Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences
Department of Neuromedicine and Movement Science

Sondre Hovda Dahlskås

Validation of the NTNU HAR-model on complex physical activity and detecting ballgame periods from the model predictions.

Master's thesis in Human Movement Science
Supervisor: Paul Jarle Mork & Ellen Marie Bardal
December 2020

NTNU
Norwegian University of
Science and Technology

Sondre Hovda Dahlskås

Validation of the NTNU HAR-model on complex physical activity and detecting ballgame periods from the model predictions.

Master's thesis in Human Movement Science
Supervisor: Paul Jarle Mork & Ellen Marie Bardal
December 2020

Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences
Department of Neuromedicine and Movement Science

NTNU
Norwegian University of
Science and Technology

Abstract

Background: The NTNU Human Activity Recognition (HAR) model utilize state of the art machine learning techniques to predict the type of daily physical activity (PA) the participant is doing with 94% accuracy. Validation on other complex activity patterns like ballgames and novel analysis methods to detect ballgames from the NTNU HAR-model predictions are in demand.
Study Aim: The aim of this study is to use handball as a paradigm to assess the validity of the NTNU HAR-model as a classifier during ballgames.
We also explore the possibility of using the NTNU HAR-model predictions to detect periods of ballgames based on how many times the predicted activity changes (PAC) within a specific timeframe.
Methods: Six adolescent males equipped with two tri-axial accelerometers carried out two handball training sessions over two days with their team while being filmed from all corners of the sports hall. Accelerometers were kept on between training sessions. Using predefined PA definitions, the observed physical activity type (e.g. walking, running, standing) from the video was annotated frame-by-frame to use as a solution in validating the NTNU HAR-model. Inter-rater reliability (IRR) of the video annotation was calculated. The predictions from the NTNU HAR-model were grouped in different windows and the calculated PAC was used to detect ballgame periods.
Results: Overall accuracy of the NTNU HAR-model was 77% with sensitivity variating from 67% to 84% and precision variating from 63% to 96% in the four main activity type categories (sitting, standing, walking, running). IRR scores had Cohen's Kappa values of 0.92 and 0.95. The detecting ballgame model proved to reach an accuracy and specificity of above 90% with high resolution (7min) but with lower sensitivity and precision scores of below 70%. Lowering the resolution (30-40min) increased all parameters with up to 96% accuracy and specificity, with above 80% sensitivity and precision.
Conclusion: The NTNU HAR-model in its current state does not provide us with a valid tool to predict activity types during ballgames. It is therefore recommended to use a different HAR-model approach or altered classification method of the NTNU HAR-model if the target is to predict activity types during ballgames.
Based on the results from this thesis, we can use PAC calculated from predictions generated by the NTNU HAR-model to detect periods of ballgames. However, this method should be tested on a larger study population including other moderate to vigorous PA before used in future research.

Sammendrag

Bakgrunn: NTNU Human Activity Recognition (HAR) modellen utnytter toppmoderne maskinlæring teknikker til å predikere hvilken type daglig fysisk aktivitet (FA) deltakeren utfører med 94% nøyaktighet. Det er behov for validering av modellen på komplekse aktivitetsmønstre som ballspill, og bruk av NTNU HAR-modell prediksjonene til å oppdage ballspill.
Mål: Målet med denne studien er å bruke håndball som et paradigme for ballspill til å validere NTNU HAR-modellen på komplekse aktivitetsmønstre. Vi utforsker også muligheten til å bruke NTNU HAR-modell beregningene til å predikere ballspill basert på hvor mange aktivitetsskifter (PAC) vi har innenfor en definert tidsramme.
Metode: Seks unge (16år) menn, utstyrt med to tre-akslet akselerometre, utførte to håndball treninger over to dager sammen med deres håndballag. Laget ble filmet fra alle hjørner i sportshallen. Akselerometrene ble brukt mellom treningsøktene. Ved å bruke forhåndsbestemte FA definisjoner ble den observerte fysiske aktivitetstypen (f.eks. gå, løpe, stå) annotert bilde for bilde fra videoen for å brukes til fasit under validering av NTNU HAR-modellen. Prediksjonene fra NTNU HAR-modellen ble grupperte i ulike vinduer og den kalkulerte PAC ble brukt til å oppdage ballspillperioder.
Resultat: Den generelle nøyaktigheten til NTNU HAR-modellen var 77% med sensitivitet mellom 67% og 84%, og presisjon mellom 63% og 96% i de fire hovedaktivitetstypene (sitte, stå, gå, løpe). IRR resultatene hadde Cohen`s Kappa verdier på 0.92 og 0.95. Modellens evne til å predikere ballspill viste en nøyaktighet og spesifikasjon på over 90% med høy oppløsning (7min), men med lavere sensitivitet og nøyaktighet på under 70%. Ved å senke oppløsningen (30-40min) økte alle parameterne opp til 96% for nøyaktighet og spesifikasjon, og over 80% for sensitivitet og presisjon.
Konklusjon: NTNU HAR-modellen i nåværende tilstand gir oss ikke et gyldig verktøy til å predikere aktivitetstyper under ballspill. Det er derfor anbefalt å bruke en ulik HAR-modell metode eller en endret klassifiseringsmetode for NTNU HAR-modellen hvis målet er å predikere aktivitetstyper under ballspill. Basert på resultatene fra denne avhandlingen, kan vi bruke PAC kalkulert fra prediksjoner generert av NTNU HAR-modellen, til å oppdage perioder med ballspill. Det er likevel nødvendig å teste denne metoden på en større testgruppe, inkludert andre moderate til vigorøse FA, før det brukes i framtidige studier.

Acknowledgements

I would like to thank several people paramount to make this thesis a reality.
I would first like to thank my supervisor Paul Jarle Mork for his patience and accurate guidance in addition to being an inspiration behind this thesis. I would also like to give thanks to Ellen Marie Bardal for sharing her insight and provide feedback throughout the project.
Thank you Astrid Ustad and Hilde Bremseth Bårdstu for help with video annotation and data collection.
I would also like to thank fellow student Roar Munkeby Fenne for valuable opinions and discussions.
A thank you to Håkon Slåtten Kjærnli for his work with handling the testing process through the NTNU HAR-model, and of course to everyone at the Department of Computer and Information Sciences (IDI) responsible for developing the NTNU HAR-model.
Lastly, thank you to everyone involved and participating in the data collection and activity definition projects executed before I joined.

Table of Contents

1. Introduction

New technology and applications provide opportunities to improve physical activity characterization and enable novel analysis methods as a viable option.

It is well known that physical activity (PA) is one of the most beneficial lifestyle supplements to improve physical and psychological well-being (1-7). Health benefits of physical exercise starts immediately and will also provide long term results including reduced risk of several non-communicable diseases such as diabetes type 2, cardiovascular disease and certain types of cancer (3).The World Health Organization (4) define physical activity as "*any bodily movement produced by skeletal muscles that requires energy expenditure*", which can be further categorized depending on the amount of energy expenditure needed. The "*Metabolic equivalent of task (MET) refers to the energy expenditure required to carry out a specific activity, and 1 MET is the rate of energy expenditure while sitting at rest."* as defined by The Physical Activity Guidelines for Americans (5). The categories commonly used to describe the intensity of PA are light (>3.0MET), moderate (3.0-6.0MET) and vigorous (<6.0MET) physical activity (5, 6). The World Health Organization recommends 150 minutes of moderate PA, or 75 minutes of vigorous PA weekly for adults aged 18-64 years. For adolescents aged 5-17 years, 60 minutes of moderate to vigorous physical activity (MVPA) is recommended daily. Muscle strengthening activities should also be included 2-3 times a week. Globally, around 23% of adults and 81% of adolescents did not meet the recommended levels of physical activity in 2010 (4, 7).

With the COVID-19 pandemic, a further decrease in PA is expected due to public health recommendations including stay-at-home orders and closure of PA enabling facilities (7). Great recommendations and documentation have been made to mitigate these effects (e.g. Hammami et. al (8)), but it is hard to influence all elements affected by the pandemic. Social isolation in itself has also proven to negatively impact PA levels for multiple age groups, which is a concern in today's society (9, 10).

To understand and govern physical activity patterns in society, accurate measurement methods to correctly quantify and assess activity levels are needed. Being able to work with valid and precise measurements leads scientists and health personnel to improved health recommendations and research (11, 12).

To accurately classify physical activity volume and intensity, an objective measurement method is preferred to a subjective one (13, 14). The technological advances in accelerometry sensors and micro-electronics over the past years have made use of accelerometers in research more feasible. With their small size, high precision, enduring battery, and relatively low cost, using accelerometry makes it possible to use an objective measurement method even in larger epidemiological studies (15).

The potential to model physical activity energy expenditure from simple linear regression approaches using the accelerometer manufacturers output method; counts, have been the easiest and most common statistical post-processing approach used by researchers (15). Counts can be classified as the number of times the acceleration signal surpass a threshold within a time frame (16). The potential problem with this method is the lack of knowledge about the manufacturers signal processing, which most often is a closely guarded secret within each company. This difference makes the intensity based cut off

points for physical activity different and non-comparable between accelerometer models, which limits the research possibilities of a study (15, 16).

At the Norwegian University of Science and Technology (NTNU), a Human Activity Recognition (HAR) model that uses raw acceleration data to predict the type of physical activity being performed has been developed (17-21). Raw acceleration data is not post-processed, which give scientists the opportunity to utilize pre-collected data with future software and machine learning techniques without being held back by old post-processing methods. Raw acceleration can also provide detailed information that other objective measurement methods like heart rate (HR), counts, steps and GPS have trouble providing. By using the gravity component of the acceleration signal, researchers can detect posture of the body or a limb dependent on where the accelerometer is placed (22). Combining this with multiple accelerometers in different positions unlocks the potential to gain a complete picture of static activities. Dynamic activities can be classified by their unique acceleration patterns, but both static and dynamic activities require the use of advanced tools to be analyzed (18).

