Zhang et al. 2021). The use of SSL within HAR has also been investigated and has often shown promising results in applications where solely manual labeling is not feasible. (Oh et al. 2021, Bi et al. 2022, Singh et al. 2021, Rizve et al. 2021). Furthermore, as the focus of this study is the classification of football actions and many of these are discrete, the use of SSL is presumed to be a tenable solution.

However, literature on the use of specifically IMUs for recognition purposes in football is scarce. Cuperman et al. and Stoeve et al. have investigated the feasibility of classifying actions in football. They mainly focus on the easier distinguishable movements such as jumping, running, passing, shooting, etc. (Cuperman et al. 2022, Stoeve et al. 2021). With the exception of passing and shooting, these

are generally movements that are already often recognized within the field of HAR, and while running or jogging provide valuable information on the workload of the players involved, these movements do not provide the in-depth analysis of football actions that a player or trainer might want. Movements, such as dribbling, passing and first touch are generally harder to distinguish from each other, since the movements required see a lot of overlap. Nevertheless, these football actions provide valuable information on a player's technical ability. Moreover, most of these studies include multiple IMUs, which usually is not feasible for use in the field. Previously, the optimal sensor placements have been investigated within HAR by Rahn et al. and Xia & Sugiua 2021, and subsequently in Hockey by Shahar et al., who focused on sensors on the upper body (Rahn et al. 2021, Shahar et al. 2020, Xia & Sugiura 2021). However, no studies to date have investigated the optimal sensor placements for automatically classifying football actions.

Therefore, this study aims to combine the use of IMUs in football and the principles commonly used in HAR, to be able to predict some of the most common football actions, and herein also which placements of the sensors provide the most important information, using an SSL approach.

# Methods

## Participants

14 male and 3 female football players participated in the study (19.1 ± 6.9 yr, 174.5 ± 10.3 cm, 62.4 ± 23.1 kg, 9.7 ± 3.6 yr of football experience). The participants primarily played for teams in the best Danish senior female football league or the best youth leagues. The participants wore their own football boots and weather-appropriate clothing.

## Technical design

In this study, the state-of-the-art full-body motion capture system Xsens MVN Awinda (Xsens Technologies B.V, Enschede, the Netherlands) was used. This system consists of 17 high-performance IMUs and motion was captured using the state-of-the-art sensor fusion Kalman filter (XKF3hm), supplied by the included software, Xsens MVN Studios (version 2021.0.1). The participants were fitted with all 17 available sensors, as the software Xsens MVN Studios required, but only 7 sensors, shown in figure 1, were further used for the data analysis. These were chosen to represent the common placement of wearables used in soccer by the major player tracking companies and included sensors on the head, sternum, right wrist, right calf, right foot, left calf and left foot. Three-axis accelerometer (ACC) output (3D ± 16g), three-axis gyroscope (AR) output (±2000○/s), and three-axis magnetometer (MAG) output (± 1.9 Gauss) respectively were measured with the IMUs and four-axis quaternion (QUA) was calculated, all at 60Hz to develop a machine learning (ML) algorithm to detect football actions. The video was recorded at 1080p / 60fps (Nikon 1 J4 Model Camera with a 1 NIKKOR VR 10-30mm f/3.5-5.6 PD-ZOOM lense) and used to label each event as the ground truth for the supervised ML algorithm to train and test. The cameras were placed approximately 5 meters from the playing field, with an angle of roughly 15° relative to the ground, to make the whole exercise area visible during the recording.

## Practical design

A self-paced warm-up was instructed to the participants to minimize the risk of injury and to ensure that the participants had a full range of motion, which is also to be expected during a live game or practice session. Then the participants' body dimensions were measured

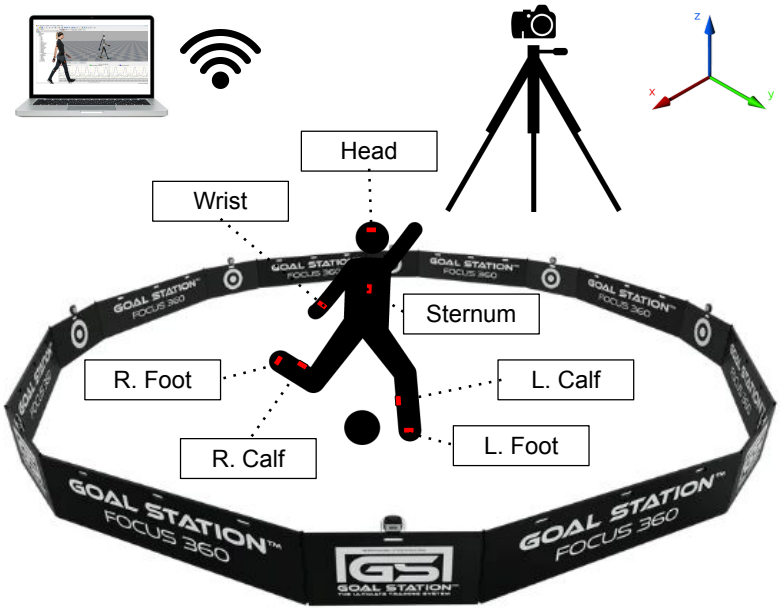


**Figure 1:** The experimental setup illustrating the computer with the Xsens MVN Studios and a WIFI connection to the Xsens MVN Awinda suit mounted on the participant as illustrated by the 7 used sensors. Furthermore, the camera is illustrated pointing towards the training area surrounded by the Goal Station Focus 360.

along with being fitted with the 17 IMUs. Both the measurements and the fitting of the IMUs were executed following the guidelines from Xsens MVN Awinda (Myn et al. 2021). After being fit with the IMUs, a calibration within the Xsens MVN Studios was completed. The calibration involved the participant standing in a neutral pose (N-pose), walking straight for 4 seconds with their natural gait, before turning around and walking back to the start position, and ending the calibration in a N-pose. Only calibrations defined as good by the Xsens MVN Studios software were accepted for this study. After the calibration of the IMUs, the participants completed a training session as they would usually do. The exercises were chosen to fit a large variety of common football actions, i.e. numerous first touches, passes, dribbles, head scans, and positioning. All exercises were specialized by the coach to represent incidents the participant could expect given their playing position. The common target for the participants in all exercises using the Goal Station Focus 360 is to hit 12 walls (1.40 m x 0.8 m), that light up in a particular color, as fast as possible, each of these exercises lasting approximately 30-50 seconds. These types of exercises were chosen to simulate external output the participant had to react to, as movement patterns have demonstrated being different in reactive movements contra planned movements, the latter not being realistic in game-like situations (Wakatsuki & Yamada 2020).

In some exercises, the next wall that is to be hit lights up in a different color, which further incentives the participant to keep scanning the field During the data gathering period, the coach chose a total of 10 different exercises, however, the main goal for the participants of orientating themselves and hitting the next lit-up board as fast as possible

never changed. The differences in the exercises were mainly the locations of the boards. To ensure synchronization between the IMUs and cameras, interrupting periods were implemented at the beginning and end of each recording. During the interruptions, the participants were instructed to stop all movements for 5 s, then do three quick jumps and subsequently stand still for 5 s. These periods were easily detectable on IMU recordings, which enabled synchronization of the video and the IMU recordings.

## Data Processing

The data used in this study were sampled from 186 complete trials. 10 trials were omitted due to erroneous data, such as sensor fallout or interference with the video recordings. The raw data from the IMUs were hereafter exported to .xlsx files. The IMU and video recordings were thoroughly gone through frame by frame, and subsequently labeled manually in Excel by what action the participant was performing at a given frame. Furthermore, each file was given a unique participant as well as an occurrence ID. The actions involved were as follows: a pass, positioning, first touch, dribbling, head scanning the field (head scan), and not head scanning of the field (not head scan). Definitions of each action are given in Table 1. The labeling of head scan was done independently of the other labels, meaning that a given frame could be labeled pass, positioning, first touch, or dribbling as well as head scan or not head scan. The labels are illustrated in figure 2 as the mean values of all football actions used in this study, normalized between 0 - 100 %, for each of the outputs from all IMUs. All calculations and code were written using the Python programming language (version 3.11.2), in the open-source scientific environment Spyder (version 5.3.3). The data was split (50%/50%) into data that was labeled for the supervised ML algorithm and data kept unlabeled for the pseudo-labeling. Furthermore, the labeled data was split (80%/20%) into a training and test set and additionally, one-hot encoding was implemented as a method to convert the categorical data to numerical, as to ensure that higher numbered classes did not carry more weight (Potdar et al. 2017).

**Table 1:** Each football action included and definitions on when the action begins and ends.

| Action | Defintion |
|---|---|
| Pass | Begins with the passing foot leaving the ground, and ends when the passing foot makes contact with the ground again |
| First touch | Begins when the foot doing the first touch leaves the ground, and ends when the foot makes contact with the ground again |
| Dribble | Begins when the foot doing the dribble leaves the ground, and ends when the foot makes contact with the ground again |
| Positioning | Begins the participant is not in possession of the ball, and is therefore moving around to get in a better position, and ends when the participant is lifting a foot to prepare for a first touch or a pass |
| Head scan | Head scan denotes an active head movement where a participant's face is temporary directed away from the ball to gather information in preparation for subsequently engaging with the ball (Jordet et al. 2020). |

