Detecting Drinking Behavior in Social Settings Using Chest-Mounted Accelerometer Data

< Thomas Baeten[1]>
Supervisor(s): <Hayley Hung[1]>, <Litian Li, Stephanie Tan [1]>
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: <Thomas Baeten>
Final project course: CSE3000 Research Project
Thesis committee: <Hayley Hung>, <Litian Li, Stephanie Tan>, <Julian Urbano Merino>

Abstract

This study investigates the feasibility of detecting drinking behavior in social environments using chest-mounted accelerometer data. A dataset collected during a conference is used, consisting of accelerometer data and annotated video recordings of 48 participants. After preprocessing, a random forest classifier is trained using statistical features: mean, correlation, maximum, minimum, and covariance derived from the y- and z-axes of acceleration data. Evaluation through sixfold cross-validation for one participant yields an accuracy of 79%, while a leave-one-out cross-validation across participants achieves 70% accuracy. Qualitative analysis of false predictions reveals that actions like nodding, walking while drinking, or movement of the drinking hand towards the face can mimic drinking behavior. These findings demonstrate that accelerometer data contains detectable signals of drinking behavior even in noisy real-world conditions. However, further improvements require more diverse training data, consistent annotation, and possibly the inclusion of additional movement categories. The results support the potential of wearable accelerometers for drinking monitoring in social settings.

1 Introduction

Understanding noisy interactions is a particularly difficult task for intelligent systems. The field of behavioral science in social interaction could benefit from more research on uncontrolled interactions between people. To find how this would differ from controlled experiments, the methods used in controlled experiments are applied to drinking behavior during social gatherings. An experiment has been conducted where a group of people were measured using an accelerometer in a social interaction.[1] This accelerometer was close to the heart, because people normally wear conference badges around their neck during conferences, and this makes the experiments reproducible. An accelerometer is more reliable than radio-based tracking systems, such as ultra-wideband, in busy places due to signal disturbances from interference, reflections, and weakening.[2] With this experiment, the participants were already annotated when they were drinking, which makes the experiment useful to apply multimedia analysis and supervised machine learning techniques. How can this data be processed to find useful results? Looking at earlier research on different experimental setups using accelerometer data may provide insight. Looking at earlier research on different experimental setups using accelerometer data may provide insight.

Currently, there is already some research on accelerometer data, which can be relevant for research. An example of relevant research is an experiment that uses accelerometers in watches to detect different types of activity. [3] Another experiment testing the drinking behavior of cows with an accelerometer close to the neck. [4]

This previous work matches in questioning whether an accelerometer is a feasible method to detect some type of activity or, in the second case, even drinking. The biggest difference is that the experimental setup is way more optimal than when done in an interaction. Although this is influential in the outcome, some of the methodology used in previous studies can also be used to draw a conclusion.

The main goal of this research will be to show how implementing a model on accelerometer data, while being in the suboptimal environment of measuring data, which is a conference in this case, can predict drinking. The results of this study can later become relevant in behavioral studies or can be used for similar experiments.

This leads to the following research question:

How accurately can chest-mounted accelerometer data detect drinking events in natural social environments based on a random forest model?

To give guidance to answer the research questions, the following sub-questions are derived:

- How well does a machine learning model perform in detecting drinking for a single individual?
- How well does a machine learning model perform in detecting drinking between multiple participants?
- In what cases does the machine learning model fail, and how can this be explained by studying video material?

In Section 2, the dataset will be explained in more detail. After this, in Section 3, the previous work is studied. Then, in Section 4, the methodology is explained. In Section 5, the results of the implementation of this methodology are shown, which are discussed in Section 6. In the end, the paper is concluded in Section 8.

2 Dataset

The conflab dataset[1] is a dataset in a conference setting. The dataset contains data from 48 participants. The participants are studied with regard to their drinking behavior. The annotation is done by 1 to 3 different annotators per segment of 2 minutes. Furthermore, for every participant, accelerometer data is available in the x-direction, y-direction, and z-direction. The dataset also contains the 2-minute segments as top-down, muted videos.

2.1 Accelerometer

The definition of an accelerometer helps to understand what the measurements exactly do, and understanding the directions of the accelerometer can be helpful later for feature extraction. ”Accelerometers are devices that measure acceleration and convert it to an electrical signal. The output is commonly in units of gravity (g)”.[5] Understanding direction gives an idea of how helpful a single direction can be in identifying drinking. Figure 1 shows what the directions look like for the rectangular plane in the chest.

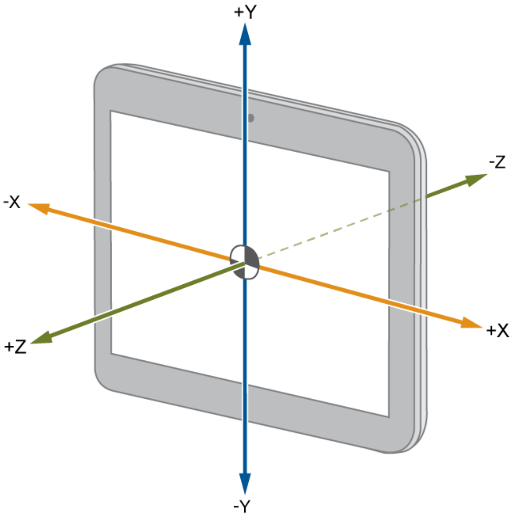

Figure 1: Accelerometer in a rectangular plane that is connected to the chest. The x-direction tracks movement to the left and the right, the y-direction tracks upward and downward movement, and the z-direction tracks forward and backward movement. (From [6])

This means that the x-direction is horizontal movement or movement to the left and to the right. The y direction is vertical movement or upward and downward movement. The z-direction shows forward and backward movement.

2.2 Annotations

An example of how 3 different annotators annotate on a segment of 2 minutes for 1 participant helps to give insights into how annotations will be handled. An example is shown in Figure 2.