The NTNU HAR-model utilizes state of the art machine learning techniques and can predict the type of daily physical activity the participant is doing with 94% accuracy (19). The NTNU HAR-model studies and other research (23) provides us with a baseline accuracy of above 80% as acceptable, and above 90% as excellent. This includes static activities like sitting, standing, lying down, and dynamic activities like walking, running, bending and jumping. The NTNU HAR-model achieves this by analyzing raw acceleration signals from two tri-axial accelerometers placed on the lower back and middle of the right thigh, and recognizing patterns using advanced calculations based on multiple time and frequency domain features. The NTNU HAR-model has also been developed to encompass different challenges and goals, as identifying sensor no-wear time with above 97% accuracy (21), detecting sleep and wake periods with above 94% accuracy (17), and specialization toward activity recognition in stroke patients achieving above 93% accuracy (20).

The purpose of the NTNU HAR-model is to analyze accelerometer data from more than 58 000 participants that were given the choice to wear the two tri-axial accelerometers for a week during the latest iteration of The Nord-Trøndelag health study - HUNT4. A total of above 38 000 chose to wear the accelerometers during the study (24). Because of this target group, the NTNU HAR-model has been developed with datasets focused on free-living daily activities in adults. With the NTNU HAR-model developed for daily activities in thousands of people, it does have some limitations. For example, activities that are mainly performed in specialized settings instead of normal daily activities tend to include movement patterns the developers never trained the model to recognize. The time window of each prediction might also be a limiting factor in these settings, as the model predicts what activity the person has done the most of within a 5 second window (19).

Some activities where this could become detrimental can be classified as complex physical activities (CPA) (e.g., football, handball, dance, gymnastics) (25, 26). Viewing ballgames as a paradigm for CPA, needing multiple movement patterns to achieve the players goal within a short time window is common (27, 28). This could result in a loss of important data for researchers analyzing PA based on the NTNU HAR-model.

Ballgames are usually classified as MVPA (27, 29, 30) but includes periods of activities recognized as sedentary or light (e.g. standing, walking, sitting, lying, bending). Differences in human activity recognition and physical response between healthy adults and

adolescents during ballgames should be minor (25, 26). With the potential challenges of scientists recognizing a bout of ballgames as MVPA based solely on the NTNU HAR-model predictions, a solution utilizing the possibilities of the existing model to further increase gained knowledge without the need of making changes to the model is needed.

The aim of this study is to use handball as a paradigm to assess the validity of the NTNU HAR-model as a classifier during ballgames.

We also explore the possibility of using the NTNU HAR-model predictions to detect periods of ballgames based on how many times the predicted activity changes (PAC) within a specific timeframe.

1. Method

Data collection used in this thesis is part of a larger validation study on CPA and Human Activity Recognition. The study protocol was executed by scientists at NTNU with approval from NSD – Norwegian Center for Research Data.

2.1 Participants

Six healthy adolescent participants where included in the data collection used for this thesis. Characteristics of the participants are found in table 2.1. The participants were recruited through their handball organization with explicit information that participation will not impact the subject's relationship with their organization or support staff. All interested team members received written information prior to participation and signed a written consent upon inclusion in the study.

Table 2.1: Characteristics of the study participants. Values are mean ± standard deviation (range).

	Boys (n=6)
Age (years)	16 ± 0,0 (16 - 16)
Weight (kg)	77,2 ± 8,0 (62 - 85)
Height (cm)	182,5 ± 4,5 (174 - 186)

2.2 Data collection and equipment

Two scientists from NTNU with two observing master students guided the participants though a pre-planned protocol in February 2019. Data collection started in the evening. Anthropometry data was first collected, followed by the mounting of two Axivity AX3 accelerometers placed on the lower back and middle of the right thigh as presented in figure 2.1. The subjects were also equipped with a Polar m400 heart rate (HR) -watch with HR-belt. Four GoPro cameras were set-up in a quadrant to film one subject at a time from all horizontal angles. Synchronization of cameras and each of the subjects equipped AX3 accelerometers and HR watch followed.

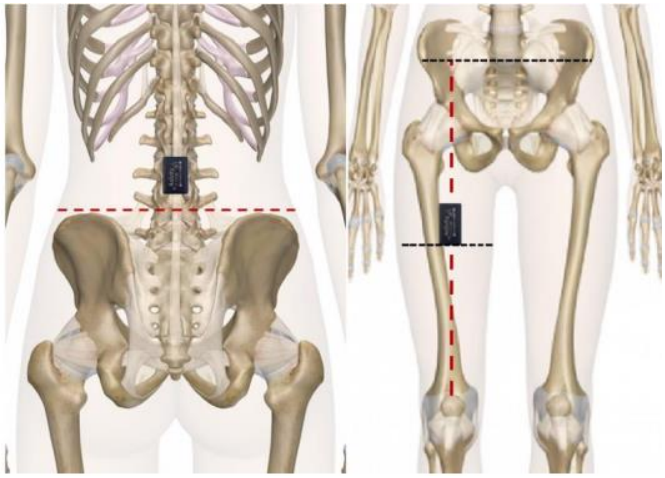


Figure 2.1: Anatomically correct placement of AX3 on lower back and thigh.

The cameras where then shifted to each corner of the empty room before the subjects followed a structured protocol of common movements in handball to increase baseline data. The cameras where then moved and placed in each corner of the sports hall (figure 2.2), where the subjects joined the rest of their handball team for a normal practice. After practice, the subjects turned in their HR-watch and belt but kept the accelerometers on.

On day two, the subjects arrived back in the sports hall for normal practice at approximately the same time of day as day one. After handing out HR-equipment and a short interview to assess if there were any sensor no-wear periods of accelerometers or notable bouts of physical activity between the practices, another round of synchronization ensued before the subject joined their normal practice. After the practice, all the equipment was returned to the scientists for analysis.

2.3 Video annotation

The video annotation was done frame-by-frame from the 25fps GoPro video in ANVIL (version 6.0). The program is a video and audio annotation tool where user defined specifications and features can be added to suit the annotation objective. The annotators use the program to register which of the 14 defined main activities (appendix 1) the participant is doing every 0,04s during the ballgame session as illustrated in figure 2.2. The acceleration and pulse signals are not visible while annotating. The completed annotations were exported to text format to be used in later validation analysis.

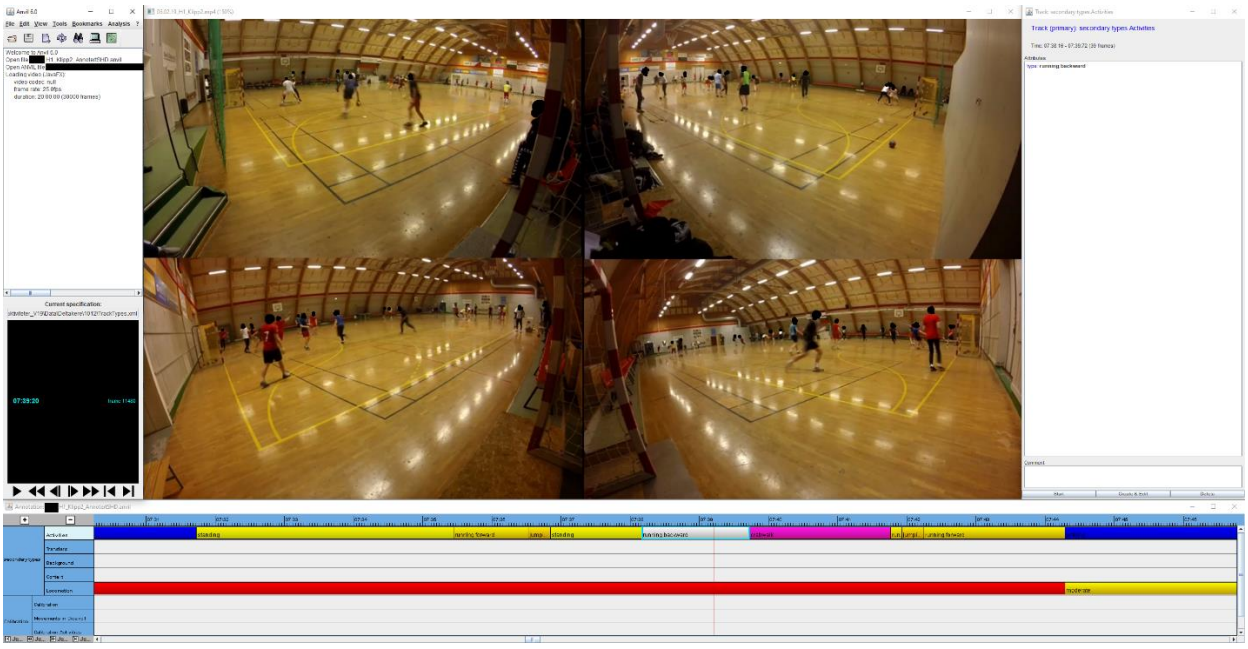


Figure 2.2: Illustration of ANVIL with censored data

2.3.1 Inter-rater reliability

The Inter-rater reliability (IRR) of the video annotation was calculated using an in-built feature of ANVIL. IRR was calculated between annotator 1 and 3, as well as annotator 2 and 3. By numbering each 20min video that annotator 1 and 2 annotated, annotator 3 used

google random number generator (31) to randomly pick out which video was to be used for IRR by both annotators. IRR performance was analyzed by Cohen's kappa coefficient (κ). All 14 unique main activities (appendix 1) were included in the IRR analysis.

2.4 NTNU HAR-model

The acceleration signals and annotations were handed over to the Department of Computer and Information Sciences (IDI) at NTNU for processing using the NTNU HAR-model. The NTNU HAR-model utilize a supervised machine learning approach, where the model observes input-output pairs to learn what determines an outcome (e.g. The model knows that a chosen 10sec period consist of running from the annotated data, and tries to determine rules in the acceleration signal to distinguish the activity from other known examples) (32).