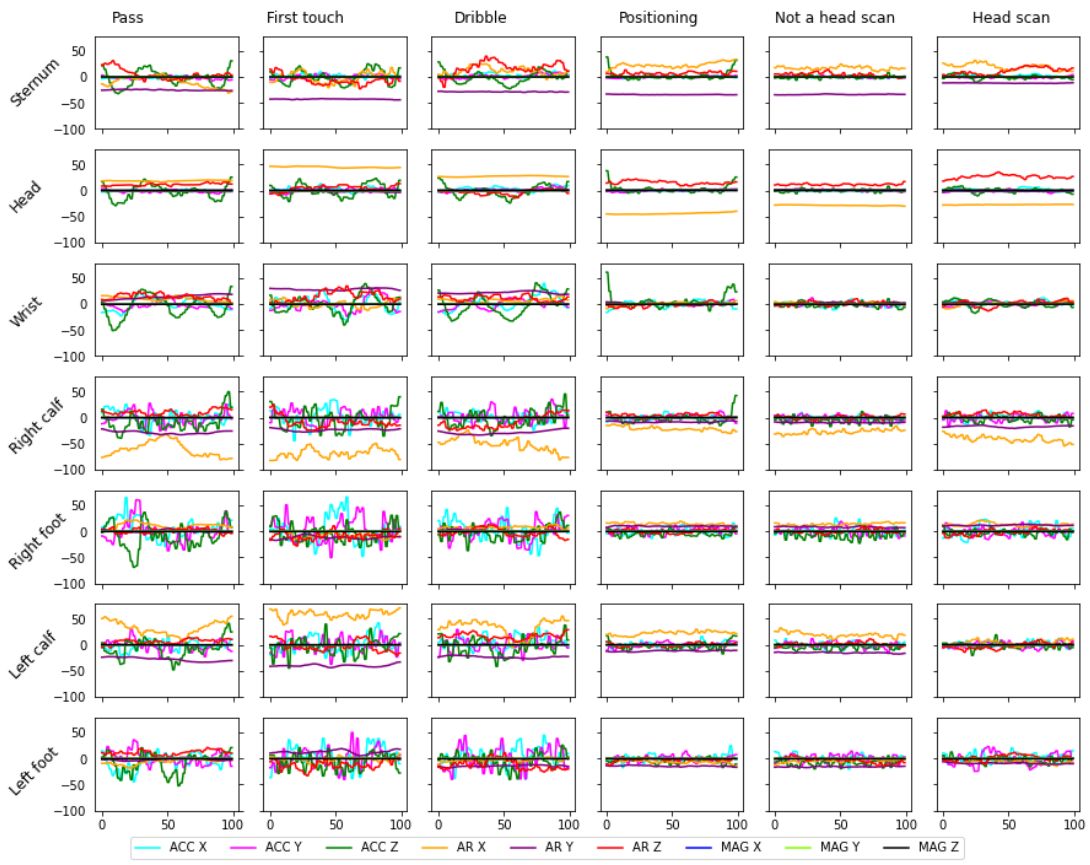


**Figure 2:** Shows the 9 sensor outputs (3D accelerations, 3D angular rate, and 3D Gauss) measured by the 7 IMUs (Sternum, Head, Wrist, Right calf, Right foot, Left calf, and Left foot) for the four football actions as well as the binary classification detecting head scans, all movements normalized between 0 to 100 %.

## Balancing the data

To overcome any imbalances in the data sets, which is expected due to the low interference within the training, an adaptive synthetic sampling approach (ADASYN) was used (He et al. 2008). ADASYN is an oversampling approach, which artificially generates more data points from the minority classes, and puts more weight on those minority classes that are harder to learn, compared to the majority class that are easier to learn (Santos et al. 2018).

## Feature Extraction

When handling signal-type temporal data, manual feature extraction is a common technique to improve the performance of the algorithm (Preece et al. 2008, Cust et al. 2019, Kanjilal & Uysal 2021). This is why the derivatives were calculated from the raw data, as Hamäläinen et al. showed that the jerk-type features are especially productive in developing a robust algorithm (Hamäläinen et al. 2011). From both the raw data and the derivatives, the magnitude was calculated as the square root of the squared sum of the 3-axis (4-axis for quaternion). After constructing each data vector, a rolling window of 12 frames with 50 % overlap was used. The 12 frames were chosen as Jaén-Vargas et al. demonstrated that moderate to large windows yielded the highest-performing algorithms, and suggested a 1/5 of the Sampling Frequency (60 Hz) is the minimum size window for optimal performance (Jaén-Vargas et al. 2022). The 50 % overlap that was chosen as the feature extraction in this project was mainly based upon the features proposed in Zhu et al. (Zhu et al. 2017) where they used 67 % overlap, however, Dehghani et al. showed no increase in performance with higher overlap (Dehghani et al. 2019), which is why the

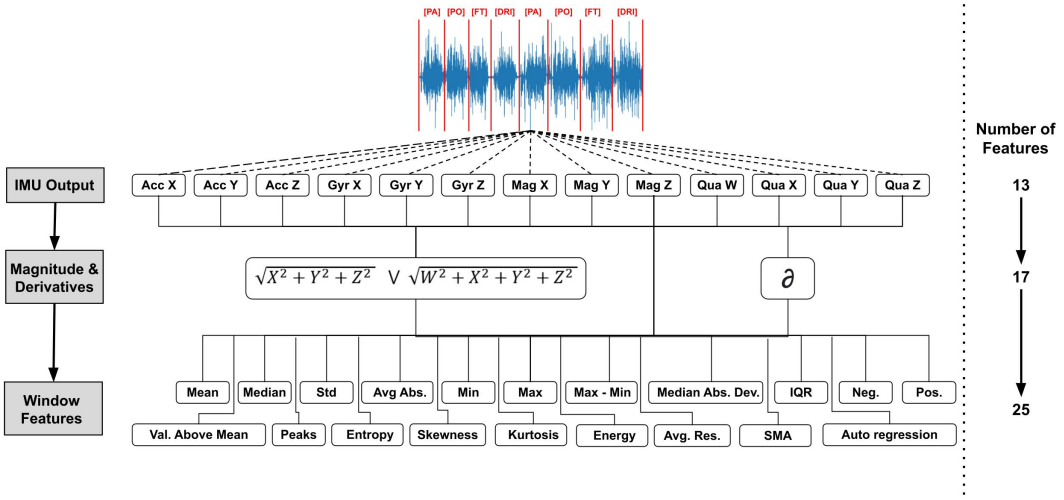


**Figure 3:** The Feature Extraction with the 13 raw data output being X, Y, and Z axis of the accelerometer, gyroscope, and magnetometer as well as the W, X, Y, and Z axis quaternion output. The derivatives and magnitudes of the raw output were used to calculate the final features, being mean, median, standard deviation, average absolute difference, minimum, maximum, difference between minimum and maximum, median absolute deviation, interquartile range, 10th and 90th percentile, negative count, positive count, values above the mean, peaks, entropy, skewness, kurtosis, energy, average resultant acceleration, signal magnitude area and autoregression coefficients with Burg order equal to four correlation coefficients between two signals respectively.

slightly lower 50 % was chosen. Furthermore, Zhu et al. also demonstrated that keeping the data from both the accelerometer, gyroscope, and magnetometer as well as the quaternion output is the most optimal, and they also suggest that the features should be kept in the time domain (Zhu et al. 2017). Over each window the following features were calculated; mean, median standard deviation (STD), average absolute difference, minimum, maximum, difference between minimum and maximum, median absolute deviation, 10th and 90th percentile, interquartile range (IQR), negative count, positive count, values above the mean, peaks, entropy, skewness, kurtosis, energy, average resultant acceleration, signal magnitude area (SMA) and autoregression coefficients with Burg order equal to four correlation coefficients between two signals, illustrated in figure 3. This gives a total of 750 features from each of the 7 IMUs and thereby 5250 features for each window. All features were normalized between 0 and 1 to ensure that larger values won't carry more weight than smaller values.

## Development of the algorithm

To find the best performing algorithm, the following ML algorithms were tested: random forest, XGBoost, gaussian naive bayes, manhattan distance k-nearest-neighbor, multinomial logistic regression, support vector machine with radial basis function kernel, deep learning artificial neural network (DNN), bidirectional long short-term memory (biLSTM) neural network, and finally a multilayered 2D-convolutional neural network (CNN). These were chosen as they are some of the most commonly used algorithms for HAR (Narayanan et al. 2020, Cust et al. 2019, Herold et al. 2019, Kwapisz et al. 2011, Ariza-Colpas et al. 2022, Dehghani et al. 2019, Garcia-Gonzalez et al. 2020, Jaén-Vargas et al. 2022, Kanjilal & Uysal 2021, Khanal et al. 2021, Nunavath et al. 2021, Preece et al. 2008, Prasad et al. 2021, Stoeve et al. 2021, Walse et al. 2016). All models were tested, as illustrated in table 2, on the following metrics: accuracy, F1-score, precision, recall, Krippendorf's alpha for the multiclass labels, Cohen's Kappa for the binary labels, and finally, five-fold cross-

**Table 2:** Showing the performance metrics (accuracy, precision, recall, F1-score, Krippendorf's Alpha/Cohen's Kappa and a five fold cross-validation) for each of the algorithms tested for both the prediction of football actions and head turns.

| Football action algorithm | Accuracy | Precision | Recall | F1-score | Krippendorf's Alpha | Cross-Validation |
|---|---|---|---|---|---|---|
| Random Forrest | 0.68 | 0.56 | 0.68 | 0.62 | 0.34 | 0.60 |
| Artificial Neural Network | 0.73 | 0.70 | 0.73 | 0.71 | 0.54 | 0.77 |
| XGBoost | 0.58 | 0.62 | 0.58 | 0.59 | 0.35 | 0.66 |
| Gaussian Naive Bayes | 0.48 | 0.59 | 0.38 | 0.48 | 0.35 | 0.35 |
| K-Nearest-Neighbour | 0.58 | 0.57 | 0.58 | 0.57 | 0.43 | 0.54 |
| Multinomial Logistic Regression | 0.53 | 0.59 | 0.53 | 0.43 | 0.03 | 0.62 |
| Support Vector Machine | 0.58 | 0.60 | 0.58 | 0.59 | 0.42 | 0.62 |
| Multilayered 2D Convolutional Neural Network | 0.70 | 0.69 | 0.70 | 0.69 | 0.43 | 0.64 |
| Long Short-Term Memory Neural Network | 0.75 | 0.73 | 0.75 | 0.74 | 0.55 | 0.74 |
| Head scan algorithm | | | | | Cohen's Kappa | |
| Random Forrest | 0.70 | 0.72 | 0.70 | 0.70 | 0.23 | 0.72 |
| Artificial Neural Network | 0.78 | 0.84 | 0.76 | 0.78 | 0.61 | 0.79 |
| XGBoost | 0.67 | 0.66 | 0.67 | 0.66 | 0.48 | 0.63 |
| Gaussian Naive Bayes | 0.53 | 0.64 | 0.43 | 0.53 | 0.40 | 0.40 |
| K-Nearest-Neighbour | 0.60 | 0.78 | 0.50 | 0.58 | 0.23 | 0.58 |
| Multinomial Logistic Regression | 0.63 | 0.67 | 0.65 | 0.65 | 0.33 | 0.67 |
| Support Vector Machine | 0.71 | 0.73 | 0.71 | 0.71 | 0.33 | 0.71 |
| Multilayered 2D Convolutional Neural Network | 0.73 | 0.69 | 0.72 | 0.70 | 0.47 | 0.69 |
| Long Short-Term Memory Neural Network | 0.74 | 0.81 | 0.71 | 0.74 | 0.54 | 0.78 |

validation. The best performance was defined as the algorithm that produced the highest F1-score. Furthermore, all of these models underwent hyperparameter tuning using a systematic grid search from the scikit-learn library, to make sure the most optimal version of each model was tested. The best-performing model for classifying the 4 different football actions was the deep learning biLSTM (LSTM) with an F1-score of 0.74 and the best-performing model for the classification of head scans was the DNN with an F1-score of 0.78.