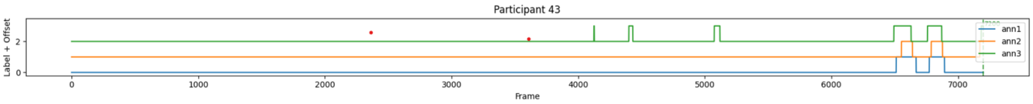

Figure 2: Annotation for drinking behavior of participant 43.

This example shows two cases that are important for the choices in the method later.

The first thing that stands out is that between around 4200 and 5200 frames, annotator 3 finds three moments where the participant is drinking, and the other two annotators do not recognize these drinking moments. In this case, the video shows that the participant does raise their glass but does not sip from their glass. Situations where the annotators do not agree on whether a participant is drinking are studied in the video material because this is something that does not happen often it is feasible to study. Later in Section 4, a judgment is made for each of these situations. In this case of not sipping, as in figure 2, the moments are not seen as drinking. However, in other cases, some annotators do not notice some subtle drinking behavior that other annotators do notice, and the moments are seen as drinking.

When all three annotators annotate something as drinking, there is a difference in the start and end frames of the annotations. This has 2 main reasons. The first reason is the delay caused by human annotation. Humans are nowhere near perfect and cannot give a frame on the dot when someone starts or finishes drinking. Another reason is that there is a different perception of when someone starts or finishes drinking. Studying the videos shows that some annotators start their annotation when somebody raises their glass, and other annotators start their annotation when a participant starts sipping. A voting system must be chosen to find positive samples. This is done later in Section 4.

3 Related Work

Two different types of literature must be studied to decide which related work is important to find how to use chest-mounted accelerometer data to develop a model to classify drinking. Firstly, it is important to know how accelerometer data is used to classify drinking behavior. Then, it is also important to look at the work done in using chest-mounted accelerometer data for activity recognition. After describing this, the previous work is compared in terms of how it can be useful with the current dataset previously discussed in Section 2.

3.1 How accelerometer data is used to classify drinking

In previous research on the classification of drinking with accelerometer data, it immediately arose that most of the research was performed on accelerometers connected to the wrist. [7] Studies of wrist accelerometer data generally agree that drinking can be divided into 3 or more stages. All of these papers decide on at least the following 3: raising the glass or hand-to-mouth, sipping, and lowering the glass or mouth-to-hand. For accelerometers connected to the arm, the easiest way to recognize drinking is to search for hand-to-mouth movement. Although it is much harder for a chest-mounted accelerometer to detect hand-to-mouth movement, one thing that is interesting to take from this is that the term 'drinking' may be interpreted differently by annotators. Will the annotators see drinking as the moment the participant starts with the hand-to-mouth movement or when the participant starts sipping their drink? This will be studied in depth in Section 4. Also, feature selection can be helpful because although the hand-mounted accelerometer will track different movements in the drinking behavior, it also tries to find statistical features that track a certain intensity of movement from accelerometer data. This paper uses correlation, mean, maximum, and covariance for the accelerometer. Research that will be more helpful for detecting drinking from arm-mounted accelerometer data is research that compares drinking with eating.[8] This research cannot depend on hand-to-mouth movement to identify drinking, because this is the same for drinking and eating.

While wrist movements may appear similar for eating and drinking, other papers, like the research by Gomes et al.[8] still use wrist movements in the hand-to-mouth phase in drinking from accelerometer data to classify their data.

3.2 How chest-mounted accelerometer data is used in activity recognition

The study of chest-mounted accelerometer data for other activity recognition helps to find a proven model. Important in choosing a model is to find a model that can handle scarce data and overfitting well. The research by Hosseinian[9] uses a random forest model, a support vector machine, and as a baseline a decision tree model to identify the difference between sitting, standing, lying, and walking. Research by Logacjov et al.[10] uses Support Vector Machine (SVM), k-Nearest Neighbor (k-NN), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (Bi-LSTM), and Multi-Resolution Convolutional Neural Network (Multi-res CNN) to classify sitting, standing, lying, walking, stair ascending, stair descending, and cycling. Research by Twomey et al.[11] uses Random Forest (RF), Logistic Regression (LR), Multi-layer Perceptron (MLP), Conditional Random Fields (CRFs), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) to classify sitting, standing, lying, and walking.

The scarcity of drinking moments is taken into account when choosing a fitting model for the conflab dataset. A random forest model provides the best solution for this. A random forest model handles a small sample setting well with proper variable screening and weighting, while hyperparameter tuning matters less.[12] Furthermore, a random forest model also works well to prevent overfitting.[13] This happens because while a single decision tree is sensitive to overfitting, averaging multiple uncorrelated trees reduces this overfitting. Reducing overfitting is important for this dataset because participants are not drinking much more frequently than they do.

Although chest-mounted accelerometers are studied a lot for activity recognition and arm-mounted accelerometers are used for drinking detection, there has been no notable research on drinking detection with chest-mounted accelerometers. With methods like the random forest model already used in different research for activity recognition, this is executed in Section 4.

4 Method

Multiple steps need to be taken beforehand to implement the random forest model. At first, the annotations need to be studied. The annotations then need to be preprocessed to get positive and negative samples as input to the model. Then there will be an explanation of how the samples are drawn. Lastly, the feature extraction will be explained.

4.1 Annotations

As explained in Section 2.2, a voting system should be implemented for the annotations. Three different voting systems are possible: AND voting, majority voting, and OR voting. The advantage of choosing an AND voting system is that there is a higher chance of the start and end of the window found by the AND voting being in the sample or a higher certainty of the window only containing sipping. For OR voting, the advantages are that delay from one annotator will not lead to missing part of the start of the drinking moment and that part of the hand-to-mouth phase and mouth-to-hand phase as described in Section 3.1. The majority voting has some of the advantages and disadvantages of both. Weighing the advantages against the disadvantages, an OR voting system is the most fitting. The biggest reason for this is that the transition from not drinking to drinking and vice versa also gives information about whether someone is drinking, which is also proven by previous research using the transition from not drinking to drinking as an indicator of whether someone is drinking. Although in this research the accelerometer was arm-mounted, and this transition gives more information to the arm-mounted accelerometer, there is still movement in this phase, which can be helpful for the chest-mounted accelerometer. Movement of the head and chest normally does not begin at the moment when someone sips their drink.