Decisions are made by the model through a random forest (RF) method, which is a collection of multiple decision trees (DT) (33). A DT uses predetermined rules learned from supervised learning to determine the correct answer of a classification, regression, ranking or probability estimation problem (e.g. You have a glass of red liquid. The DT can answer wine or water. Because of a rule stating that water cannot be red, the DT determines the liquid to be wine). A vital element to make these methods viable is the signal features that are extracted from the acceleration signals. The NTNU HAR-model uses 138 features to explain characteristics of the acceleration signals, including time domain features like mean, standard deviation and range, as well as frequency domain features like amplitude statistics.

The RF decision trees are independent of each other, which means that we can utilize bagging (34) to give each tree a random subset of the data and combination of features. This creates diversity in the DT that are created, and if done right, usually leads to a more robust model (33) (e.g. With our earlier wine example, juice is introduced as an answer and positive alcoholic content is introduced as a feature. The original DT without knowing the additional feature answer 50% wine and 50% juice. A new DT that knows the additional feature answer 100% wine). In the end, a majority-voting approach is used to determine the final prediction. (e.g. With our wine example, if we have an even number of the different DTs in our RF, wine would be chosen as the final prediction with 75% of the votes).

2.5 Statistical analysis

Results from IDI were presented as a confusion matrix (appendix 2), which is a table that display every predicted activity in relation to its true label counterpart from annotation data. We define the distribution of classifications as:

1) True positive (TP) – Correct prediction of chosen activity.
2) False positive (FP) – Wrong prediction of chosen activity.
3) True Negative (TN) – Prediction of another activity when chosen activity is wrong.
4) False Negative (FN) – Prediction of another activity when chosen activity is correct.

After receiving the predictions back from IDI, the accuracy, sensitivity, precision, specificity and a combined total of the predictions in relation to our annotations were calculated using equations:

$$Accuracy = \frac{Correct\ predictions}{Total\ predictions} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{Correct\ positive\ predictions}{Correct\ positive + false\ negative\ predictions} = \frac{TP}{TP + FN}$$

$$Precision = \frac{Correct\ positive\ predictions}{False\ positive + correct\ positive\ predictions} = \frac{TP}{TP + FP}$$

$$Specificity = \frac{Correct\ negative\ predictions}{False\ positive + correct\ negative\ predictions} = \frac{TN}{FP + TN}$$

$$Total = Accuracy + Sensitivity + Precision + Specificity$$

Eq1: *Equation used during NTNU HAR-model performance analysis and detecting ballgames.*

Since the NTNU HAR-model gives predictions in 5sec windows, the true label from annotation is calculated as the most featured activity in the same 5sec windows. For ease of use and increased validity, some labels and predictions have been combined or excluded.
Running forward and backward from true labels has been combined into running to match predictions.
Shuffling has been labeled as standing, so all shuffling predictions has been added to standing predictions.
Stairs ascending and descending has been combined to stairs for both predictions and true label.
Lying prone, supine, right and left have all been included in lying for both predictions and true label.
Cycling sitting and standing has been combined to cycling for both predictions and true label.
Picking, vigorous activity, non-vigorous activity, transport and commute have been excluded from predictions because the category does not exist in labels.
Undefined, crabwalk, other activity and skipping have been excluded from labels because the category does not exist in predictions.
Heel drop have been excluded because the category is only usable as an indicator of start and end of annotation in the acceleration signal.

2.6 Detecting ballgame

Data manipulation was done in Excel (Microsoft Office 2016) by excluding all the predictions before and after the first heel drop on day one and the last heel drop on day two (signifies the start and end of data collection).

Using multiple for-loops in MATLAB (9.8.0 R2020a, Mathworks Inc., Natick, MA, USA), the predictions were then grouped in windows of 2-60min with 1min steps and returned the number of changes within each window (PAC). A window of 2 minutes includes 24 predictions as one prediction is 5sec. The windows where then blocked (added together) in every combination possible from 1 to 20 while never exceeding 60min total (e.g. 5min

window with 1 block gives a total of 5min, 2 block gives 10min … 12 block gives 60min. Or a 6min window with 1 block gives 6min … 10 block gives 60min). Where fractions are needed to reach 60min, the number of blocks were adjusted down to the nearest integer. Every combination of window size and blocking was then run through analysis with different thresholds for how many changes of activity a window*block combination needed for the window*block to be considered ballgame. The looped threshold values were calculated with a step-size of five to run from:

$$Lower\ limit = \frac{window\ size}{6} * block \rightarrow Upper\ limit = \frac{window\ size}{6} * 4 * block$$

As small changes were discovered with individually increasing the threshold by one, we picked out the five strongest values and repeated the calculation for every step between them. With each iteration we receive a 2x2 confusion matrix for each subject. These confusion matrices were then combined into a single 2x2 confusion matrix before analyzing the performance using Eq1.

The analysis was then repeated while incorporating an overlap with one and two elements. This was done by having the last one and two elements of the previous window copied to be the first one and two elements of the next window. The time variable was then adjusted as we are adding five and ten seconds to each window. Smaller tests with overlap from 10-50 with a step size of 10 was done in the end. The overlap never exceeded 20% of the window size.

3. Results

A total of 12824 predictions (17h 48min 40sek) covers the annotation work used in the initial NTNU HAR-model performance analysis from IDI (appendix 2), while 10064 predictions (13h 58min 40sek) was used after exclusion. In our analysis for detecting out ballgames, a total of 108763 predictions (151h 3min 35sek) was used with no exclusions.

3.1 Participation notes

All six participants completed the first handball practice, while subject 1007 did not participate in the second practice.
Participant 1008 had a broken finger and did alternative footwork drills during both practice sessions.
Participant 1009 and 1012 reported a 1h 15min handball practice with timestamps between practice one and two, while participant 1010 reported strength training right after school but without timestamps. The annotated strength training session is therefore an assumption based on school schedule, predicted activities and PAC.
All participants reported wearing the accelerometers for the entire duration of the study, which results in 0% sensor no-wear time.

3.2 Annotation

Amount of total video annotation done by each annotator was 18% for annotator one, 27% for annotator two and 55% for annotator three. Table 3.1 present the IRR between annotator one and three, and annotator two and three. Cohen's kappa coefficients were 0,95 and 0,92 respectively.

Table 3.1: Calculated Cohen's kappa, number of unique activities, and number of activity changes.

Annotator	κ-value	Unique activities	Activity changes
1 vs. 3	0.95	11 vs. 11	166 vs. 154
2 vs. 3	0.92	10 vs. 10	214 vs. 219

3.3 NTNU HAR-model

The NTNU HAR-model correctly classified 6437 of 12824 instances achieving an accuracy of 50% before exclusion, and the accuracy increased to 77% with 7798 of 10064 correct instances after exclusion. Statistical analysis of overall accuracy, sensitivity, precision and specificity for every included activity after exclusion is presented in table 3.2.

Table 3.2: Overall accuracy, sensitivity (Sens.), precision (Pre.), specificity (Spe.) and number of predictions (N) for every included activity.

Activity	Sens.	Pre.	Spe.	N
Walking	0.81	0.81	0.79	5319
Running	0.84	0.63	0.93	1669
Stairs	0.00	0.00	0.99	54
Standing	0.67	0.80	0.95	1824
Sitting	0.84	0.96	0.99	969
Lying	0.00	0.00	0.99	117
Transition	0.14	0.02	0.99	47
Bending	0.09	0.45	0.99	38
Cycling	0.00	0.00	0.99	27
Jumping	0.00	0.00	1.00	0
Accuracy	**0.77**			

Specificity was 0.99-1.00 for every activity with >1000 predictions, 0.95 for standing, 0.93 for running and 0.79 for walking. The sensitivity was >80% for walking, running and sitting, with standing following suit at 67%. While precision was >80% for walking, standing and sitting, with running following suit at 63%. Transition and bending had very few correct predictions (1 and 17) which is reflected in a very low sensitivity and precision. Stairs, lying and cycling had zero correct predictions, which in turn gave them 0% sensitivity and precision. Jumping was never predicted. The distribution of predictions is presented in table 3.3.

Table 3.3: Confusion matrix for the NTNU HAR-model predictions. The green cells represent the amount of correctly identified instances of each activity. Rows represent the labeled activities, and the columns represents predictions from the NTNU HAR-model. The column second most to the right is the total number of instances annotated in each category. The second most bottom row is the total number of instances detected by the model. The cells with red text represent the wrong positive (bottom) and negative (right) instances in each category. Excluded activities was not included in the table.

Annotated activity	Predicted activity											
	Walking	Running	Stairs	Standing	Sitting	Lying	Transition	Bending	Cycling	Jumping	Total	Wrong
Walking	4353	575	48	316	12	0	29	13	9	0	5355	1002
Running	191	1047	3	1	0	0	0	0	0	0	1242	195
Stairs	0	0	0	0	0	0	0	0	0	0	0	0
Standing	644	32	1	1454	11	0	7	8	5	0	2162	708
Sitting	33	1	0	21	926	114	10	0	3	0	1108	182
Lying	0	0	0	0	0	0	0	0	0	0	0	0
Transition	4	1	0	0	0	0	1	0	1	0	7	6
Bending	90	10	2	32	20	3	0	17	9	0	183	166
Cycling	0	0	0	0	0	0	0	0	0	0	0	0
Jumping	4	3	0	0	0	0	0	0	0	0	7	7
Total	5319	1669	54	1824	969	117	47	38	27	0	10064	N/A
Wrong	966	622	54	370	43	117	46	21	27	0	N/A	2266

3.4 Detecting ballgame

Table 3.4 show combined results with an even distribution of window, block, overlap and threshold combinations for detecting ballgames during our 151h 3min 35sek of possible predictions. The chosen combinations are based on overall performance. Accuracy proved to be excellent (>90%) for most combinations with a window*block size of more than 10min, while also staying above 90% with a resolution as high as 7min with 84*1 (window*block) combination. Sensitivity and precision changes drastically when adjusting the threshold, but the relationship between them is not linear. With 360*1 (window*block) and 2x overlap combination and a threshold of 175 instead of 200 (table 3.4), we get an accuracy of 95%, sensitivity of 94% and precision of 70%. If we increase the threshold to 225, we get an accuracy of 95%, sensitivity of 71% and precision stays at 80%.