## Feature Reduction

One of the issues with a vast amount of data, comprehensive feature extraction, and deep neural networks is the computational power required. Therefore, the feature importance of each feature was calculated for both the LSTM and DNN through the TensorFlow (2.12.0) library, to ensure no redundant features were used. However, no clear outliers were detected, but some features outperformed others and the five most important features are illustrated in Table 3. Another way to reduce the dimensionality of the data is the state-of-the-art method of Principal Component Analysis (PCA). PCA separates itself by not looking at the feature importance but by investigating the variance of the data. PCA is a technique that transforms high-dimensionality data into a lower-dimensionality while retaining as much information as possible. The PCA does this by finding the directions of maximum variance in high-dimensional data and projecting it onto a new subspace with fewer dimensions (Abdi & Williams 2010). PCA has been widely used in HAR, where the main advantage is the reduction in required computational power, while still maintaining a high accuracy (Aljarrah & Ali 2019, 2021, Fergani et al. 2013, Walse et al. 2016, Ariza-Colpas et al. 2022). For some applications, PCA has even improved the results, mainly by reducing the noise (Chen et al. 2017, Zhao et al. 2022). In this study, the first 948 principal components were used out of the 5250 features, as they were able to explain >95% of the variance, which is in the common range for PCA in HAR (80 - 95 %) (Aljarrah & Ali 2019, 2021, Fergani et al. 2013, Walse et al. 2016, Ariza-Colpas et al. 2022). The PCA reduced the computational time needed by 56 and 47 % from the LSTM (3 hours, 16 minutes) and DNN (2 hours, 27 minutes) respectively, while not lowering the F1-score, as illustrated in figure 4.

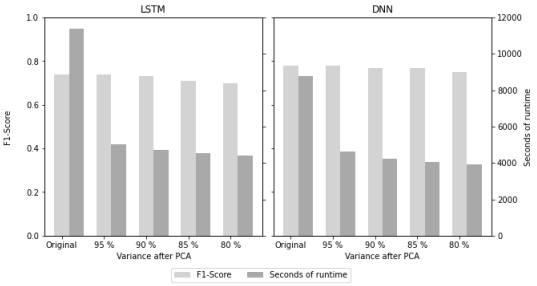


**Figure 4:** Shows the F1-score and computational time of both the LSTM and DNN run with the original dataset as well as different amounts of variance kept after the PCA. Both were run mainly on the GPU (AMD Radeon RX Vega 7), with fewer smaller computations on the CPU (AMD Ryzen 7 4700U).

## Semi-supervised learning

In SSL, portions of the data are kept unlabeled and different methods are then used to predict the unknown labels. One of the most used methods is pseudo-labeling, where you train an algorithm on the labeled data, use the algorithm to predict the class for the unlabeled data, and then train another algorithm on the combined dataset of the manually labeled data and the predicted labels (Thapa et al. 2023, Fu et al. 2021, Xu et al. 2022). There is no clear consensus on the split between labeled and unlabeled data, however, Xu et al. demonstrated that larger proportions of labeled data yields higher perfor-

**Table 3:** Showing the feature importance, mean, and standard deviation (std) of the five most important features with normalization, for both the LSTM and the DNN, followed by which sensor and its output (MAD = Mean absolute deviation).

| LSTM | IQR sternum acceleration (Z) | Entropy sternum acceleration (Z) | Entropy right calf Gauss (Y) | IQR right calf (Y) | Median SMA right foot jerk (X) |
|---|---|---|---|---|---|
| Feature Importance | 1.60 e-8 | 1.54 e-8 | 1.33 e-8 | 1.27 e-8 | 9.87 e-9 |
| Pass | $0.09 \pm 0.05$ | $2.06 \pm 0.90$ | $3.12 \pm 0.29$ | $0.22 \pm 0.23$ | $4.99 \pm 3.21$ |
| First touch | $0.04 \pm 0.06$ | $2.09 \pm 0.87$ | $2.02 \pm 0.07$ | $0.15 \pm 0.12$ | $1.61 \pm 7.30$ |
| Dribble | $0.07 \pm 0.06$ | $2.09 \pm 0.88$ | $2.48 \pm 0.11$ | $0.12 \pm 0.11$ | $4.62 \pm 3.20$ |
| Positioning | $0.05 \pm 0.05$ | $2.23 \pm 0.73$ | $2.28 \pm 0.09$ | $0.13 \pm 0.11$ | $8.97 \pm 1.32$ |
| DNN | MAD Sternum acceleration (Z) | MAD Sternum quaternion (Z) | Entropy Head acceleration (X) | MAD Sternum quaternion (Y) | MAD Head jerk (Y) |
| Feature Importance | 1.80 e-8 | 1.28 e-8 | 9.77 e-9 | 9.72 e-9 | 9.11 e-9 |
| Head scan | $236 \pm 537$ | $1.08 \pm 2.28$ | $1.98 \pm 0.97$ | $0.55 \pm 0.69$ | $6.99 \pm 7.58$ |
| Not a head scan | $101 \pm 360$ | $0.52 \pm 1.56$ | $2.08 \pm 0.88$ | $0.28 \pm 0.41$ | $2.82 \pm 3.02$ |

mance, which is why the data in this study was split (50%/50%) (Xu et al. 2022). The type used in this study is an uncertainty-aware pseudo-label selection method with a confidence threshold of 0.70, run over 10 iterations as proposed by Rizve et al. (Rizve et al. 2021). This method only introduces data where the model is more certain (> 70 %), which is a trade-off, as it unavoidably leads to fewer data points (22 %), but also less noise introduced. Additionally, the process is repeated for 10 iterations, where the accuracy of the pseudo-labeling is constantly improving. To ensure that the unlabeled data was not solely introducing noise, 10 % of the unlabeled data was labeled and the LSTM and DNN predicted the labels of the 4 football actions and head scans respectively. The results were an accuracy of 0.79 for the LSTM and an accuracy of 0.84 for the DNN, indicating that the majority of the pseudo-labeled data would supply the algorithms with valuable information. However, this is slightly less than the usual accuracy recommended for SSL (> 90 %), but the method was substantiated as the F1-score of the LSTM was increased from 0.73 to 0.74 after the introduction of the uncertainty-aware pseudo-labeled data. Similarly, the F1-score of the DNN increased from 0.76 to 0.78.

## Post-processing

Due to the discrete nature of the movements, the windows were relatively short and a soft voting approach with a sliding window over 3 frames was therefore used. This sliding window calculated the mean of probabilities over 3 frames, with 50 % multiplied by the present frame and 25 % multiplied by the previous and subsequent frames. This made the predictions more rigid to change, while still emphasizing the current frame more, as illustrated for the LSTM in figure 5

## The algorithms

The foundation of the LSTM for predicting football actions and the DNN for predicting head scans were based on a systematic grid search but were subsequently fine-tuned using PyTorch (2.0) to find the optimal ones. Usually, in SSL different hyperparameters are needed for the model to predict the pseudo-labels and the final model. However, in this study, as only 50 % of the data was kept unlabeled and 82 % of the unlabeled data was utilized by using the uncertainty-aware pseudo-labeled technique, the difference in data sizes was not substantial enough to lead to different hyperparameters needed for optimal performance, which is why only a single LSTM and DNN is illustrated in figures 6 and 7.

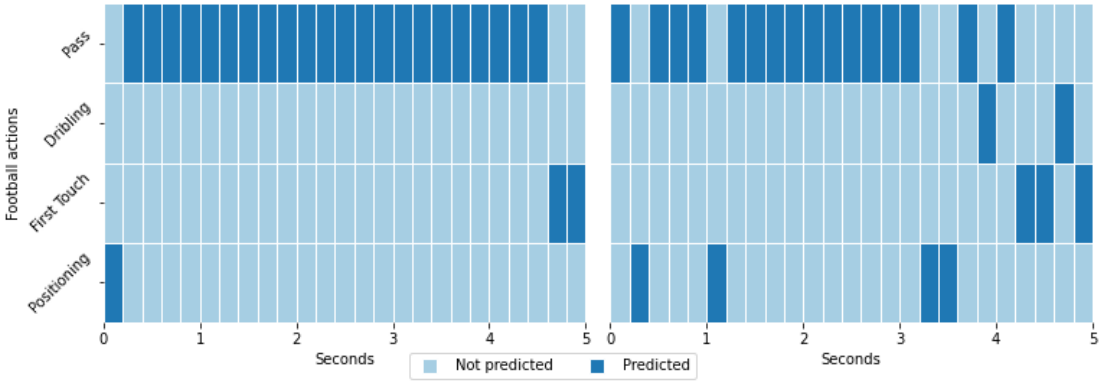


**Figure 5:** Shows the data soft voting smoothed over a 3-frame sliding window (left) and the raw data (right).

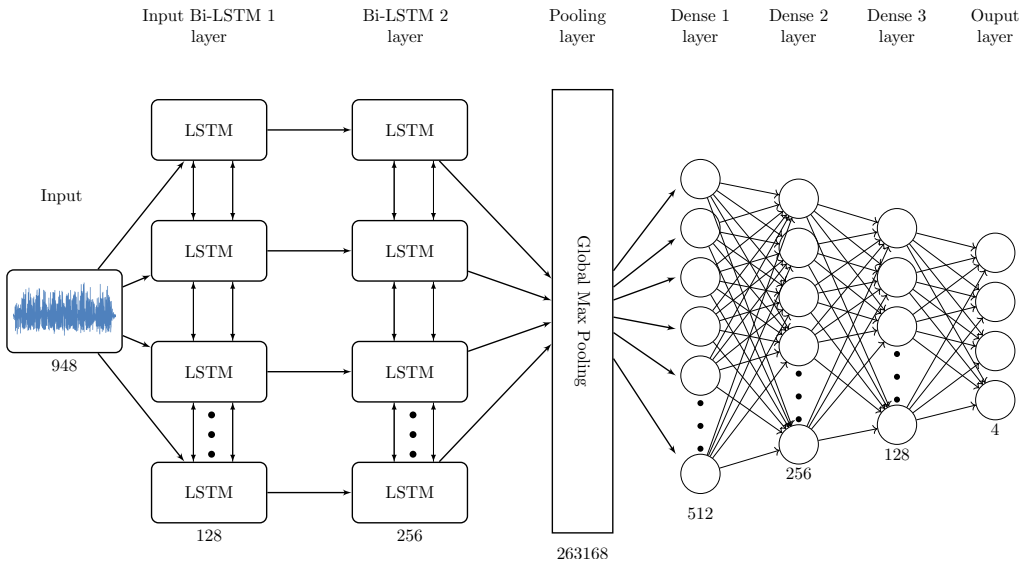


**Figure 6:** Shows the architecture of the LSTM to predict football actions. Under each layer is denoted the number of neurons in the respective layer.