Only four participants had positive samples when applying the OR voting and leaving out annotated drinking moments when the participants were not sipping. The way their samples look is shown in Table 1.

Table 1: Summary of Participant Drinking Activity. With the number of frames in which they are drinking. The percentage of total time they are drinking. and the number of occasions were they are drinking.

Participant	**Frames drinking**	**Percentage of time drinking**	**Different drinking occasions**
Participant 1	221	3.07%	3
Participant 5	151	2.10%	2
Participant 20	610	8.47%	1
Participant 43	868	6.03%	8

4.2 Preprocessing

Now that positive samples are drawn, some negative samples need to be drawn. A set of rules is decided for choosing negative samples to take unambiguous samples of whether the sample contains any frames of drinking behavior (also containing the lifting of a glass and lowering of a glass):

- Every negative sample must have as large a distance to the other frames as possible. This is especially needed for the distance with drinking cases, as negative samples being close to the positive samples can lead to negative samples containing drinking behavior.
- Every negative sample must contain only acceleration data annotated as not drinking. This means that the samples left out in the annotation phase by reviewing the videos can also not be part of the negative samples.
- The size of every negative sample must be close to the size of the positive samples of that participant to simulate events that are not completely different from drinking behavior.

For every participant, twice as many negative samples as positive samples are chosen to reduce overfitting and still make use of the many samples available. The following preprocessing steps are followed:

- Negative samples are drawn according to the set of rules described.
- The samples are converted to fit the accelerometer data. This is done by first converting the frames to seconds by dividing by 60. Then, it is also normalized to fit the accelerometer data. This is done by multiplying by the ratio 936.98 / 938.1. The accelerometer data spans over 936.98 seconds, while the annotations are executed over 938.1 seconds. This is the case because the videos are not exactly 2 minutes.

- The accelerometer data is collected for every sample and is ready to be used for constructing a feature vector.

Table 2 shows information on all samples. Such as the start and end frame in the video, the video the sample is drawn from, the accelerometer data the sample covers, the stretch of the video the column is talking about, if the participant is drinking or not, and the participant the sample is about. The video times and videos are added to make the qualitative analysis reproducible, or in other words, to watch the content of the videos for analysis purposes. In Figure 3, the distribution of the samples is presented graphically. The mean and median of the distribution are the same, while the distribution of the samples is slightly different. This is done to eliminate systematic bias based on the length of an activity. The samples also seem to have a close to normal distribution between half and three seconds, with an outlier at ten seconds.

Table 2: Table of all samples with the following information: the first column shows sample indices for later reference. The second column shows the start and end frames of each sample. The third column indicates the video where the sample appears. The fourth column provides the corresponding time range (in seconds) from the accelerometer data. The fifth column lists the sample's time range in the video (in seconds). The sixth column contains the label (1 for drinking, 0 for not drinking). The final column shows the participant the sample is about.

Index	Sampleframes (start, finish)	Video	Sampletime (during accelerometer data)	Sampletime video	Label	Participant
0	(1227,1305)	vid3_seg4	(597.802, 599.101)	(20.45,21.75)	1	1
1	(1549,1624)	vid3_seg4	(603.162, 604.411)	(25.82,27.07)	1	1
2	(3566,3634)	vid3_seg4	(636.739, 637.871)	(59.43,60.57)	1	1
3	(272,355)	vid3_seg4	(581.904, 583.286)	(4.53,5.92)	0	1
4	(611,686)	vid3_seg4	(587.548, 588.796)	(10.18,11.43)	0	1
5	(2103,2173)	vid3_seg4	(612.385, 613.55)	(35.05,36.22)	0	1
6	(2684,2752)	vid3_seg4	(622.056, 623.188)	(44.73,45.87)	0	1
7	(4801,4870)	vid3_seg4	(657.298, 658.446)	(80.02,81.17)	0	1
8	(5922,5997)	vid3_seg4	(675.959, 677.207)	(98.70,99.95)	0	1
9	(1142,1215)	vid3_seg6	(836.134, 837.349)	(19.03,20.25)	1	5
10	(7121,7199)	vid3_seg6	(935.665, 936.963)	(118.68,119.98)	1	5
11	(503,570)	vid3_seg6	(825.497, 826.612)	(8.38,9.50)	0	5
12	(3088,3171)	vid3_seg6	(868.528, 869.91)	(51.47,52.85)	0	5
13	(4802,4871)	vid3_seg6	(897.061, 898.21)	(80.03,81.18)	0	5
14	(6177,6252)	vid3_seg6	(919.95, 921.199)	(102.95,104.20)	0	5
15	(6086,6697)	vid3_seg2	(438.975, 449.146)	(101.43,111.62)	1	20
16	(2192,2823)	vid3_seg2	(374.153, 384.657)	(36.53,47.05)	0	20
17	(3938,4531)	vid3_seg2	(403.218, 413.09)	(65.63,75.52)	0	20
18	(790,913)	vid3_seg2	(350.814, 352.862)	(13.17,15.22)	1	43
19	(2764,2895)	vid3_seg2	(383.675, 385.855)	(46.07,48.25)	1	43
20	(3028,3110)	vid3_seg2	(388.069, 389.434)	(50.47,51.83)	1	43
21	(4445,4532)	vid3_seg2	(411.658, 413.106)	(74.08,75.53)	1	43
22	(6803,6913)	vid3_seg2	(450.911, 452.742)	(113.38,115.22)	1	43
23	(6493,6668)	vid3_seg3	(565.607, 568.52)	(108.22,111.13)	1	43
24	(6759,6892)	vid3_seg3	(570.035, 572.249)	(112.65,114.87)	1	43
25	(7172,7199)	vid3_seg3	(576.91, 577.36)	(119.53,119.98)	1	43
26	(362,486)	vid3_seg2	(356.141, 358.055)	(6.03,8.10)	0	43
27	(1319,1446)	vid3_seg2	(381.128, 383.042)	(21.98,24.10)	0	43
28	(1912,2078)	vid3_seg2	(397.408, 399.439)	(31.87,34.63)	0	43
29	(2319,2465)	vid3_seg2	(416.219, 418.3)	(38.65,41.08)	0	43
30	(3341,3427)	vid3_seg2	(427.173, 428.538)	(55.68,57.12)	0	43
31	(3767,3845)	vid3_seg2	(434.181, 435.713)	(62.78,64.08)	0	43
32	(4832,4903)	vid3_seg2	(463.546, 465.61)	(80.53,81.72)	0	43
33	(5359,5422)	vid3_seg2	(479.477, 481.591)	(89.32,90.37)	0	43
34	(5858,5966)	vid3_seg2	(489.348, 492.112)	(97.63,99.43)	0	43
35	(6374,6490)	vid3_seg2	(496.124, 498.554)	(106.23,108.17)	0	43
36	(1110,1225)	vid3_seg3	(513.137, 514.568)	(18.50,20.42)	0	43
37	(2611,2726)	vid3_seg3	(520.228, 521.527)	(43.52,45.43)	0	43
38	(3589,3711)	vid3_seg3	(537.957, 539.139)	(59.82,61.85)	0	43
39	(4719,4844)	vid3_seg3	(546.73, 547.779)	(78.65,80.73)	0	43
40	(5377,5459)	vid3_seg3	(555.037, 556.834)	(89.62,90.98)	0	43
41	(5798,5890)	vid3_seg3	(563.626, 565.557)	(96.63,98.17)	0	43