Increasing the overlap up to 50 when increasing window size did not improve overall performance. For 360*1 (window*block) with 50 overlap, sensitivity saw an overall improvement of 8%, but accuracy, precision and specificity slightly decreased in performance by 1%, 7% and 1%.

Individual details with the combination 84*1 (window*block) and a threshold of 54 is presented in figure 3.1 and table 3.5. Accuracy proved to be excellent (>90%) for every participant except 1010 which scored 84%. Sensitivity varied from 58% to 80% while precision varied from 43% to 100%. Specificity was excellent (>90%) for every participant except 1010 who scored 85%. The system overestimated the amount of ballgame by 44% for participant 1007 and 79%* for participant 1010 and underestimated the amount of ballgame by 43%, 17%, 5% and 13% for participants 1008, 1009, 1011 and 1012. In total 71820sec was predicted by the system, with 71590sec being played.

*13% if counting strength training as ballgames.

Table 3.4: Combined results. Window size (Win), block size (Block), Time, threshold (Thr.), accuracy (Acc.), sensitivity (Sens.), precision (Pre.), specificity (Spe.) and Total statistics for detecting ballgame. Separated by number of overlaps and organized with guidelines by number of blocks.

		No overlap						1x overlap						2x overlap					
Win*block	Time	Thr.	Acc.	Sens.	Pre.	Spe.	Total	Thr.	Acc.	Sens.	Pre.	Spe.	Total	Thr.	Acc.	Sens.	Pre.	Spe.	Total
84*1	7min	56	0.90	0.61	0.58	0.94	3.04	54	0.91	0.70	0.64	0.94	3.19	54	0.92	0.69	0.68	0.95	3.25
120*1	10min	75	0.93	0.79	0.66	0.95	3.32	70	0.94	0.86	0.70	0.95	3.45	70	0.94	0.85	0.71	0.95	3.45
240*1	20min	150	0.95	0.80	0.75	0.96	3.46	145	0.94	0.80	0.73	0.96	3.44	150	0.95	0.76	0.78	0.97	3.46
360*1	30min	170	0.95	0.97	0.69	0.94	3.55	175	0.95	0.94	0.71	0.95	3.55	200	0.96	0.82	0.80	0.97	3.55
480*1	40min	265	0.97	0.95	0.76	0.97	3.65	275	0.97	0.90	0.78	0.97	3.62	268	0.98	0.95	0.83	0.98	3.73
30*4	10min	44	0.89	0.73	0.49	0.91	3.01	45	0.91	0.76	0.64	0.93	3.24	40	0.90	0.71	0.60	0.93	3.14
60*4	20min	95	0.94	0.76	0.71	0.96	3.37	85	0.94	0.88	0.72	0.95	3.49	92	0.94	0.73	0.75	0.97	3.38
90*4	30min	144	0.95	0.84	0.73	0.96	3.49	150	0.96	0.78	0.81	0.98	3.52	134	0.96	0.84	0.79	0.97	3.57
120*4	40min	184	0.97	0.92	0.86	0.98	3.73	175	0.97	0.92	0.83	0.97	3.69	180	0.95	0.81	0.78	0.97	3.51
15*8	10min	38	0.86	0.66	0.35	0.88	2.75	40	0.91	0.75	0.62	0.93	3.22	42	0.88	0.42	0.53	0.95	2.77
30*8	20min	80	0.91	0.75	0.56	0.93	3.15	70	0.93	0.92	0.68	0.93	3.47	80	0.91	0.65	0.67	0.95	3.18
45*8	30min	134	0.95	0.81	0.74	0.97	3.47	125	0.95	0.78	0.81	0.97	3.51	100	0.93	0.84	0.66	0.94	3.37
60*8	40min	160	0.95	0.95	0.69	0.96	3.55	170	0.97	0.92	0.83	0.98	3.70	162	0.96	0.85	0.85	0.98	3.64

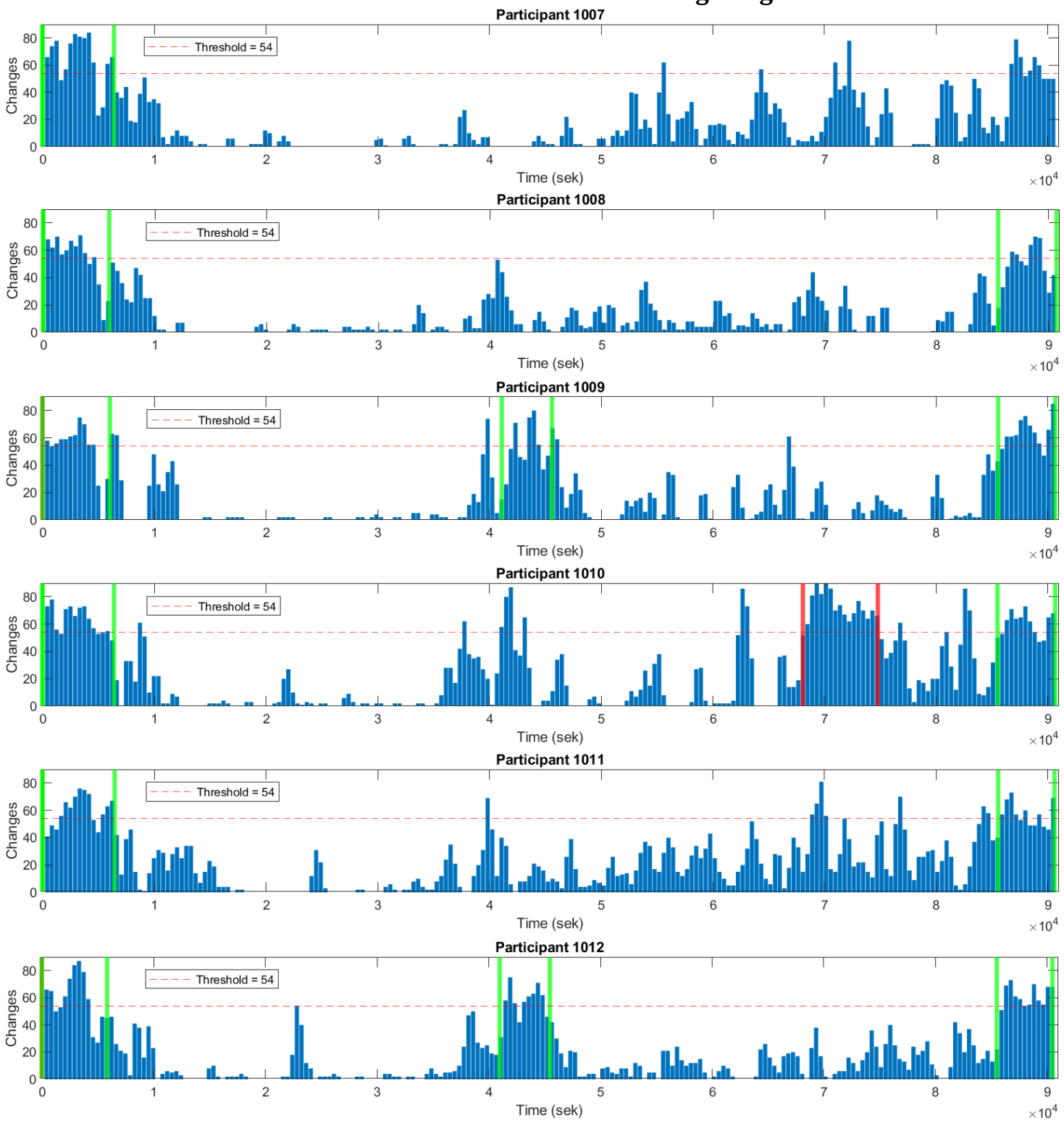


*Figure 3.1: Bar graph PAC per 84*5sec windows with 2x overlap. Space between green bars represent time spent during ballgames, while space between red bars represent time spent during strength training.*

Table 3.5: Individual statistics including window size (Win), block size (Block), threshold (Thr.), accuracy (Acc.), sensitivity (Sens.), precision (Pre.), specificity (Spe.), Total statistics and Predicted vs. Actual time in seconds and %-over or underestimation (%est.) of time spent detecting ballgame with 2x overlap.

Subject	Win*block	Thr.	Acc.	Sens.	Pre.	Spe.	Total	Pred./actual (sec)	%est.
1007	84*1	54	0.94	0.80	0.55	0.95	3.24	9240/6420	144
1008	84*1	54	0.95	0.58	1.00	1.00	3.53	6300/11120	57
1009	84*1	54	0.92	0.69	0.81	0.97	3.39	13020/15695	83
1010	84*1	54	0.84	0.78	0.43	0.85	2.90	20580/11510*	179*
1011	84*1	54	0.91	0.63	0.65	0.95	3.15	10920/11495	95
1012	84*1	54	0.94	0.71	0.96	0.99	3.61	11760/15350	77
Total	84*1	54	0.92	0.69	0.68	0.95	3.25	71820/71590	100

*6740sek spent during strength training: 20580/18250 – 113%.