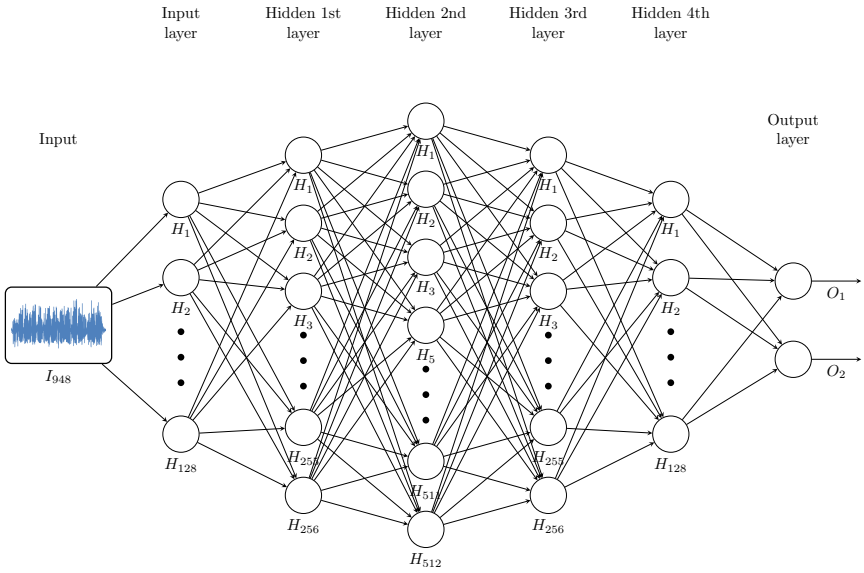


**Figure 7:** Shows the architecture of the DNN to predict the head scans. Under each layer is denoted the number of neurons in the respective layer.

### LSTM

The LSTM, shown in figure 6, was written using the application programming interface Keras (2.12.0), which can be imported into the system of choice through the ML platform TensorFlow (Chollet et al. 2015). Built into Keras are multiple ML tools that were used in this study. The LSTM consisted of 2 bidirectional LSTM layers (biLSTM) and 4 dense neural layers, each one using the parametric rectified linear unit (ReLU) activation function, except for the last output layer which used the softmax activation function. A dropout of 0.4 was implemented on all hidden layers, the input layer, and an early stop of 35 epochs to prevent overfitting. Focal categorical cross-entropy was chosen as the loss function and Adam was chosen as the optimizer. These variables were chosen based on a systematic grid search, which chose the combination of variables that yielded the best-performing algorithm.

### DNN

As with the LSTM, the DNN, illustrated in figure 7, was based on a systematic grid search. The DNN consisted of 4 hidden neural network layers, each one using the parametric rectified linear unit (ReLU) activation function, except for the last output layer which used the sigmoid activation function. To prevent overfitting, a dropout of 0.6 was implemented on all hidden layers, the input layer, and an early stop of 50 epochs. Focal categorical cross-entropy was chosen as the loss function and Root Mean Square Propagation (RMSProp) was chosen as the optimizer.

## Results

Presented in table 4 is the performance scores of the final SSL LSTM algorithm on predicting football actions.

**Table 4:** Performance metrics scores of the final semi-supervised LSTM model.

| Football Action Prediction Model | |
|---|---|
| Performance metrics | Score |
| Accuracy | 0.75 |
| Precision | 0.73 |
| Recall | 0.75 |
| F1-score | 0.74 |
| Krippendorff's alpha | 0.55 |
| Cross Validation | 0.74 |

The accuracy of the algorithm was found to be 0.75, indicating that it correctly classified 75% of the instances. Precision was calculated to be 0.73, which is a measure of the proportion of true positives among all positive predictions. The recall was calculated to be 0.75, which represents the proportion of true positive predictions among all actual positive instances. To provide a comprehensive evaluation of the algorithm, the F1-score was calculated, yielding a value of 0.74. The F1-score is the harmonic mean between precision and recall, providing a balanced measure of the algorithm's performance. To calculate the inter-rater reliability, Krippendorff's alpha was calculated, resulting in a value of 0.55, which indicates a moderate agreement between the four classes (Wong et al. 2021, Landis & Koch 1977). Furthermore, a cross-validation score of the algorithm was calculated to assess the algorithm's generalizability. This score was calculated to be 0.74, indicating the algorithm's consistency in performing across out-of-set data.

Illustrated in figure 8 is a confusion matrix based on the predictions of the football actions made by the LSTM. As it illustrates, the algorithm performed better when it had to predict the positioning and passing actions, but had more trouble when it had to predict the first touch and dribbling actions.

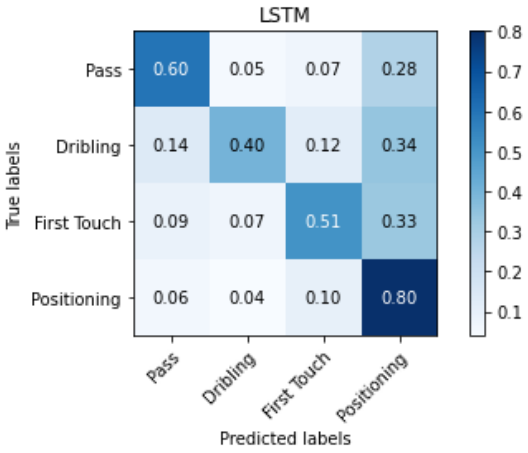


**Figure 8:** A normalized confusion matrix of the LSTM, with predicted events on the x-axis, and actual events on the y-axis. The hue indicates the share of the prediction.

This is further illustrated in the AUC-ROC curve in figure 9, where the AUC-score is greatest for the pass and positioning actions, scoring an AUC-score of 0.87 and 0.81, respectively, when comparing the prediction of the single action vs. the rest. The dribbling and first-touch actions had slightly worse AUC scores of 0.75 and 0.77, respectively.

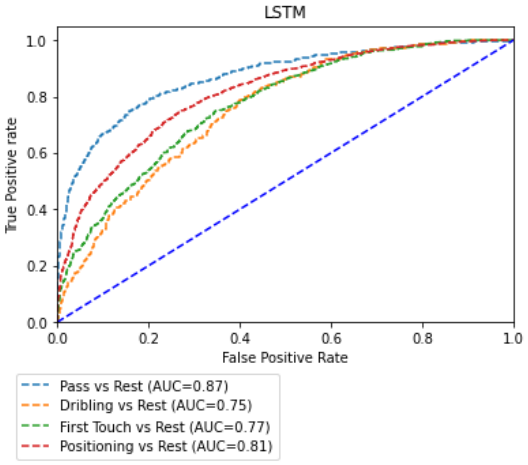


**Figure 9:** A one vs rest AUC-ROC curve of the LSTM, with the false positive rate on the x-axis and the true positive rate on the y-axis.

Similarly were the performance metrics calculated for the DNN's ability to predict head scans, as shown in table 5. The accuracy of the model was found to be 0.78, the precision was calculated to be 0.84, the recall was calculated to be 0.76, the F1-score was calculated to be 0.78, the Cohen's Kappa was calculated, resulting in a value of 0.61, which indicates a substantial agreement between the two classes (Wong et al. 2021, Landis & Koch 1977). And finally, a cross-validation score of the model was calculated to be 0.79. Likewise was a confusion matrix constructed for the DNN, as shown in figure 8. The confusion matrix indicates that the DNN performed slightly better when classifying not head scans compared to head scans. And finally, a AUC-ROC curve is illustrated in figure 11, which indicates the DNN's overall decent ability to distinguish between the two classes.

For predicting the football actions using only a single sensor, figure 12 shows the F1-score of the LSTM and DNN for the data from each of the IMU sensors separately. It shows that the IMU with the most valuable information for classifying football actions was the sensor on the right and left calves, with an F1-score of 0.65 and 0.63, respectively, and the IMU with the most valuable information for detecting head scans was the head sensor with an F1-score of 0.69, followed by the sternum with an F1-score of 0.66.

**Table 5:** Performance metrics scores of the final semi-supervised DNN model.

| Head Scan Prediction Model | |
|---|---|
| Performance metrics | Score |
| Accuracy | 0.78 |
| Precision | 0.84 |
| Recall | 0.76 |
| F1-score | 0.78 |
| Cohen's Kappa | 0.61 |
| Cross Validation | 0.79 |

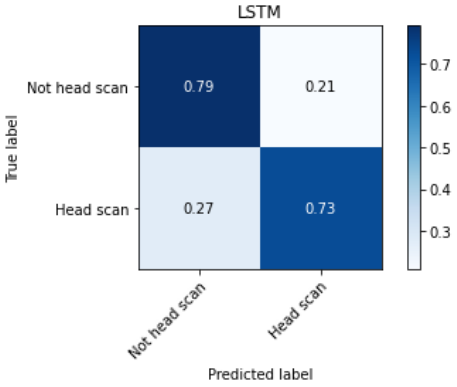


**Figure 10:** A normalized confusion matrix of the DNN, with predicted events on the x-axis, and actual events on the y-axis. The hue indicates the sum of the prediction.