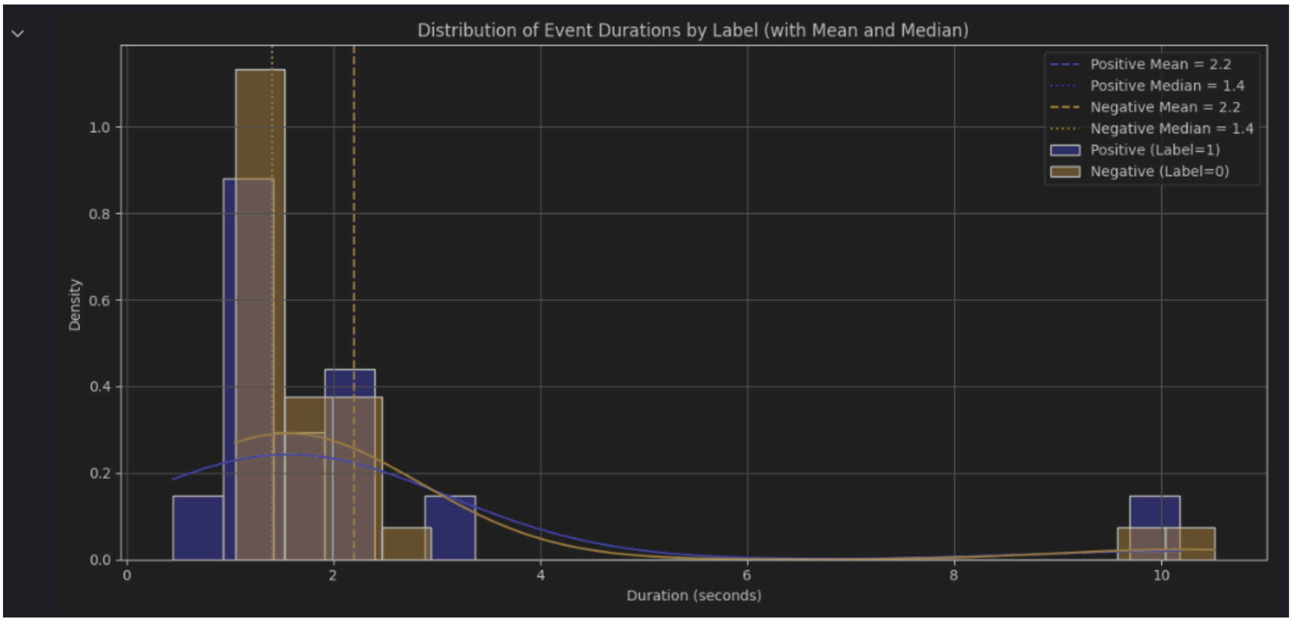

Figure 3: Statistical distribution of positive and negative samples with their mean and distribution.

4.3 Feature extraction

Different statistical features are considered and compressed into a feature vector to construct one from the available samples. First, it is important to consider the directions of the accelerometer that are also described in Section 2.1. The x-direction is irrelevant for detecting drinking because people do not have specific movement in the horizontal plane when drinking, whereas movement in the y- and z-directions gives information about drinking. This means that statistical features are calculated in the y- and z-directions. The different statistical features, how they are calculated, and why the features are chosen are shown in Table 3 and Table 4.

Table 3: Feature vector components and how they are calculated.

Component	**Calculation Formula**
Mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Correlation	$r = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$
Maximum	$\max(X) = \max\{x_1, x_2, \ldots, x_n\}$
Minimum	$\min(X) = \min\{x_1, x_2, \ldots, x_n\}$
Covariance	$\text{Cov}(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

Table 4: Feature vector components and their purpose.

Component	**Why It Is Used**
Mean	Shows average intensity of drinking movement, which helps to detect the posture of a person when they are drinking or not drinking.
Correlation	Captures the degree to which two axes vary together over the time of the movement.
Maximum	Captures extremes of movements, which can help identify drinking.
Minimum	Captures extremes of movements, which can help identify drinking.
Covariance	Captures the degree to which two axes vary together.