4. Discussion

The aim of this study was to use handball as a paradigm to assess the validity of the NTNU HAR-model as a classifier during ballgames, and explore the possibility of using the NTNU HAR-model predictions to detect periods of ballgames based on PAC. The main results show that the NTNU HAR-models accuracy performed below the acceptable threshold of 80%, with large fluctuations between the different activities. Detecting ballgames solely through the NTNU HAR-model predictions is achievable, but with large variations in individual over- and underestimation.

4.1 NTNU HAR-model performance

The NTNU HAR-model achieved below acceptable results with an accuracy of 77%, even after exclusion of PA types as described in section 2.5 Statistical analysis. The exclusion process makes the results comparable to previous NTNU HAR-model studies. Reinsve (19) proved the same model to have an accuracy of 94% in adults during daily activities which is leaps and bounds above our results.

The accuracy of the NTNU HAR-model is dependent on how well the predictions match with the video annotation, so poor annotation would result in decreased accuracy. In this study however, the fact that three annotators collaborated during the annotation process can be used to pseudo validate the quality of annotations with the use of IRR. The IRR between annotator 3 and the other two annotators were very satisfying according to interpretations of Kappa values needing to equal or exceed 0.82 (35). It was not possible to calculate IRR between annotator 1 and 2, but we can predict that the IRR between annotator 1 and 2 should be higher than 1-(0,08+0,05)=0,87 considering both annotators had a Cohen's Kappa coefficient of 0,92 and 0,95 with annotator 3. This gives us an indication of having close to zero annotation bias and is a corner stone for future analysis. The IRR score and the fact that annotator 3 was not part of developing the activity definitions indicate the robustness of the definitions used in this study. A key strength of using the Kappa value to signify IRR is that the algorithm considers the possibility that the annotators were guessing during annotation, which eliminates possible overestimation with other methods like percent agreement (35, 36).

With the discussed support behind our annotations, closer analysis is needed to understand the accuracy of the NTNU HAR-model during ballgames. The model predicted most instances of walking (5319), standing (1824), and running (1669). Walking was the only activity out of the three with an acceptable sensitivity and precision of 81% each, but we also have the activity sitting with 969 instances reaching a sensitivity of 84% and a precision of 96%. We can see from the distribution of activities related to standing in table 3.3 and the low sensitivity in table 3.2 that the NTNU HAR-model slightly underestimates standing due to instances of walking. This can be influenced by shuffling being defined as standing during annotation, but it has probably more to do with the definition of walking. The definition of walking requires "locomotion toward a destination, one stride or more", but based on the annotations, movement back and forth without a clear destination happens often during ballgames. This can make the annotators interpret the movements as shuffling, while the acceleration signals and patterns are more consistent with the NTNU HAR-models training data for walking.

According to the distribution of activities in table 3.3 and the low precision in table 3.2, the NTNU HAR-model also slightly overestimates running due to instances of walking.

The can be influenced by the typical MVPA nature of ballgames (27-29, 37), where movements observed as walking could have features from the acceleration signal more reminiscent of running.

However, the largest impact in performance should most likely come from ballgames not being a good fit for the NTNU HAR-models' specifications. The current iteration of the NTNU HAR-model classifies which PA the participant is performing by calculating signal features from the raw acceleration data collected with two tri-axial accelerometers. This classification is a prediction with a window size of 5sec due to achieving increased accuracy during earlier testing with different datasets where participants mostly did one activity within the 5sec window (18, 19). However, during CPA like ballgames, the amount of information within a 5sec window can include more than the current NTNU HAR-model is able to process. During annotation we saw that our participants could do five different activities in a 5sec window, and 3 different activities in a 1sec window (figure 2.2). McInnes et. al (28) reported an average change in activity every 2sec during basketball, solidifying our results. Decreasing the window size of the NTNU HAR-model could therefore enable better detection and possibilities for distinguishing smaller bouts of different PA, which in turn should increase performance for CPA. This is something that should be tested in future studies, but revisions like these would have an impact on every aspect of the NTNU HAR-model program. Challenges with comparing to previous NTNU HAR-model studies, decreased performance during daily life, and incorporating HAR-additions like sensor no-wear time (21) and sleep-wake classification (17) would all need to be overcome. A better solution could be to detect periods where a shorter window size can bring beneficial results and run those periods with a second HAR-model designed for similar data characteristics.

4.2 Detecting ballgame

For detecting ballgame, our best results with a relatively high resolution of 7min windows is the combination with 84*1 (window*block), 54 threshold and 2x overlap that yielded an accuracy of 92%, a sensitivity of 69% and a precision 68%. As we will discuss later, even with excellent accuracy, sensitivity and precision should be our main priority. These values are considered good enough to give us a valuable indication of when ballgames were played but having both above 80% is preferable as acceptable standards to use in future research. Even though a 7min resolution provide below preferable results with our data, the resolution might be a better choice if all of our assumptions discussed in section "4.2.2 Analysis based on assumptions" are correct. A solution with a window*block size of 7min with 2x overlap should therefore not be immediately excluded as a future choice. Based purely on table 3.4 though, if we want both sensitivity and precision to exceed 80%, we need to use a lower resolution. Either the 360*1 (window*block) with 2x overlap that yielded an accuracy of 96%, a sensitivity of 82% and a precision of 80%, or one of our 40min windows that also yielded akin to or better results.

Increasing the overlap from x0, x1 and x2 until an overlap of up to x50 did not provide better overall performance, but rather increased overestimation of the system. This could be caused by having smoother transitions before and after periods with many PAC and could therefore be a possible solution for studies where underestimation becomes an issue.

The ability to detect ballgame from the existing NTNU HAR-model results provide scientists valuable information that can be used to determine periods of MVPA. We can see from table 3.2 that the existing NTNU HAR-model classified running only 17% of the

practice time and misclassified cycling 0,3%. These are the only possible MVPA periods classified, and researchers could only classify this period of 17,3% as MVPA if no other information is provided. Based on Póvoas et al. (27), elite handball players stayed above 60% of maximum HR for 93% of the time during matches. This indicates that ballgames should provide more than 17,3% time spent as MVPA and is supported by Leek et. al (30) where adolescents spent above 50% of soccer practice as MVPA.

One of the most valuable metrics for researchers from a method like detecting ballgames from PAC should be how much time the participant spends engaging in ballgames, which can be related to how much time the participants spends in MVPA. An accurate measurement can be achieved if we have an even distribution of sensitivity and precision. As we see in table 3.5, our chosen metric from table 3.4 with an even distribution have close to no difference in time spent during ballgames compared to the reported and labeled time. If we dig deeper though, our individual scores have great variability.

The individual statistics from figure 3.1 and table 3.5 give us great insight about our combined results in table 3.4. We can read from figure 3.1 that periods of ballgames have a lot more PAC than regular daily activities. However, we need a high threshold because multiple periods of daily activities also achieve high amounts of PAC. Without these periods, we should be able to lower our threshold without the effect of reducing our precision, which again would increase sensitivity and overall performance. If these periods of daily activity could be identified during data collection, we could look at a possibility to extend the model from detecting ballgame to e.g. detecting MVPA. This all depends on the results from data collection and is further explored in 4.2.2 analysis based on assumptions.

Another strength for researchers being able to detect ballgames, is the possibility to exclude these periods from the original NTNU HAR-model and use a more specialized HAR-model developed for CPA. As CPA includes small bursts of different PA, CPA can largely be compared to the activity pattern of children during play (38). With this, the possibilities for improvement and usability stretches even further, but it all depends on the detecting ballgames model being able to detect the correct periods, and how well a new HAR-model would perform.

Our approach to detect periods of ballgames relies on how often the NTNU HAR-model predicts the participant switching from one activity to another. We do not put different weights on different activities as data from other studies, especially the dataset Trondheim Free Living used in earlier NTNU HAR studies, suggest daily activities does not include as many PAC as ballgames has proven to include (19, 26, 28). But we still need to keep this in mind, as changing between sitting and lying count the same as changing between walking and running.

We had relatively few participants (six) in this study, but the amount of data is still quite large. This makes it hard for a human to calculate the most optimal window and block size. The solution was to give this task to a computer that can systematically calculate every combination within our specified limits. We chose a window of 2min – 60min with blocking from 1 – 20 while never exceeding the 60min maximal combined window*block size.

The reason we start at 2min is based on a trail where 2min is the preferred window size for the most accurate amount of changes during ballgames in relation to vigorous PA measured by pulse. This concept will be further discussed in 4.3 future research. The reason we end at 60min as our maximum combined window*block size is because a larger

window*block size would make smaller periods of ballgames harder to distinguish from daily activities, and overestimation could become prevalent if the period is detected.

The optimal lower and upper threshold limits were chosen after early probing resulted in every single optimal threshold value being found within these limits. Because of a small probing size, the possibility of having to run the script a second time with different limits were evaluated on every combination before choosing the five strongest values. This was done instead of trusting probability analysis like confidence interval, as the time saved with probability analysis was minor. The only reason for choosing limits like this was to reduce run-speed in MATLAB. Another solution to increase run-speed was to combine the six 2x2 individual confusion matrices before running analysis, and only run individual analysis on chosen examples.

4.2.1 Statistical understanding

As with the NTNU HAR-model's performance discussed in section 4.1, using accuracy, sensitivity, precision and specificity are great measurements of performance for detecting ballgames.

Accuracy measures the selection of correct time-sensitive predictions. A correct prediction results in the system (our combination of window size, block size, threshold and overlap) predicting whether the participant is engaging in ballgames at a specific point in time. This translates to accuracy being a definition of the number of windows where the system guessed the right activity at the right time, both positive and negative, divided by the total amount of guesses. It is our most important parameter but might need to be higher than other studies because most of the 25h duration of our study is spent far below our threshold (e.g. sedentary while sleeping). This makes the amount of correct negative predictions skyrocket during the night, which pushes accuracy higher compared to studies where a higher amount of predictions has closer to equal chance to be both positive and negative. We should still consider our >90% score on most combinations in table 3.4 and table 3.5 to be an excellent indication of the potential behind detecting ballgames based on PAC (23).