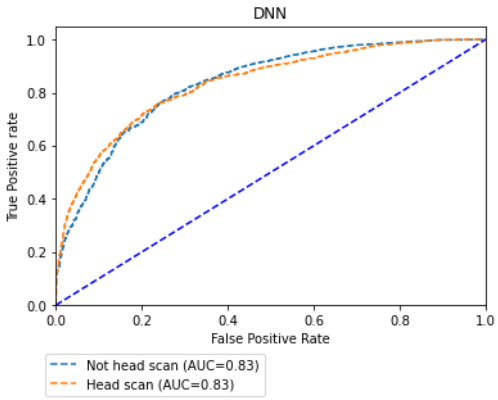


**Figure 11:** A AUC-ROC curve of the DNN, with the false positive rate on the x-axis and the true positive rate on the y-axis.

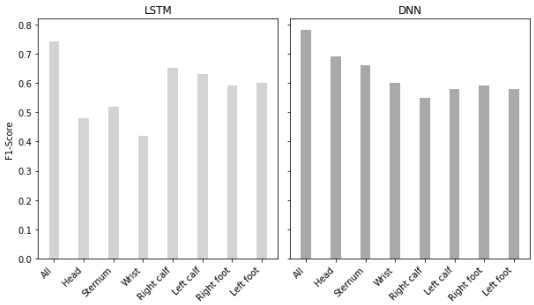


**Figure 12:** Shows the F1-score of the 7 different sensors isolated for both the LSTM and DNN, as well as all the sensors combined.

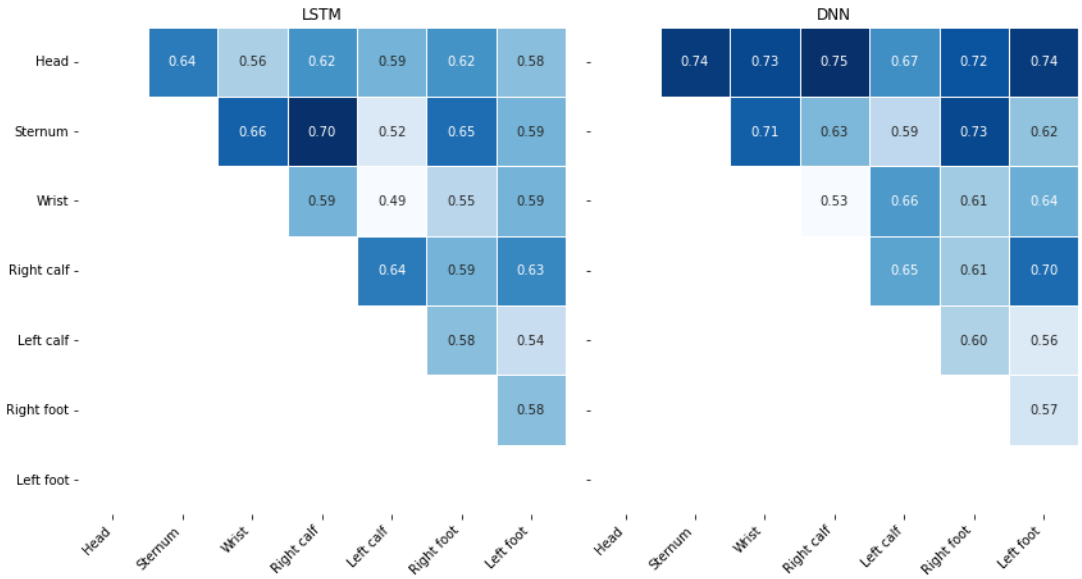


**Figure 13:** Shows F1-score of the pairwise combinations of both the LSTM and DNN.

Finally, for the pairwise prediction of football actions, figure 13 shows the best combination of the use of a two-sensor setup. The results show that the best combination of IMUs for predicting football actions was the sternum and right calf sensor with an F1-score of 0.70, and for the prediction of the head scan, the best combination of sensors was found to be the head and right calf IMUs, producing an F1-score of 0.75.

## Discussion

The results from this study showed that the combined use of IMUs, ML algorithms, and the principles commonly used in HAR has some capability of predicting football actions, hereunder head scans, with F1-scores ranging from 0.74 - 0.78. The literature on predicting specific football actions using the same principles as in this study is scarce, but predicting specific actions in other sports have been conducted, e.g. predicting tennis strokes, table tennis strokes, ball-related actions in volleyball, etc. (Blank et al. 2015, Dokic et al. 2020, Kautz et al. 2017, Connaghan et al. 2011). These studies achieved accuracies in the range of 82.5% and 96.7%, meaning that they scored higher accuracies than the ones presented in this study, but in different sports. The sports that scored the best results, were tennis and table tennis stroke detection, which have little to no overlap with football actions, as they are done with the upper extremities. This means that actions such as running and changing directions, in general actions with ground reaction forces, won't impact the sensors on the upper extremities as they would on the sensors on the lower extremities.

However, two studies have investigated action prediction in football using IMUs. Cuperman et al. achieved excellent results with an accuracy as high as 98.3% (Cuperman et al. 2022), also by using IMUs and deep learning, the included actions being sprinting, jogging, shooting, jumping, and passing. These

actions were, however, done in a controllable fashion, where participants would, e.g. perform 10 sprints followed by a shot. Cuperman et al. also used movements such as jogging and sprinting, which already have been shown to be recognizable movements by using ML and HAR principles before (Ahmadi et al. 2014, Ghazali et al. 2018). The football actions included in this study, i.e. passing, first touch, and dribbling show a great amount of overlap and are very complex with multiple movements happening at the same time both involving upper and lower limb segments. They are therefore harder to distinguish, even for humans, which is why the definitions of the movements were based around the feet contact times as they where easily visible on the video. The results of the prediction of the first touch and dribbling also showed the lowest accuracy, as illustrated in figures 8 and 9, which further shows the complexity and difficulty of predicting the first touch and dribbling actions. Methodologically, the data gathering in Cuperman et al., is further away from a real training or competition setting compared to our study, which could be a contributing factor to their better results. Furthermore, the sliding window approach used in Cuperman et al. is 1 s long, meaning a prediction only occurs once every second. This gives more stable predictions, but for fast and discrete movements, such as the ones included in this study, this approach is not preferable as the movements were as short as 150 ms and multiple football actions therefore might happen in the same window. However, the end-to-end method used by Cuperman et al. seems promising, and further research could look into if this method is also applicable with less distinguishable movement.

Another study by Schuldhaus et al. tried to differentiate between passes and shots in football, achieving an accuracy of 84.2% (Schuldhaus et al. 2015). Although still achieving a higher accuracy than this study, Schuldhaus et al. only had to perform a binary classification of either a pass or a shot. When dealing with predictions, the accuracy of the model will tend to fall when the number of possible predictions increases (Goodfellow et al. 2016). Having to predict four actions compared to two can have a big impact on the accuracies, as seen in this study on the difference between predicting football actions and head scans.

With regard to head scans, previous studies have shown the importance of them in elite-level football, increasing the chances of a successful follow-up pass if a head scan has been performed prior to receiving the ball (Jordet et al. 2020, 2013, McGuckian et al. 2019, Aksum et al. 2021). The literature specifically on predicting head scans is very scarce, but McGuckian et al. investigated the importance of head scans using an IMU similar to the ones used in this study to investigate the outcomes after a head scan (McGuckian et al. 2018). However, no details about the accuracy of their prediction model are to be found, the author only stating that it had been previously validated. This makes it difficult to compare the accuracy found on head scans in this study with others. An accuracy score of 0.78 would mean that most of the head scans a player performs will be detected. Furthermore, no studies have to date tried to detect head scans using a single sensor setup on the trunk, where all FIFA approved wearables are currently placed, however, this was tested in this study. Although the results shown in figure 12 indicate that the most important sensor placement for detecting head scans is the head with an F1-score of 0.69, the sensor placed on the sternum was close behind with an F1-score of 0.66. This shows the feasibility of detecting head scans using the single sensor setup most Sports Tech companies use today, which pre-

viously have not been demonstrated. As the earlier mentioned studies showed, the relationship between when a head scan and a pass is performed is extremely relevant and as the method proposed in this study showed some capability in predicting these, with 60 % of the passes and 73 % of the head scans accurately predicted as shown in figure 8 and 10 respectively. This could provide valuable information for players trying to improve their awareness and passing accuracy (Jordet et al. 2020, Aksum et al. 2021).

Being on the topic of discussing movements, a discussion of the confinements of the movements is also relevant. Even though the data for this study was attempted to be collected in a low interference way compared to similar studies, all participants were still in a controlled environment with no opposition. This is of course not relevant when comparing it to some real-life settings, i.e. training sessions or competitions, but implementations were applied to account for this by having lit-up boards function as a target to pass to. This accomplishes an aspect of orientation that the participant also has to account for in a training session or competition. Furthermore, by implementing time as a performance metric, the participants performed the actions as fast as possible, which is also an important factor for the participants if they were in a match, which should make the movements performed by the participants as close to a competitive setting as possible. This was important for this study, as the aim was to use as close to a competitive setting as possible, but without including the risk of opposing players damaging the sensors.

With regard to the algorithm and preprocessing steps, it could also be argued that other state-of-the-art methods could have been used. For example, Kanjilal & Uysal investigated whether or not manual feature extraction performs better than using raw temporal data (Kanjilal & Uysal 2021). They achieved better results using raw temporal data with a combination of a convolutional neural network (CNN), which learns features from the dataset without being told which features are important (Goodfellow et al. 2016). One of the drawbacks of using this method is that it needs a vast amount of data for it to be efficient (Kanjilal & Uysal 2021). Furthermore, the understanding of what features are important for HAR will decrease when using CNN algorithms, as the way the model will learn patterns will be unrecognizable for humans, meaning a deeper understanding of what is important for predictions will be practically impossible. It can make sense to use a CNN model when working with pictures or other spatial data that needs predicting, as it will be possible to get a visual representation of what the model deems to be important to learn. But when working with temporal data, as in this study, there will be a need to transform the input data into an arbitrary feature space with multiple dimensions. This gives no possibility to produce a feature map that provides any valuable information, and this "black-box" problem has for a long time been one of the greater issues with using these models, and a call has been made to switch to more interpretable models instead (Rudin 2019). Especially when exploring uncharted territory, using interpretable models will increase understanding and how to conduct research within the section.