These features lead to a feature vector that contains the following eight statistical features for every sample: mean in the y direction, mean in the z-direction, correlation between the y-direction and the z-direction, maximum in the y-direction, maximum in the z-direction, minimum in the y-direction, minimum in the z-direction, and covariance in the y-direction and the z-direction.

4.4 Experimental setup

Now that the feature vector is extracted and a random forest model is decided, an experimental setup is selected. First, for participant 43, cross-validation is used on the 24 samples. Participant 43 is chosen because they have the most different occasions when they are drinking, which gives more training and test samples. In a study by Pang and Jung[14], different cross-validation methods are used, where, for their research, tenfold cross-validation keeps the balance between not having too much bias per single split and not having too much variance between sets. Although for their research the sample size is 200, this experiment uses 24 samples, so a six-fold cross-validation is implemented. The sixfold cross-validation works well in this case because the number of samples is divisible by six, which means that every fold contains four samples. For the sixfold cross-validation, the 24 samples are split into six folds of four samples on beforehand. Each fold is used interdependently as a test set, where the other five folds are used as a training set. In this way, every sample is used for testing.

Second, there will be a leave-one-out cross-validation (LOOCV), which is according to Cha et al.[15]: ”useful when evaluating the performance of machine learning when a dataset or category value is small.”. For this dataset, this means that every participant in itself is used as a test set, where the other three participants are the training set. This method generates more data for the small dataset. This creates four different simulations in which one participant is simulated to be a new participant with no annotations available, which helps to answer the second sub-question.

4.5 Parameters

Different parameters are chosen for the models to optimize their performance. The following parameters are chosen and explained how they were picked.

- The random tree forest uses 100 estimators. This number was hyperparameter tested on both models against 50 and 200 estimators and performed slightly better.
- A y prediction threshold of 0.33 for the single participant model and 0.3 for the across participant model is chosen by testing thresholds 0.3, 0.33, and 0.5. The threshold means that if the model sees the probability as higher than the threshold to be 1, it will predict 1. The parameters of 0.3, 0.33, and 0.50 were chosen because 0.5 has no bias in prediction, 0.33 balances the fitting to the balance of the sample sizes(two negatives for every positive). 0.3 is a value chosen for more bias towards drinking prediction, but not extreme. This is done to counteract the overfitting to the negative samples.
- A random seed of 42 is chosen to keep the results the same for each time the random forest, as well as the sixfold, is trained, and to make the results of the random forest model reproducible.
- The random forest model has parameter: ”class_weight = ”balanced””. This helps to adverse not drinking predictions because there are more non-drinking samples than drinking samples.

4.6 Evaluation metrics

The right evaluation metrics need to be chosen to get valuable results. Precision is helpful to detect whether the model has many false positives or how many actual instances the model correctly classified. Recall helps detect false negatives, or detect how many of the actual classifications it found correctly. The F1 score balances precision and recall. A classification report is made to assess precision, recall, and the F1 score for negative cases, positive cases, the weighted average between them, and the unweighted average between them. Section 5 will show a table with precision, recall, and F1 score for every different rendition of the sixfold cross-validation and LOOCV. In addition, a ROC curve will be shown that compares the model with a random guess. The higher the area under this curve (AUC), the better the performance of this model. In addition to the quantitative metrics, the false positives and false negatives are described in a table for both models.

5 Result

This section details the results of the experiments according to the methodology presented in Section 4. First, experiments are carried out on the samples of participant 43 with a sixfold cross-evaluation. Finally, the experiments are performed with all participants with LOOCV.

5.1 Experiments on individual

The sixfold cross-validated random forest model is evaluated in a classification report in Table 5 and the ROC curve in Figure 4 to answer whether drinking can be detected for a single individual. The quantitative analysis done here also helps to study the cases where the model fails, which is useful for qualitative analysis by studying samples that the model mispredicts. These are shown in Table 6. The exact meaning of the results and how they help answer these questions is explained in Section 6.

Table 5: Classification Report Metrics of the sixfold cross-validation of participant 43 with Mean

Fold	**Class 0**				**Class 1**				**Accuracy**	**Macro Avg**				**Weighted Avg**			
No.	Prec.	Rec.	F1-score	Supp.	Prec.	Rec.	F1-score	Supp.		Prec.	Rec.	F1-score	Supp.	Prec.	Rec.	F1-score	Supp.
1	1.00	1.00	1.00	3	1.00	1.00	1.00	1	1.00	1.00	1.00	1.00	4	1.00	1.00	1.00	4
2	1.00	0.67	0.80	3	0.50	1.00	0.67	1	0.75	0.75	0.83	0.73	4	0.88	0.75	0.77	4
3	0.50	0.50	0.50	2	0.50	0.50	0.50	2	0.50	0.50	0.50	0.50	4	0.50	0.50	0.50	4
4	1.00	1.00	1.00	2	1.00	1.00	1.00	2	1.00	1.00	1.00	1.00	4	1.00	1.00	1.00	4
5	0.67	0.67	0.67	3	0.00	0.00	0.00	1	0.50	0.33	0.33	0.33	4	0.50	0.50	0.50	4
6	1.00	1.00	1.00	3	1.00	1.00	1.00	1	1.00	1.00	1.00	1.00	4	1.00	1.00	1.00	4
Mean	**0.86**	**0.81**	**0.83**	**2.67**	**0.67**	**0.75**	**0.69**	**1.33**	**0.79**	**0.76**	**0.78**	**0.76**	**4.00**	**0.81**	**0.79**	**0.79**	**4.00**