Sensitivity measures the portion of correct positive predictions during the time period labeled as ballgames. Since sensitivity only includes instances where changes should be above our ballgames threshold, this parameter will tell us how well the system can predict windows that should, according to our hypothesis, include a high amount of PAC. If the threshold is too low, sensitivity will approach 0%. If the threshold is too high, sensitivity will approach 100%. Sensitivity is one of our most important measurements, but only combined with precision to tell a complete story.

Precision measures the portion of positive predictions that are correctly classified. A high threshold will make precision approach 100%, while a low threshold makes precision approach 0%. Combining sensitivity and precision paints a more complete picture of our results because sensitivity utilizes false negatives while precision utilizes false positives. If precision is low while sensitivity is high, we can conclude that the systems predictions are skewed toward overestimating instances of ballgames. An underestimation of ballgames will occur if the opposite happens. Using sensitivity and precision, we can see that there are big differences in the different combinations presented in table 3.4. During analysis, manipulating sensitivity and precision using different thresholds provided big changes while not having a large impact on our accuracy. This supports our discussion earlier about the

vast amount of negative instances compared to positive instances having an impact on our accuracy.

Specificity measures the portion of negative predictions that are correctly classified. With a low threshold, specificity approaches 100% because false positive predictions approach 0. This value will be a lot less sensitive if the relationship between time spent during ballgames is far more or less than the time spent not engaging in ballgames. This is evident in our results, as specificity constantly stays as the best performing statistic. In our study, sensitivity and precision are more significant in answering our study aim. This is reflected in specificity behaving the same way as accuracy discussed earlier.

This thesis implements a **total** statistic. The optimal result for every statistic is 100% which equates to 1.0 for the four previously mentioned statistics in section 4.2.1 Statistical understanding. This makes it easy to use our total statistic as an overview of total performance where 100% equates to 4.0. During analysis, accuracy was favored as the most important statistic to select the best performing combination of adjustable factors. Having a total statistic as a secondary value of performance helped us faster understand overall performance as the analysis results were processed manually. We could also add weights to the statistics depending on their importance for our goal but chose not to do so because the total is only used as a different representation of our results.

Resolution can be explained as how many times we can predict if the participants are engaging in ballgames. Decreasing the window size increases resolution, while the opposite is also true. With a total collection time of approximately 25h 7min per person, our resolution with a 7min window will become 215 predictions. Multiplying this with 6 participants gives us 1290 predictions in total. If we increase the window size to 40min we only get 220 predictions in total. By adding elements through overlap, we will increase the predictions of our 7min window by 30 with 2x overlap. Having a high resolution decreases the amount of possible changes within the given timeframe. This in turn has the natural effect of reducing the possible difference in changes between ballgames and other daily activities. Decreasing the window size is positive because we can more accurately predict the start, end, and total time spent during ballgames. If there are longer breaks as well, a higher resolution will let us classify the active parts before and after as ballgames, and the break as non-ballgames. With a lower resolution, depending on the length of the break and the number of changes in the period before and after the break, the break can be classified as ballgame, or the period before and/or after the break can be classified as non-ballgame. A limitation of reduced resolution includes the increased possibility of short bouts of other activities being classified as ballgames, and parts of a ballgame session that does not include as many PAC can be classified as non-ballgames. Because of this "give and take" nature of high and low resolution, choosing a correct balance depends on the study and research question. With a study where the aim is to estimate the number of practices/matches of a certain length is desirable, choosing a low resolution will provide more accurate results. If the study is researching the time of day, or amount of time spent during ballgames, then a high resolution is preferable.

Changing the window size of the NTNU HAR-model should have a similar impact as changing the resolution has on detecting ballgames, and the give and take relationship that resolution have should be largely mitigated. Decreasing the window size of the NTNU HAR-model would most likely increase the amount of PAC for every non-static PA, but it should also widen the gap between the amount of PAC predicted between daily activities and CPA

(19, 26, 28). This would in return increase our accuracy when detecting ballgames, even if our previous assumption about an increased accuracy of the NTNU HAR-model would be wrong. If this proves to be correct, adjusting the window size of the NTNU HAR-model might be preferrable depending on the study aim.

4.2.2 Analysis based on assumptions

Looking back on the information we have gathered, some assumptions could explain parts of our results.

Participant 1007 reported not participating in practice on day two of the study, but we can assume that he engaged in some physical activity based on the knowledge we have acquired from the predicted activities and PAC in figure 3.1. Overestimation for participant 1007 could possibly be explained based by this assumption.

Participant 1008 had a broken finger and did alternative footwork drills for both practice sessions. This can explain the underestimation of our system as the practice drills where repetitive with less complex movements.

Participant 1010 reported a dynamic strength training session after school. The NTNU HAR-model predicted a lot of changes between activities during this period, so we can assume the participant was almost always moving and changing positions during this time.

The threshold could be lowered to achieve better overall performance without these outliers, but we should consider the outliers to be representative incidents for our target population based on these results. Further research with a larger study population is needed before we can make other assumptions.

The outliers can also indicate that the NTNU HAR-model does not need a regular CPA to predict a lot of different activities, possibly because the window size of the NTNU HAR-model is 5sec. This in turn leads us to consider a new definition of CPA tailored for the NTNU HAR-models 5sec windows, as changing movements more often than every 5sec does not make a difference in the model. We can then assume that more activities than initially thought could be indistinguishable from ballgames. Our strength training session is solid evidence for this, but it is impossible to conclude anything without detailed information about the session and more test-data. What could be concerning are PA not moderate to vigorous in nature being indistinguishable to ballgames. Our study population were six adolescent boys over approximately 25h 7min during winter, so we can assume that we do not have any data including activities like light housework or garden work. If these activities would be recognized as ballgames by our system, then that is a big limitation.

4.3 Future research

Based on our assumptions about detecting ballgames, future studies can use already existing annotated data from training and validating the NTNU HAR-model to further explore limitations including house and garden work with a larger study population.

Exploring different feature sets and window sizes within the NTNU HAR-model based on data using CPA might also provide different opportunities. This includes potentially increased performance on children, detecting ballgames, and new analysis opportunities during ballgames.

During this study, a third aim to determine if the participants were performing MVPA during ballgames was roughly explored. The method uses PAC in a similar fashion as detecting ballgames from the NTNU HAR-model predictions. To validate if the participant

was engaging in MVPA, the collected HR-data was used. As HR-data comes from a delayed physical response, the synchronization was only visually performed to fit the model while validating with annotation video in ANVIL. No protocol was developed but using a more documented best-fit method for each participant might be a viable solution. Early results suggest an accuracy of around 80% when using 2min window sizes without overlap on the annotated data, and an accuracy of around 50% on the predicted data from the NTNU HAR-model. More interesting is the fact that the predicted amount of time spent in MVPA during the practice should come within 20% of the actual time. Participant 1010 and 1012 spent 70% and 46% of the first practice in MVPA according to HR-data, while using NTNU HAR-predictions and PAC we achieved 65% and 56%. This can be used as proof of concept for further research.

4.4 Strengths and limitations

Some strengths and limitations of this study have already been discussed, but there are more that should be thoroughly explored. To my knowledge, this is the first study to explore the characteristics of PAC from HAR-model predictions to classify periods of different activity types.

With only 6 male participants of approximately the same age and body size our group is very homogenous. This can be both positive and negative as our results are less influenced by individual differences within a population. While it is impossible to compare our results directly with a general population, the accuracy for our target group will have increased validity. We could also use data from different studies together with our results to form a well-documented hypothesis for other population groups. Although a challenge of comparing our results with other studies are a lack of studies using CPA or PAC comparable to ours. As the NTNU HAR-model is originally developed for use in HUNT4, we should consider the importance of reaching the highest accuracy in the largest target population group. Young Norwegian men are according to Statistics Norway (39) the most active group engaging in ballgames.

More participants would give us the opportunity to make random groups when processing data, which would make it possible to create a threshold for one group and validate with another. Right now, we are using the entire study population to create our threshold, while doing the validation individually.

As we have unreported periods of high amount of PAC on 2/6 participants, the accuracy and sensitivity could be higher with a larger study population or a different study protocol that made the participants record every case of PA. We also have a period of strength training for one participant that was classified as ballgames by our detecting ballgame model. Based on the unrecorded periods and our strength training period, we can estimate that our system will pick out any period of multiple activities done in short succession, not just what is defined as CPA. This might be mitigated by shortening our prediction window in the NTNU HAR-model as discussed earlier, because it should provide us with a larger discrepancy of PAC between ballgames and daily activities.

Unfortunately, some problems occurred with synchronization between acceleration signals and annotations during testing of the NTNU HAR-model at IDI. This resulted in the researchers having to manually synchronize the data based on timestamps from the accelerometers, HR-watches, GoPro-files, and annotation timeline in ANVIL. This was done

by researchers with experience in using this method, but it is still possible that it could cause a change in performance.

A strength of this study is that the NTNU HAR-model is open source and uses raw acceleration signals. This ensures full transparency and provide opportunities for other researchers to confidently compare results.

Another strength is the use of objective measurements from an accurate method of observation with video recordings (40) in combination with robust activity definitions (appendix 1) as the groundwork for NTNU HAR-model validation.

4.5 Conclusion

The NTNU HAR-model in its current state does not provide us with a valid tool to predict activity types during ballgames. It is therefore recommended to use a different HAR-model approach or altered classification method of the NTNU HAR-model if the target aim is to predict activity types during ballgames.

Based on the results from this thesis, we can use PAC calculated from predictions generated by the NTNU HAR-model to detect periods of ballgames. However, this method should be tested on a larger study population including other moderate to vigorous physical activities before used in future research.