## Sensor Placement

The other main focus of this study was to investigate which sensor placements supplied the most important information to predict football actions. As figure 12 shows, the most important sensor placements were on the right calf to predict football actions, and on the head to predict head scans. The drop-off from using only one sensor compared to

all sensors was an F1-score of 0.09 for the LSTM and DNN algorithms. It is therefore especially relevant, as most HAR research today is based on a single wearable, as the drop off in usability is high when including multi-sensor setups (Cust et al. 2019). This further suggests why the research on optimal sensor placements is very important and this has earlier been investigated within HAR (Rahn et al. 2021). The results by Rahn et al. showed that the optimal sensor placement is very task specific. They also demonstrated that some pairwise combinations of sensors can be higher performing than the combination of all sensors, due to the removal of noise. This, however, was not the case in this study, as the performance of all pairwise combinations of sensors was lower for both algorithms than the combination of all sensors. Nevertheless, this study shows how it is still possible to achieve decent results in some applications by using fewer, but more relevant sensors. Meaning that some of the sensors provide little to no valuable information and the removal of data from the 5 most irrelevant of the 7 sensors only decreased the performance of the LSTM by 0.04 and DNN by 0.03. This is especially relevant for creating a system with the intent of usability in the field, where the setup used in this study with 7 sensors simply is not feasible. And as all common hardware for movement classification today is based on a single or two sensors setup, researchers, coaches, players, etc. should consider usability and appropriate accuracy thresholds when choosing which setup to use.

Schuldhaus et al. showed an accuracy of 82.4% by using sensors in the boots when predicting passes and shots from other events (Schuldhaus et al. 2015). They stated that future research should investigate the use of different sensor placements, which was done in this study, and it shows that placing sensors on the foot of the players might not provide the most relevant information, but suggests placing sensors on the calves gives more relevant information. This might be due to the data quality, as slightly lower accuracies on IMUs have been demonstrated at higher movement velocities (Taylor et al. 2017). However, as this study did not include full power movements, such as a shot, the possible decrease in accuracy should be minimal.

## Data Quantity

When using ML algorithms, data quantity is one of the most important factors to keep in mind, as these algorithms require enormous amounts of data to be able to achieve optimal performance (Obermeyer & Emanuel 2016). Exactly how big of a quantity of data is needed is difficult to assess, as the amount of data required is very case specific. The DNN is generally considered inferior to the LSTM when dealing with time series data. However, in this study, the DNN and LSTM performed similarly. LSTMs usually need vast amounts of data to perform optimally (Sarker 2021), and the similar performance might therefore be an indicator that there were an insufficient amount of data for the optimal performance of the LSTM in this study. Meanwhile, collecting vasts amount of data will be beneficial for the performance of the ML algorithms, even when using an SSL approach, large quantities of this data have to be labeled, which can be time-consuming and thereby not realistic in a real-world setting. This is especially true when dealing with temporal data, where recording is often done with a high-frequency sampling rate device and subsequently each time step has to be labeled manually. Furthermore, as this study investigates discrete movements, where a participant can perform multiple movements within a second, the time to manually label increases significantly. An option to boost the data quantity without more time assigned to labeling is utilizing one or

multiple data augmentation techniques, such as Variational Autoencoder (VAE) or Generative Adversarial Networks (GANs), as proposed by Talavera et al. for time series data (Talavera et al. 2022). It was also possible possible to increase the amount of data for each window by using the data from all 17 IMUs. However, this would stray away from the focus on investigating the common sensor placements and the feasibility of using fewer sensors, why other ways to boost the amount of data seems preferable. Another similar approach to data augmentation is the one used in Cuperman et al., where the combination of a high sampling rate (500 Hz) and overlap (up to 99 %) functioned as a form of data augmentation (Cuperman et al. 2022). Nevertheless, overall the data quantity in this study seems to be sufficient, as the results of both predicting models show decent accuracy, while also withholding a high cross-validation score. A low score of cross-validation could be an indicator of the model overfitting, and while the root cause for an overfit model can be manyfold, it is often a lack of data that causes the overfitting (Ying 2019). Even though an increase in data quantity would always be beneficial, the time required to manually label the data would be unrealistic. This is the main reason for using SSL in this study, as it was then possible to collect a large amount of data without manually labeling it. In this study, the labeled and unlabeled data was split 50-50, however, other studies show great results splitting labeled and unlabeled data as high as 5-95 (Tarvainen & Valpola 2017, Shi et al. 2018). The key consideration centers around whether or not there is a way to collect data in big amounts without being too time-consuming, and if the model beforehand has a high enough accuracy on the training data, in such a way that it can accurately pseudo-label the unlabeled data. In light of this, it prompts the inquiry if the performance of the method presented in this study could show even better results with more data. For future research, it should therefore be considered to make a cost-benefit analysis on the time it takes to manually label the data, versus the time it takes to collect data for pseudo-labeling, and how it affects the accuracy. Furthermore, the split of labeled vs. unlabeled data needs further research to establish a consensus.

## Data Quality

One of the most important steps with ML is to ensure as high data quality as possible. This is why the hardware used in this study was state-of-the-art for motion capture. This included 7 IMUs containing three-axis accelerometer output (3D ± 16g), three-axis gyroscope output (±2000∘/s), and three-axis magnetometer output (± 1.9 Gauss) respectively, which is standard rates and should be sufficient to capture football actions as the upper limit for full power kicking motions is around 12 g and 1700 ∘/s (Zhou et al. 2020, Yu et al. 2022, Kellis & Katis 2007). The sampling rate was set at the maximum possible 60 Hz, which for HAR and some sports applications are generally considered sufficient (Khan et al. 2016, Twomey et al. 2018, Zhuang & Xue 2019). However, no studies have investigated the optimal sampling rate within football, although Gómez-Carmona et al. suggest that a sampling rate of 10 Hz is insufficient in estimating workload in football, and a sampling rate of $\geqslant$ 100 Hz is sufficient, however, no testing of the sampling rates in between was done (Gómez-Carmona et al. 2021). Stoeve et al. and Cuperman et al. used sampling rates of 200 and 500 Hz respectively, which might be a contributing factor to their overall better results, although Khan et al. indicate that such high sampling rates might be excessive (Stoeve et al. 2021, Cuperman et al. 2022, Khan et al. 2016).

Another common issue when dealing with

SSL is the uncertainty of the level of label noise in the pseudo-labels (Lokhande et al. 2020). An attempt was made to overcome this problem by using the uncertainty-aware SSL method and tested with a random sample of 10 % showing an accuracy of 0.82 and 0.87 for the LSTM and DNN, respectively. This inevitably introduces noise, which generally is tolerated as the efficacy of SSL has been demonstrated with up to 30 % label noise, but as a rule of thumb, the accuracy needs to be over 90 % (Lokhande et al. 2020). Nevertheless, the noise might be too high in this study, although this approach overall improved the results. However, no consensus has to date been established and further studies investigating the required accuracy for pseudo-labeling within HAR are needed.

Furthermore, the efficacy of the features used in this study is uncertain. They were mainly based on Zhu et al., who focused on HAR and mainly continuous movements. They suggest keeping both the accelerometer, gyroscope, magnetometer, and quaternion output as well as solely using time-domain features, however, they computed their features over a 3-second sliding window, which is 15 times larger than the window used in this study (Zhu et al. 2017). Thereby it is questionable whether the same features are the most optimal for the recognition of discrete motions, as the ones included in this study.

## Data Balancing

In this study, an attempt was made to interfere as little as possible during the training sessions. This was done to ensure as high fidelity of the data as possible, which is generally wanted in ML as it theoretically improves the robustness of the algorithm. However, this approach was for this application double-edged, as the movements were quite unbalanced. This is a problem; for example, the DNN could get a > 95 % accuracy by simply predicting 'not a head scan' all the time. Therefore, it was chosen to put more weight on the minority classes by utilizing the state-of-the-art oversampling algorithm ADASYN. The efficacy of ADASYN has demonstrated decent results in previous studies (Irvine et al. 2019, Chen et al. 2021). Oversampling, however, set up a paradox, as the focus was to ensure high fidelity data by low interference, but then subsequently it was needed to synthetically upsample data, where the fidelity is hard to ensure (Santos et al. 2018). The most optimal solution is to undersample the dataset, but this is rarely done in real-life applications, due to the usual scarcity of data, especially for more complex algorithms, such as LSTMs (Lemaître et al. 2017, Ramyachitra & Manikandan 2014). Furthermore, undersampling has also been demonstrated as being inferior to oversampling techniques such as ADASYN for some applications, which is why ADASYN was eventually selected for this study (Mohammed et al. 2020). However, Alharbi et al. suggest cluster-based oversampling methods are more efficient in HAR than Synthetic Minority Oversampling Technique (SMOTE), which is a slightly simpler, but similar technique to ADASYN and might warrant further investigation (Alharbi et al. 2022).

Another imbalance in the dataset was regarding the participants, as 14 were male but only 3 were female, and likewise only 3 participants were left-footed compared to 14 right-footed participants. This might be an issue as Tuncer et al. have demonstrated > 99 % accuracy in differentiating between genders in HAR (Tuncer et al. 2020), and furthermore, significant differences have been demonstrated between male and female movements in football (Sørensen et al. 2022, Sakamoto et al. 2013). The robustness of the algorithms for classifying female movements might therefore be subpar to classifying male movements. Likewise, there could

be a difference in the classification of right-footed and left-footed individuals, as the acceleration traces and angular rates are mirrored. However, no studies have investigated this difference within HAR and until such evidence is presented is the efficacy of the prediction of left-footed players unknown. This is one of the common issues in ML applications, i.e. the bias versus variance trade-off. In this study, both the LSTM and DNN were introduced to a moderate amount of bias, mainly through the dropout implemented. This was done to ensure the models were not overfitting and thereby increasing the generalization of the models (Jankowsky & Schroeders 2022). In general, as much variance as possible in the participants is needed to properly represent the target group. However, this was not possible with only 17 participants. They were all Danish, under 30, and with decent football experience, and thereby a larger and more diverse group of participants is needed to boost the generalizability of the models.