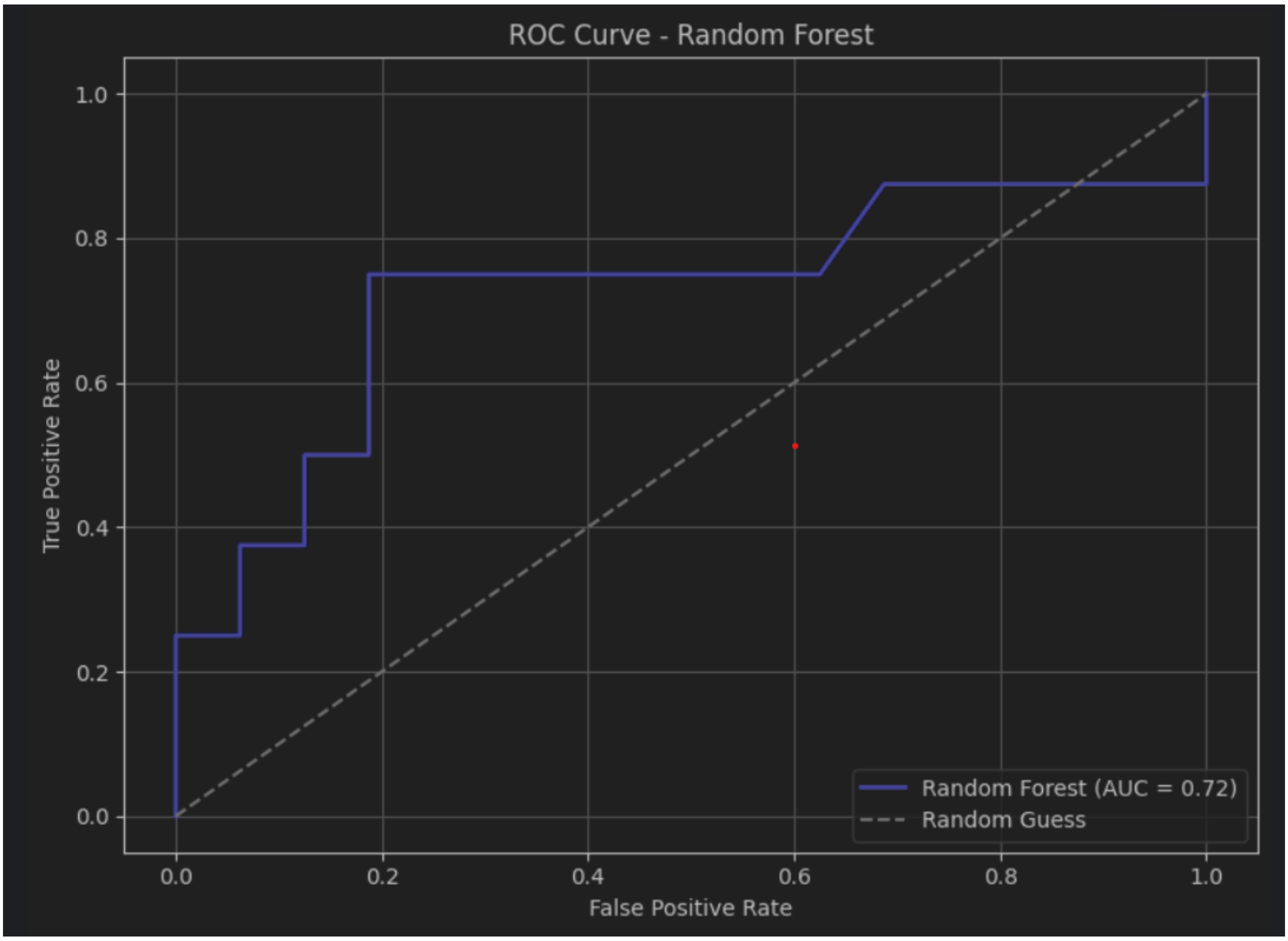

Figure 4: ROC curve of the sixfold cross-validation of participant 43.

Table 6: Table showing mispredictions of the model for specific samples. The first column shows the relevant sample number. The second column shows whether the participant drinks (1) or not (0). The last column describes the behavior shown by the participant in the sample.

Sample Number	**Actual behavior**	**Description**
23	1	The participant is walking while drinking.
25	1	The sample is very short and is cut off when the video ends. This means that the entire motion of drinking is not captured.
29	0	The participant nods twice in agreement with their conversation partner. This motion is comparable to the motion of drinking.
35	0	The participant nods twice in agreement with their conversation partner. This motion is comparable to the motion of drinking.
39	0	The participant nods in a conversation while standing still. The acceleration of this movement can be comparable to that of drinking.

5.2 Experiments across participants

The leave-one-out cross-validated random forest model is evaluated in a classification report in Table 7 and the ROC curve in Figure 5 to answer whether drinking can be detected between different participants. In Table 8, the behavior of false positives and false negatives is described. The exact meaning of the results and how they help answer these questions is explained in Section 6.

Table 7: Classification Report Metrics of LOOCV between participants of a random forest model with the mean of all results.

Part No.	**Class 0**				**Class 1**				**Accuracy**	**Macro Avg**				**Weighted Avg**			
	Prec.	Rec.	F1-score	Supp.	Prec.	Rec.	F1-score	Supp.		Prec.	Rec.	F1-score	Supp.	Prec.	Rec.	F1-score	Supp.
1	0.80	0.67	0.73	6	0.50	0.67	0.57	3	0.67	0.65	0.67	0.65	9	0.70	0.67	0.68	9
5	0.80	1.00	0.89	4	1.00	0.50	0.67	2	0.83	0.90	0.75	0.78	6	0.87	0.83	0.81	6
20	0.67	1.00	0.80	2	1.00	0.00	0.00	1	0.67	0.83	0.50	0.40	3	0.78	0.67	0.53	3
43	0.82	0.56	0.67	16	0.46	0.75	0.57	8	0.62	0.64	0.66	0.62	24	0.70	0.62	0.63	24
Mean	**0.77**	**0.81**	**0.77**	**7.00**	**0.74**	**0.48**	**0.45**	**3.50**	**0.70**	**0.76**	**0.65**	**0.61**	**10.50**	**0.76**	**0.70**	**0.66**	**10.50**