References

1. Warburton DE, Nicol CW, Bredin SS. Health benefits of physical activity: the evidence. Cmaj. 2006;174(6):801-9.
2. Reiner M, Niermann C, Jekauc D, Woll A. Long-term health benefits of physical activity–a systematic review of longitudinal studies. BMC public health. 2013;13(1):1-9.
3. Lee I-M, Shiroma EJ, Lobelo F, Puska P, Blair SN, Katzmarzyk PT, et al. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. The lancet. 2012;380(9838):219-29.
4. Organization WH. Physical activity: WHO; 2020 [Available from: https://www.who.int/news-room/fact-sheets/detail/physical-activity.
5. Piercy KL, Troiano RP, Ballard RM, Carlson SA, Fulton JE, Galuska DA, et al. The physical activity guidelines for Americans. Jama. 2018;320(19):2020-8.
6. Organization WH. Global recommendations on physical activity for health: World Health Organization; 2010 [Available from: http://apps.who.int/iris/bitstream/10665/44399/1/9789241599979_eng.pdf.
7. Hallal PC, Andersen LB, Bull FC, Guthold R, Haskell W, Ekelund U, et al. Global physical activity levels: surveillance progress, pitfalls, and prospects. The lancet. 2012;380(9838):247-57.
8. Hammami A, Harrabi B, Mohr M, Krustrup P. Physical activity and coronavirus disease 2019 (COVID-19): specific recommendations for home-based physical training. Managing Sport and Leisure. 2020:1-6.
9. Hall G, Laddu DR, Phillips SA, Lavie CJ, Arena R. A tale of two pandemics: How will COVID-19 and global trends in physical inactivity and sedentary behavior affect one another? Progress in Cardiovascular Diseases. 2020.
10. Hawkley LC, Thisted RA, Cacioppo JT. Loneliness predicts reduced physical activity: cross-sectional & longitudinal analyses. Health Psychology. 2009;28(3):354.
11. Bonomi AG, Goris AH, Yin B, Westerterp KR. Detection of type, duration, and intensity of physical activity using an accelerometer. Medicine & Science in Sports & Exercise. 2009;41(9):1770-7.
12. Warren JM, Ekelund U, Besson H, Mezzani A, Geladas N, Vanhees L. Assessment of physical activity–a review of methodologies with reference to epidemiological research: a report of the exercise physiology section of the European Association of Cardiovascular Prevention and Rehabilitation. European Journal of Cardiovascular Prevention & Rehabilitation. 2010;17(2):127-39.
13. Troiano RP, Berrigan D, Dodd KW, Masse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. Medicine and science in sports and exercise. 2008;40(1):181.
14. Kavanaugh K, Moore JB, Hibbett LJ, Kaczynski AT. Correlates of subjectively and objectively measured physical activity in young adolescents. Journal of Sport and Health Science. 2015;4(3):222-7.
15. Troiano RP, McClain JJ, Brychta RJ, Chen KY. Evolution of accelerometer methods for physical activity research. British journal of sports medicine. 2014;48(13):1019-23.
16. Bassett Jr DR, Rowlands AV, Trost SG. Calibration and validation of wearable monitors. Medicine and science in sports and exercise. 2012;44(1 Suppl 1):S32.
17. Hay A. Machine learning methods for sleep-wake classification using two body-worn accelerometers: NTNU; 2019.
18. Hessen H-O, Tessem AJ. Human activity recognition with two body-worn accelerometer sensors: NTNU; 2016.
19. Reinsve Ø. Data analytics for hunt: Recognition of physical activity on sensor data streams: NTNU; 2018.
20. Vågeskar E. Activity recognition for stroke patients: NTNU; 2017.

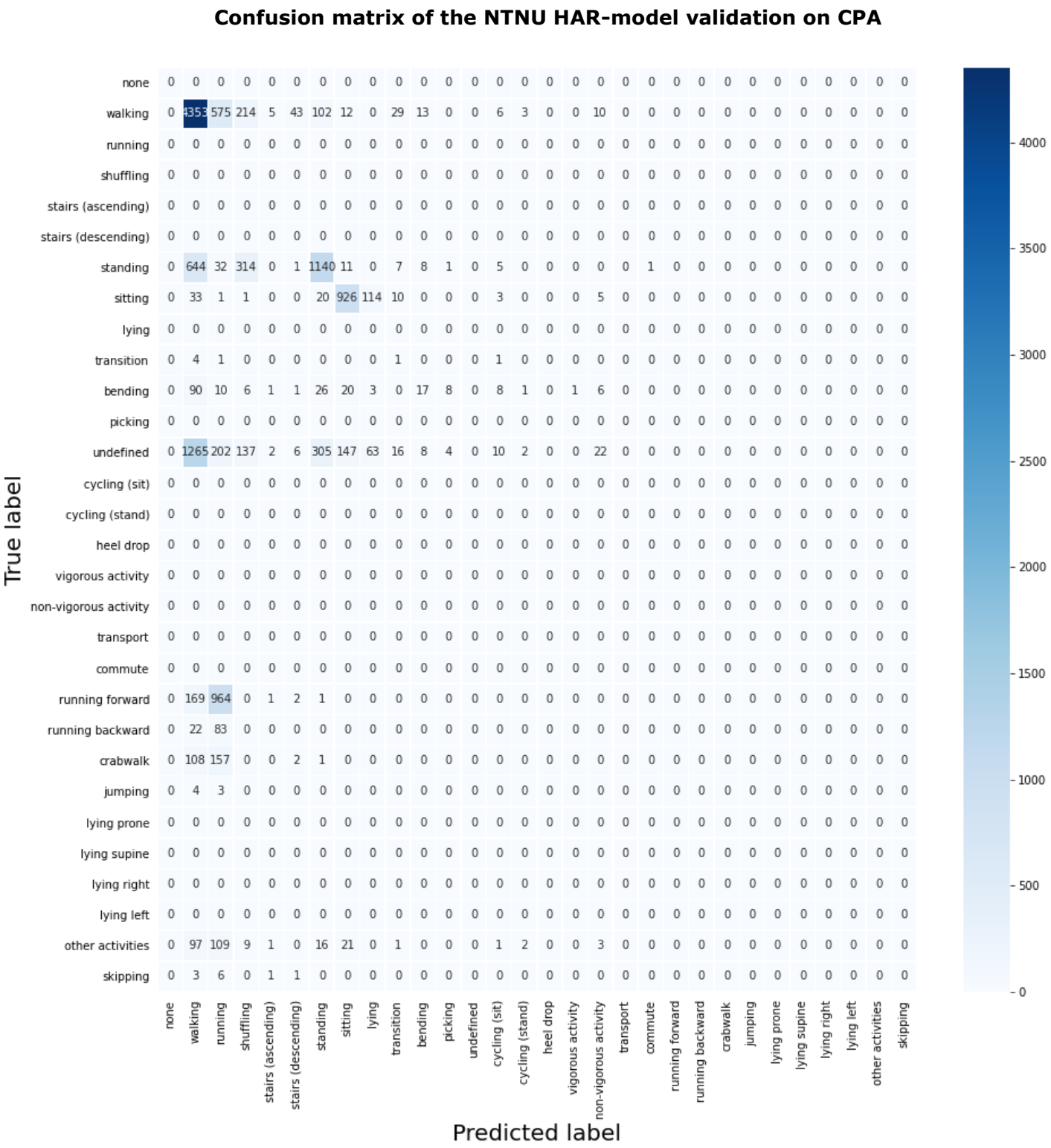
21. Wold T, Skaugvoll SAE. Ensemble Classifier Managing Uncertainty in Accelerometer Data within Human Activity Recognition Systems: NTNU; 2019.
22. Rowlands AV, Olds TS, Hillsdon M, Pulsford R, Hurst TL, Eston RG, et al. Assessing sedentary behavior with the GENEActiv: introducing the sedentary sphere: Lippincott Williams & Wilkins New York; 2014.
23. Trost SG, Zheng Y, Wong W-K. Machine learning for activity recognition: hip versus wrist data. Physiological measurement. 2014;35(11):2183.
24. NTNU. HUNT4 Norwegian University of Science and Technology 2020 [Available from: https://www.ntnu.no/hunt/hunt4.
25. Becker DR, Grist CL, Caudle LA, Watson MK. Complex Physical Activities, Outdoor Play, and School Readiness among Preschoolers. Global Education Review. 2018;5(2):110-22.
26. Østvang TK. A pilot study to explore the composition of complex physical activity in adolescents: NTNU; 2018.
27. Póvoas SC, Seabra AF, Ascensão AA, Magalhães J, Soares JM, Rebelo AN. Physical and physiological demands of elite team handball. The Journal of Strength & Conditioning Research. 2012;26(12):3365-75.
28. McInnes S, Carlson J, Jones C, McKenna M. The physiological load imposed on basketball players during competition. Journal of sports sciences. 1995;13(5):387-97.
29. Fröberg A, Raustorp A, Pagels P, Larsson C, Boldemann C. Levels of physical activity during physical education lessons in Sweden. Acta Paediatrica. 2017;106(1):135-41.
30. Leek D, Carlson JA, Cain KL, Henrichon S, Rosenberg D, Patrick K, et al. Physical activity during youth sports practices. Archives of pediatrics & adolescent medicine. 2011;165(4):294-9.
31. Google. Random number generator Google 2020 [Available from: https://www.google.com/search?q=random+number.
32. Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering. 2007;160(1):3-24.
33. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.
34. Breiman L. Bagging predictors. Machine learning. 1996;24(2):123-40.
35. Landis JR, Koch GG. The measurement of observer agreement for categorical data. biometrics. 1977:159-74.
36. McHugh ML. Interrater reliability: the kappa statistic. Biochemia medica: Biochemia medica. 2012;22(3):276-82.
37. Di Salvo V, Gregson W, Atkinson G, Tordoff P, Drust B. Analysis of high intensity activity in Premier League soccer. International journal of sports medicine. 2009;30(03):205-12.
38. Bailey RC, Olson J, Pepper SL, Porszasz J, Barstow TJ, Cooper DM. The level and tempo of children's physical activities: an observational study. Medicine and science in sports and exercise. 1995;27(7):1033-41.
39. SSB. Vi trener mer enn før Statistisk sentralbyrå 2016 [updated 23.11.2016. Available from: https://www.ssb.no/kultur-og-fritid/artikler-og-publikasjoner/vi-trener-mer-enn-for.
40. Aminian K, Robert P, Buchser E, Rutschmann B, Hayoz D, Depairon M. Physical activity monitoring based on accelerometry: validation and comparison with video observation. Medical & biological engineering & computing. 1999;37(3):304-8.