## Summary

Based on this study, the SSL algorithm seems as a decent method to predict various football actions (pass, dribbling, first touch, positioning, and head scans) with only temporal data from IMUs. The results showed an F1-score ranging from 0.74 to 0.78, with the highest accuracy being the binary detection of head scans and the lowest accuracy the similar movements dribbling and first touch. The LSTM should be best suited for temporal-type data, but the amount of data might be insufficient in this study and further studies are therefore needed. Furthermore, the information from the 7 most used sensor placements was investigated and the right calf gave the most important information for the prediction of football actions, whereas the head was most important for head scans, however, they both decreased the F1-score of 0.09. In case a two-sensor setup is possible, the F1-score can generally be maintained with only a decrease of 0.04 for the classification of football actions with sensors both on the sternum and right calf as well as a decrease of 0.03 for the classification of head scans with sensors on the head and right calf. This suggests that setups more feasible for coaches and players might be advantageous if a slight decrease in accuracy is tolerable.

# Bibliography

Abdi, H. & Williams, L. J. (2010), 'Principal component analysis', *Wiley interdisciplinary reviews: computational statistics* **2**(4), 433–459.

Ahmadi, A., Mitchell, E., Richter, C., Destelle, F., Gowing, M., O'Connor, N. E. & Moran, K. (2014), 'Toward automatic activity classification and movement assessment during a sports training session', *IEEE Internet of Things Journal* **2**(1), 23–32.

Aksum, K. M., Pokolm, M., Bjørndal, C. T., Rein, R., Memmert, D. & Jordet, G. (2021), 'Scanning activity in elite youth football players', *Journal of Sports Sciences* **39**(21), 2401–2410.

Alharbi, F., Ouarbya, L. & Ward, J. A. (2022), 'Comparing sampling strategies for tackling imbalanced data in human activity recognition', *Sensors* **22**(4), 1373.

Aljarrah, A. A. & Ali, A. H. (2019), Human activity recognition using pca and bilstm recurrent neural networks, *in* '2019 2nd International Conference on Engineering Technology and its Applications (IICETA)', IEEE, pp. 156–160.

Aljarrah, A. A. & Ali, A. H. (2021), 'Human activity recognition by deep convolution neural networks and principal component analysis', *Further Advances in Internet of Things in Biomedical and Cyber Physical Systems* pp. 111–133.

Ariza-Colpas, P. P., Vicario, E., Oviedo-Carrascal, A. I., Butt Aziz, S., Piñeres-Melo, M. A., Quintero-Linero, A. & Patara, F. (2022), 'Human activity recognition data analysis: History, evolutions, and new trends', *Sensors* **22**(9), 3401.

Aughey, R. J. (2011), 'Applications of gps technologies to field sports', *International journal of sports physiology and performance* **6**(3), 295–310.

Barron, D. J., Atkins, S., Edmundson, C. & Fewtrell, D. (2014), 'Accelerometer derived load according to playing position in competitive youth soccer', *International Journal of Performance Analysis in Sport* **14**(3), 734–743.

Bayat, A., Pomplun, M. & Tran, D. A. (2014), 'A study on human activity recognition using accelerometer data from smartphones', *Procedia Computer Science* **34**, 450–457.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A. & Raffel, C. A. (2019), 'Mixmatch: A holistic approach to semi-supervised learning', *Advances in neural information processing systems* **32**.

Bi, H., Perello-Nieto, M., Santos-Rodriguez, R., Flach, P. & Craddock, I. (2022), 'An active semi-supervised deep learning model for human activity recognition', *Journal of Ambient Intelligence and Humanized Computing* pp. 1–17.

Blank, P., Hoßbach, J., Schuldhaus, D. & Eskofier, B. M. (2015), Sensor-based stroke detection and stroke type classification in table tennis, *in* 'Proceedings of the 2015 ACM International Symposium on Wearable Computers', pp. 93–100.

Chen, J., Sun, Y. & Sun, S. (2021), 'Improving human activity recognition performance by data fusion and feature engineering', *Sensors* **21**(3), 692.

Chen, Z., Zhu, Q., Soh, Y. C. & Zhang, L. (2017), 'Robust human activity recognition using smartphone sensors via ct-pca and online svm', *IEEE transactions on industrial informatics* **13**(6), 3070–3080.

Chollet, F. et al. (2015), 'Keras'.
**URL:** *https://github.com/fchollet/keras*

Connaghan, D., Kelly, P., O'Connor, N. E., Gaffney, M., Walsh, M. & O'Mathuna, C. (2011), Multi-sensor classification of tennis strokes, *in* 'SENSORS, 2011 IEEE', IEEE, pp. 1437–1440.

Cuperman, R., Jansen, K. M. & Ciszewski, M. G. (2022), 'An end-to-end deep learning pipeline for football activity recognition based on wearable acceleration sensors', *Sensors* **22**(4), 1347.

Cust, E. E., Sweeting, A. J., Ball, K. & Robertson, S. (2019), 'Machine and deep learning for sport-specific movement recognition: A systematic review of model development and performance', *Journal of sports sciences* **37**(5), 568–600.

Dehghani, A., Sarbishei, O., Glatard, T. & Shihab, E. (2019), 'A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors', *Sensors* **19**(22), 5026.

Dokic, K., Mesic, T. & Martinovic, M. (2020), Table tennis forehand and backhand stroke recognition based on neural network, *in* 'Advances in Computing and Data Sciences: 4th International Conference, ICACDS 2020, Valletta, Malta, April 24–25, 2020, Revised Selected Papers 4', Springer, pp. 24–35.

Fergani, B. et al. (2013), Evaluating a new classification method using pca to human activity recognition, *in* '2013 International Conference on Computer Medical Applications (ICCMA)', IEEE, pp. 1–4.

Franklyn-Miller, A., Richter, C., King, E., Gore, S., Moran, K., Strike, S. & Falvey, E. (2017), 'Athletic groin pain (part 2): a prospective cohort study on the biomechanical evaluation of change of direction identifies three clusters of movement patterns', *British journal of sports medicine* **51**(5), 460–468.

Fu, Z., He, X., Wang, E., Huo, J., Huang, J. & Wu, D. (2021), 'Personalized human activity recognition based on integrated wearable sensor and transfer learning', *Sensors* **21**(3), 885.

Garcia-Gonzalez, D., Rivero, D., Fernandez-Blanco, E. & Luaces, M. R. (2020), 'A public domain dataset for real-life human activity recognition using smartphone sensors', *Sensors* **20**(8), 2200.

Ghazali, N., Shahar, N., Rahmad, N., Sufri, N. J., As' ari, M. & Latif, H. M. (2018), Common sport activity recognition using inertial sensor, *in* '2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA)', IEEE, pp. 67–71.

Gómez-Carmona, C. D., Rojas-Valverde, D., Rico-González, M., Ibáñez, S. J. & Pino-Ortega, J. (2021), 'What is the most suitable sampling frequency to register accelerometry-based workload? a case study in soccer', *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology* **235**(2), 114–121.

Goodfellow, I., Bengio, Y. & Courville, A. (2016), 'Deep learning', *The reference book for deep learning models* .

Gudmundsson, J. & Horton, M. (2017), 'Spatio-temporal analysis of team sports', *ACM Computing Surveys (CSUR)* **50**(2), 1–34.

Hamäläinen, W., Järvinen, M., Martiskainen, P. & Mononen, J. (2011), Jerk-based feature extraction for robust activity recognition from acceleration data, *in* '2011 11th International Conference on Intelligent Systems Design and Applications', IEEE, pp. 831–836.

He, H., Bai, Y., Garcia, E. A. & Li, S. (2008), Adasyn: Adaptive synthetic sampling approach for imbalanced learning, *in* '2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)', IEEE, pp. 1322–1328.

Herold, M., Goes, F., Nopp, S., Bauer, P., Thompson, C. & Meyer, T. (2019), 'Machine learning in men's professional football: Current applications and future directions for improving attacking play', *International Journal of Sports Science & Coaching* **14**(6), 798–817.

Irvine, N., Nugent, C., Zhang, S., Wang, H. & Ng, W. W. (2019), 'Neural network ensembles for sensor-based human activity recognition within smart environments', *Sensors* **20**(1), 216.

Jaén-Vargas, M., Leiva, K. M. R., Fernandes, F., Gonçalves, S. B., Silva, M. T., Lopes, D. S. & Olmedo, J. J. S. (2022), 'Effects of sliding window variation in the performance of acceleration-based human activity recognition using deep learning models', *PeerJ Computer Science* **8**, e1052.

Jankowsky, K. & Schroeders, U. (2022), 'Validation and generalizability of machine learning prediction models on attrition in longitudinal studies', *International Journal of Behavioral Development* **46**(2), 169–176.

Jordet, G., Aksum, K. M., Pedersen, D. N., Walvekar, A., Trivedi, A., McCall, A., Ivarsson, A. & Priestley, D. (2020), 'Scanning, contextual factors, and association with performance in english premier league footballers: an investigation across a season', *Frontiers in psychology* **11**, 553813.

Jordet, G., Bloomfield, J. & Heijmerikx, J. (2013), The hidden foundation of field vision in english premier league (epl) soccer players, *in* 'Proceedings of the MIT sloan sports analytics conference'.

Kanjilal, R. & Uysal, I. (2021), 'The future of human activity recognition: deep learning or feature engineering?', *Neural Processing Letters* **53**, 561–579.

Kautz, T., Groh, B. H., Hannink, J., Jensen, U., Strubberg, H. & Eskofier, B. M. (2017), 'Activity recognition in beach volleyball using a deep convolutional neural network: Leveraging the potential of deep learning in sports', *Data Mining and Knowledge Discovery* **31**, 1678–1705.

Kellis, E. & Katis, A. (2007), 'Biomechanical characteristics and determinants of instep soccer kick', *Journal of sports science & medicine* **6**(2), 154.

Khan, A., Hammerla, N., Mellor, S. & Plötz, T. (2016), 'Optimising sampling rates for accelerometer-based human activity recognition', *Pattern Recognition Letters* **73**, 33–40.

Khanal, B., Rivas, P. & Orduz, J. (2021), Human activity classification using basic machine learning models, *in* '2021 International Conference on Computational Science and Computational Intelligence (CSCI)', IEEE, pp. 121–126.

Kwapisz, J. R., Weiss, G. M. & Moore, S. A. (2011), 'Activity recognition using cell phone accelerometers', *ACM SigKDD Explorations Newsletter* **12**(2), 74–82.