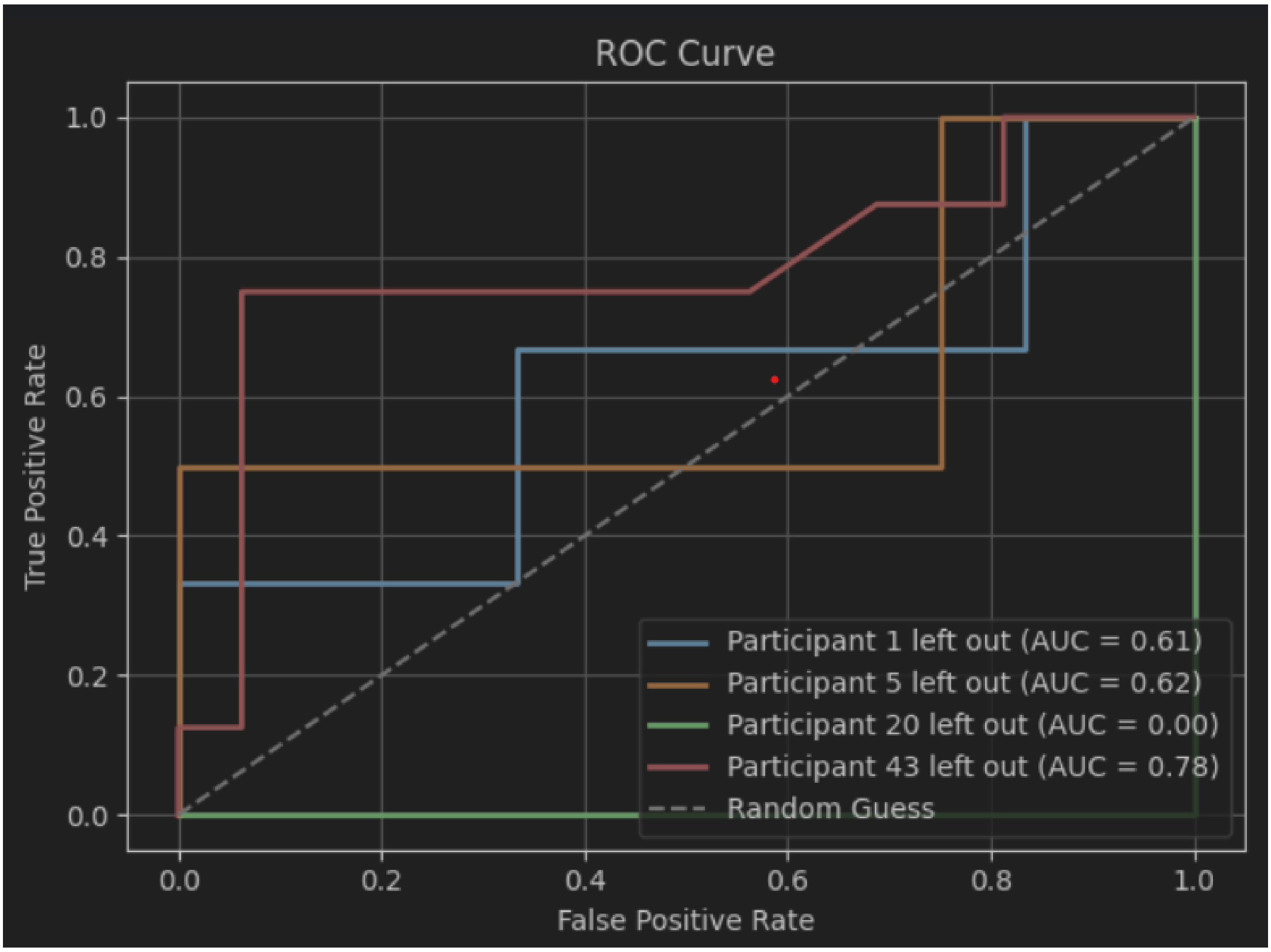

Figure 5: ROC curve of the LOOCV on all participants. Every different participant, as a test set against the three other participants as a train set, has its own ROC curve as well as a random guess model that functions as a baseline.

Table 8: Table showing mispredictions of the model across participants. The first column shows the relevant participant. The second column shows the sample number corresponding to Table 2. The third column shows whether the participant drinks (1) or not (0). The last column describes the behavior shown by the participant in the sample.

Participant	**Sample Number**	**Actual behavior**	**Description**
1	2	1	Participant takes a quick sip of their drink.
1	5	0	Participant raises their drinking hand to scratch their eyebrow.
1	8	0	Participant listens to their conversation partner and has little movement.
5	9	1	It is hard to see if the participant is drinking. The top camera view places the participant's lips the same height as the coffee he is drinking from. There is no certainty of correct annotation, so the model can be right, in this case.
20	15	1	Participant finishes their drink.
43	18	1	Participant drinks with most movement from the wrist and almost no head movement.
43	21	1	Participant drinks with most movement from the wrist and almost no head movement.
43	30	0	Participant listens to their conversation partner and has little movement.
43	31	0	Participant listens to their conversation partner and has little movement.
43	32	0	Participant is explaining something to their conversation partner by pointing with what used to be their drinking hand.
43	33	0	Participant leans a bit forward to watch the phone of their conversation partner.
43	34	0	Contestant scratches their left cheek in conversation with their non-drinking hand.
43	38	0	Participant listens to their conversation partner and has little movement.
43	39	0	Participant is scratching their head with their non-drinking hand.

6 Discussion

In the discussion, the results obtained from the random forest model are studied. This discussion section begins with an examination of the results for participant 43. This initial assessment provides insight into its effectiveness in an individual context. Subsequently, the analysis expands to include results from all participants, revealing broader patterns of model behavior. Then, a direct comparison is made between the individual and collective outcomes to highlight similarities and differences. In addition, qualitative analysis is studied to come up with ideas on how to improve the model. All of this leads to an answer to the research question stated in 1.