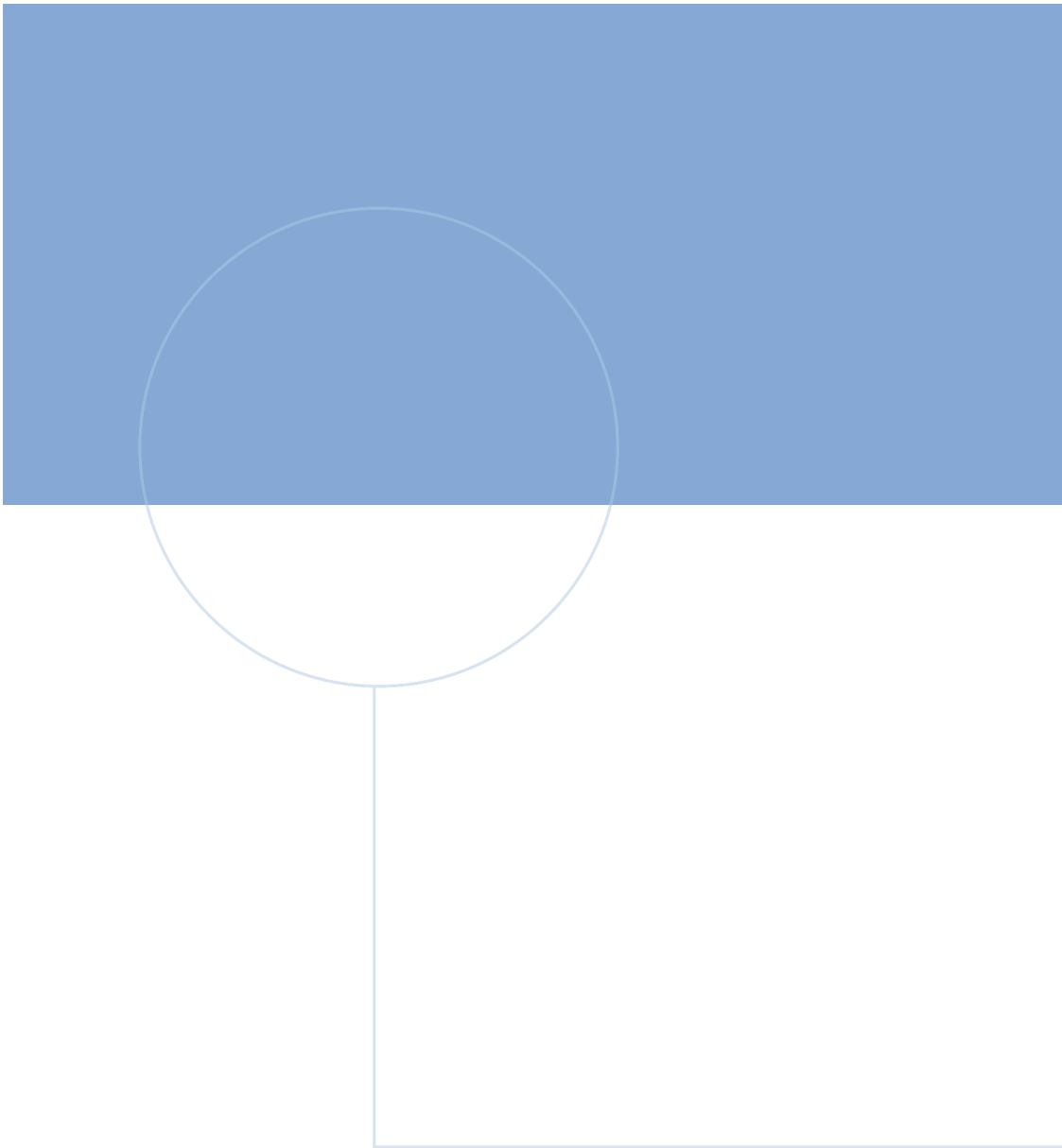
Appendix 1

DEFINITION OF ACTIVITIES	
Activity	**Description**
Sitting	When the person's buttocks are on the seat of the chair, bed or floor. Sitting can include some movement in the upper body and legs; this should not be tagged as a separate transition. Adjustment of sitting position is allowed.
Standing	Upright, feet supporting the person's body weight, some feet movement is allowed («on the spot", no substantial displacement). If both feet are lifted from the ground, another activity is inferred. Movement of upper body and arms is allowed until forward tilt and arm movement occurs below knee height. Then this should be inferred as bending.
Walking	Locomotion towards a destination, one stride or more, where both feet are lifted from the ground. Walking along a curved line is allowed. Walking can occur in all directions and with a ball. As soon as heel-off occurs, walking has started and ends when both feet are at rest or another activity is inferred.
Running	**Forward:** The movement starts when the person lifts one foot from the ground, with locomotion towards a destination in a forward direction. **Backwards/partly sideways:** The movement starts when the person lifts one foot from the ground, with locomotion towards a destination in a backwards/sideways direction. Running along a curved line is allowed. Running can be with a ball. Running ends when both feet are at rest or another activity is inferred.
Lying down	The person lies down. Adjustment after lying down is allowed if it does not lead to a change between the prone, supine, right and left lying positons. Movement of arms and head is allowed. Movement of the feet is allowed as long as it does not lead to change in posture. **Prone:** On the stomach. **Supine:** On the back. **Right side:** On right shoulder. **Left side:** On left shoulder.
Bending	Bending towards something below knee-height is tagged as bending. Steps can occur during bending. Bending ends when another activity is inferred. Bending while sitting is tagged as sitting.
Crab walking	The movement starts when the person lifts one foot sideways or backwards, with locomotion towards a destination with at least two steps. Center of gravity is lower than during walking/running, feet move at a higher speed. Feet do not necessarily leave the ground. Can occur in all directions. Crab walking ends when both feet are at rest or another activity is inferred.
Sit Cycling	Pedaling while the buttocks are placed at the seat. Cycling starts on first pedaling and finishes when pedaling ends. **For outdoor bicycling:** Cycling starts at first pedaling, or when both feet have left the ground. Cycling ends when the first foot is in contact with the ground. **Not pedaling:** Sitting without pedaling should be tagged separate as sitting.
Stand cycling	Pedaling while standing. Cycling starts on first pedaling and finishes when pedaling ends. Standing without pedaling should be tagged separate as standing.

Jumping	A bounce from the ground into the air, where both feet leave the ground. Jumping can occur from one or both feet, and in vertical and horizontal direction. The movement starts when the person`s last foot, or both feet simultaneously, leave the ground. Jumping ends when another activity is inferred.
Skipping	Step forward and jump on the same foot, with at least two steps where both feet leave the ground during each stride.
Other activities	All movements that are recognizable, but do not classify according to the definitions. This could be goalkeeper movements, falling, push-ups, hand stand etc., or activities that are uncontrolled or unintended.
Undefined	Periods until all the sensors are attached, or final adjustment made to position the video camera, can be tagged as undefined. All postures/movements that cannot be clearly identified due to blocking of the camera/view should be tagged as undefined.
TRANSITIONS	
Transition	Change from one movement to another, the period between movements should be tagged as a transition if this period is controlled and/or intended.
Transitions that will be tagged separately as a transition/undefined	
Upright to sitting	Can be from walking, running, crab walking or standing, as soon as forward trunk tilt occurs, or a lowering of the trunk, the transition has started. Steps can occur during the transition for positioning. Transition ends when buttocks are in contact with the seat of the chair, bed or floor.
Sitting to upright	Transition starts when the person's buttocks leave the chair and ends when the trunk has reached its upright position. Steps and turning can occur during the transition from sitting to upright. Can be followed by standing, walking, crab walking or running.
Upright to lying	Can be from walking, running, crab walking or standing. When the trunk flexion begins, or a lowering of the center of mass, the transition has started. Transition finishes when the person is lying flat with the trunk in a stable position.
Lying to upright	While lying, the transition begins with an upward movement of the trunk or leg movement that leads to a stable upright position or continuous walking. The trunk angle should be in a steady posture for the transition to finish. Steps can occur during the transition.
Transitions that will not be tagged separately as transitions	
Transitions between the activities walking, running, crab walking, jumping and standing	Switching between these activities can occur directly and should not be tagged as a transition. The current activity switches directly into the subsequent activity.
Jumping to sitting/lying	The movement ends when the body is in recline- or sitting position.
Sit cycling to stand cycling / stand cycling to sit cycling	When the buttocks leave the seat, stand cycling can be inferred. When the buttocks are placed at the seat, sit cycling can be inferred.

Appendix 2


Confusion matrix of the NTNU HAR-model validation on CPA
True label
Predicted label
none
walking
running
shuffling
stairs (ascending)
stairs (descending)
standing
sitting
lying
transition
bending
picking
undefined
cycling (sit)
cycling (stand)
heel drop
vigorous activity
non-vigorous activity
transport
commute
running forward
running backward
crabwalk
jumping
lying prone
lying supine
lying right
lying left
other activities
skipping
4000
3500
3000
2500
2000
1500
1000
500
0



NTNU

Norwegian University of
Science and Technology