Landis, J. R. & Koch, G. G. (1977), 'The measurement of observer agreement for categorical data', *biometrics* pp. 159–174.

Lemaître, G., Nogueira, F. & Aridas, C. K. (2017), 'Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning', *The Journal of Machine Learning Research* **18**(1), 559–563.

Lokhande, V. S., Tasneeyapant, S., Venkatesh, A., Ravi, S. N. & Singh, V. (2020), Generating accurate pseudo-labels in semi-supervised learning and avoiding overconfident predictions via hermite polynomial activations, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 11435–11443.

McGuckian, T. B., Cole, M. H., Chalkley, D., Jordet, G. & Pepping, G.-J. (2019), 'Visual exploration when surrounded by affordances: frequency of head movements is predictive of response speed', *Ecological Psychology* **31**(1), 30–48.

McGuckian, T. B., Cole, M. H., Jordet, G., Chalkley, D. & Pepping, G.-J. (2018), 'Don't turn blind! the relationship between exploration before ball possession and on-ball performance in association football', *Frontiers in psychology* **9**, 2520.

Mohammed, R., Rawashdeh, J. & Abdullah, M. (2020), Machine learning with oversampling and undersampling techniques: overview study and experimental results, *in* '2020 11th international conference on information and communication systems (ICICS)', IEEE, pp. 243–248.

Myn, U., Link, M. & Awinda, M. (2021), *Xsens mvn user manual*.

Narayanan, A., Desai, F., Stewart, T., Duncan, S. & Mackay, L. (2020), 'Application of raw accelerometer data and machine-learning techniques to characterize human movement behavior: a systematic scoping review', *Journal of Physical Activity and Health* **17**(3), 360–383.

Nunavath, V., Johansen, S., Johannessen, T. S., Jiao, L., Hansen, B. H., Berntsen, S. & Goodwin, M. (2021), 'Deep learning for classifying physical activities from accelerometer data', *Sensors* **21**(16), 5564.

Obermeyer, Z. & Emanuel, E. J. (2016), 'Predicting the future—big data, machine learning, and clinical medicine', *The New England journal of medicine* **375**(13), 1216.

Oh, S., Ashiquzzaman, A., Lee, D., Kim, Y. & Kim, J. (2021), 'Study on human activity recognition using semi-supervised active transfer learning', *Sensors* **21**(8), 2760.

Potdar, K., Pardawala, T. S. & Pai, C. D. (2017), 'A comparative study of categorical variable encoding techniques for neural network classifiers', *International journal of computer applications* **175**(4), 7–9.

Prasad, A., Tyagi, A. K., Althobaiti, M. M., Almulihi, A., Mansour, R. F. & Mahmoud, A. M. (2021), 'Human activity recognition using cell phone-based accelerometer and convolutional neural network', *Applied Sciences* **11**(24), 12099.

Preece, S. J., Goulermas, J. Y., Kenney, L. P. & Howard, D. (2008), 'A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data', *IEEE Transactions on Biomedical Engineering* **56**(3), 871–879.

Rahn, V. X., Zhou, L., Klieme, E. & Arnrich, B. (2021), Optimal sensor placement for human activity recognition with a minimal smartphone-imu setup., *in* 'SENSORNETS', pp. 37–48.

Ramyachitra, D. & Manikandan, P. (2014), 'Imbalanced dataset classification and solutions: a review', *International Journal of Computing and Business Research (IJCBR)* **5**(4), 1–29.

Richter, C., King, E., Strike, S. & Franklyn-Miller, A. (2019), 'Objective classification and scoring of movement deficiencies in patients with anterior cruciate ligament reconstruction', *PloS one* **14**(7), e0206024.

Rizve, M. N., Duarte, K., Rawat, Y. S. & Shah, M. (2021), 'In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning', *arXiv preprint arXiv:2101.06329* .

Rommers, N., Rössler, R., Verhagen, E., Vandecasteele, F., Verstockt, S., Vaeyens, R., Lenoir, M., D'Hondt, E. & Witvrouw, E. (2020), 'A machine learning approach to assess injury risk in elite youth football players', *Medicine and science in sports and exercise* **52**(8), 1745–1751.

Rudin, C. (2019), 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature machine intelligence* **1**(5), 206–215.

Sakamoto, K., Shimizu, Y., Yamada, E., Hong, S. & Asai, T. (2013), 'Difference in kicking motion between female and male soccer players', *Procedia Engineering* **60**, 255–261.

Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H. & Santos, J. (2018), 'Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches [research frontier]', *ieee ComputatioNal iNtelligeNCe magaziNe* **13**(4), 59–76.

Sarker, I. H. (2021), 'Machine learning: Algorithms, real-world applications and research directions', *SN computer science* **2**(3), 160.

Schuldhaus, D., Zwick, C., Körger, H., Dorschky, E., Kirk, R. & Eskofier, B. M. (2015), Inertial sensor-based approach for shot/pass classification during a soccer match, *in* 'KDD workshop on large-scale sports analytics', pp. 1–4.

Shahar, N., Ghazali, N., As'Ari, M. & Swee, T. (2020), Wearable inertial sensor for human activity recognition in field hockey: Influence of sensor combination and sensor location, *in* 'Journal of Physics: Conference Series', Vol. 1529, IOP Publishing, p. 022015.

Shi, W., Gong, Y., Ding, C., Tao, Z. M. & Zheng, N. (2018), Transductive semi-supervised deep learning using min-max features, *in* 'Proceedings of the European Conference on Computer Vision (ECCV)', pp. 299–315.

Singh, A., Chakraborty, O., Varshney, A., Panda, R., Feris, R., Saenko, K. & Das, A. (2021), Semi-supervised action recognition with temporal contrastive learning, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 10389–10399.

Sørensen, A., Haugen, E. C. & van den Tillaar, R. (2022), 'Is there a sex difference in technical skills among youth soccer players in norway?', *Sports* **10**(4), 50.

Stoeve, M., Schuldhaus, D., Gamp, A., Zwick, C. & Eskofier, B. M. (2021), 'From the laboratory to the field: Imu-based shot and pass detection in football training and game scenarios using deep learning', *Sensors* **21**(9), 3071.

Talavera, E., Iglesias, G., González-Prieto, Á., Mozo, A. & Gómez-Canaval, S. (2022), 'Data augmentation techniques in time series domain: a survey and taxonomy', *arXiv preprint arXiv:2206.13508* .

Tarvainen, A. & Valpola, H. (2017), 'Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results', *Advances in neural information processing systems* **30**.

Taylor, L., Miller, E. & Kaufman, K. R. (2017), 'Static and dynamic validation of inertial measurement units', *Gait & posture* **57**, 80–84.

Thapa, K., Seo, Y., Yang, S.-H. & Kim, K. (2023), 'Semi-supervised adversarial autoencoder to expedite human activity recognition', *Sensors* **23**(2), 683.

Tsunoda, T., Komori, Y., Matsugu, M. & Harada, T. (2017), Football action recognition using hierarchical lstm, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition workshops', pp. 99–107.

Tuncer, T., Ertam, F., Dogan, S. & Subasi, A. (2020), 'An automated daily sports activities and gender recognition method based on novel multikernel local diamond pattern using sensor signals', *IEEE Transactions on Instrumentation and Measurement* **69**(12), 9441–9448.

Twomey, N., Diethe, T., Fafoutis, X., Elsts, A., McConville, R., Flach, P. & Craddock, I. (2018), A comprehensive study of activity recognition using accelerometers, *in* 'Informatics', Vol. 5, MDPI, p. 27.

Wakatsuki, T. & Yamada, N. (2020), 'Difference between intentional and reactive movement in side-steps: Patterns of temporal structure and force exertion', *Frontiers in Psychology* **11**, 2186.

Walse, K. H., Dharaskar, R. V. & Thakare, V. M. (2016), Pca based optimal ann classifiers for human activity recognition using mobile sensors data, *in* 'Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 1', Springer, pp. 429–436.

Wong, K., Paritosh, P. & Aroyo, L. (2021), 'Cross-replication reliability–an empirical approach to interpreting inter-rater reliability', *arXiv preprint arXiv:2106.07393* .

Wu, F., Wang, Q., Bian, J., Ding, N., Lu, F., Cheng, J., Dou, D. & Xiong, H. (2022), 'A survey on video action recognition in sports: Datasets, methods and applications', *IEEE Transactions on Multimedia* .

Xia, C. & Sugiura, Y. (2021), 'Optimizing sensor position with virtual sensors in human activity recognition system design', *Sensors* **21**(20), 6893.

Xing, J. & Li, H. (2022), 'Study about football action recognition method based on deep learning and improved dynamic time warping algorithm', *Mobile Information Systems* **2022**.

Xu, Y., Wei, F., Sun, X., Yang, C., Shen, Y., Dai, B., Zhou, B. & Lin, S. (2022), Cross-model pseudo-labeling for semi-supervised action recognition, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 2959–2968.

Ying, X. (2019), An overview of overfitting and its solutions, *in* 'Journal of physics: Conference series', Vol. 1168, IOP Publishing, p. 022022.

Yu, C., Huang, T.-Y. & Ma, H.-P. (2022), 'Motion analysis of football kick based on an imu sensor', *Sensors* **22**(16), 6244.

Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M. & Shinozaki, T. (2021), 'Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling', *Advances in Neural Information Processing Systems* **34**, 18408–18419.

Zhao, Y., Zhou, H., Lu, S., Liu, Y., An, X. & Liu, Q. (2022), 'Human activity recognition based on non-contact radar data and improved pca method', *Applied Sciences* **12**(14), 7124.

Zhou, L., Fischer, E., Tunca, C., Brahms, C. M., Ersoy, C., Granacher, U. & Arnrich, B. (2020), 'How we found our imu: Guidelines to imu selection and a comparison of seven imus for pervasive healthcare applications', *Sensors* **20**(15), 4090.

Zhou, Z.-H. (2018), 'A brief introduction to weakly supervised learning', *National science review* **5**(1), 44–53.

Zhu, J., San-Segundo, R. & Pardo, J. M. (2017), 'Feature extraction for robust physical activity recognition', *Human-centric Computing and Information Sciences* **7**, 1–16.

Zhuang, Z. & Xue, Y. (2019), 'Sport-related human activity detection and recognition using a smartwatch', *Sensors* **19**(22), 5001.