Table 5 will be studied for the analysis of how the model performs on a single participant, which in this case is participant 43. The accuracy tells that 79 % of all classifications that were made are correct. This accuracy means that the model performs better than a random guess method, which would get a precision of 50%, and better than an all-negative prediction, which would lead to an accuracy of 67%. The weighted average and the macro average show more in-depth details of the results. The weighted average has a bias towards the negative class, which counts twice as heavily as the positive class because there are twice as many samples in the negative class. This makes the macro-average a more informative metric. This metric counts positive class predictions the same as negative class predictions. This also makes sense for evaluating the model because predicting whether someone is drinking is just as important as predicting whether someone is not drinking. The F1-score being slightly below the accuracy in the macro-average means that the model performs slightly better in predicting negative classes than in predicting positive classes.

This is also confirmed by the results of class 0, or the negative class, being slightly better than the results of class 1, or the positive class. The precision of class 0 means that it predicts the correct negative samples 86% of the time. Where the recall of class 0 means that in 81% of the cases the model predicts zero, it is zero. The F1 score of 83% shows that precision and recall are balanced and both score well. The precision of class 1 means that it predicts the correct positive samples 67% of the time. Where the recall of class 1 means that in 75% of the cases the model predicts one, it is one. The F1 score of 69% shows that precision and recall are balanced and both score decent.

The AUC of 0.72 shown in Figure 4 indicates that the model performs much better than a random model that would have an AUC of 0.5, but is still quite far from being a perfect model that would have an AUC of 1.0. This indicates that there is some discriminatory power in the model and that the model captures patterns in the data.

Generally, the model already performs very well for the amount of data fed. The slight imbalance between class 0 and class 1 is also logical because class 0 has more samples, which leads to better performance. This difference is not great enough that the model overfits class 0, but completely balancing the classes, which could be done by lowering the threshold, led to worse general performance of the model. With all the evaluation scores, it can be concluded that there is a causal relationship between accelerometer data and drinking.

Table 7 will be studied first for the analysis of how the model performs in predicting data for a new participant. The accuracy tells that 70 % of all classifications that were made are correct. This accuracy means that the model performs better than a random guess method, which would get a precision of 50%, and marginally better than an all-negative prediction, which would lead to an accuracy of 67%. The F1-score being marginally below the accuracy in the macro-average means that the model performs better in predicting negative classes than in predicting positive classes.

The results of class 0, or the negative class, being slightly better than the results of class 1, or the positive class, also confirm this. In class 0, the model scores well across all metrics, while in class 1, the recall and F1-score are much lower. Especially in predicting drinking behavior for participant 20, this is seen. The precision of 1.0 and the recall of 0.0 indicate that the model predicted all zeros. This can have two causes. The first is in the samples being much larger than the other samples, as seen in Table 2 and Table 1. This can lead to different statistical behavior for other participants. In short drinking moments, most of the drinking frame contains movement; a longer drinking moment is for a longer time in a rest state, which leads to recognizing it as not drinking. Another cause may be the small sample size. There are only three samples, and knowing that two-thirds of the samples are negative, it is not completely unlikely to predict all zeros. The best F1-score of 67% is achieved by participant 5. This also has two causes. Due to the sample size of six, there are many samples in the training set, which generally leads to better prediction. In addition, the drinking behavior of participant 5 is more similar to the drinking patterns of the other people in the set.

The AUC values shown in Figure 5 indicate how well drinking can be predicted for every participant. This indicates that there is no discriminatory power in predicting drinking behavior for participant 20, which confirms that it predicts all zeros for participant 20. For participants 1 and 5, the AUC scores show that there is marginally better predictive power to predict their behavior from others. The data for participant 43 captures most of the patterns in the data.

Generally, the model already performs decently for the amount of data fed to three of the participants. Participant 20, who has been drinking for a longer time, is harder to predict by the model. This can be improved by feeding more data from more different people into the model to generate different drinking patterns. Some people drink more or move more slowly while drinking. With the data currently fed, the model recognizes general drinking patterns, but has a harder time managing outliers.

The model performs better in predicting drinking for one individual. This is because individuals show similar patterns between different moments of drinking, while different people will show different patterns. It is easier to predict drinking for an individual when there is already data trained on them. New people's drinking behavior can sometimes be predicted, while in more atypical cases, this is harder. The prediction of new participants will improve when more different samples are gathered because this will generalize more drinking styles from taking many sips to taking one sip, from much movement in drinking to little movement in drinking.

Video analysis can offer reasoning behind the false predictions made by the model. In Table 6, the actions per sample of participant 43 are described. There were two cases of false negatives, or cases where participants who drank were predicted as not drinking. For sample 23, walking while drinking leads to acceleration in the z-direction being higher than in the participant's samples when standing still. This affects all metrics in the z-direction, and the correlation and cross-validation. To solve this, more samples where the participant is walking and drinking need to be available. It would also help to annotate cases of drinking while walking differently from drinking while standing still, as also done by Gomes et al.[7]. For the second case, it is just a case of a wrong moment where the video ends. This can also be fixed by better annotation, or by annotation on one video to solve the desynchronization issue.

For false positives, or cases where participants who did not drink were predicted as drinking, of the sixfold cross-validation for participant 43, the issue of nodding looking like drinking in the statistical features derived from the accelerometer data. This issue can be solved by using nodding samples in training the model as a third label against drinking and not drinking. Training nodding in the model is helpful to find more differences in the statistical features of the accelerometer data between nodding and drinking. It can also help with feature engineering to find new features.

The results in Table 8 also provide some insight to help strengthen the model. In false negatives, several different reasons for the mispredictions stand out. The first is sample 2, where participant 1 took a quick sip. The velocity of the sip can lead to misprediction. More drinking at a similar speed should be added to the training set. The same problem arises for sample 15. The system has no training data of a participant taking many sips in one go, and the model will predict a non-drinking case. A more diverse dataset will help. The issues in the samples of participant 43 are harder to solve. To know if this is possible, the model needs more training. At this moment, it is not possible to say with certainty whether the model can predict these subtle movements.

The false positives also have different reasons for which they are mispredicted. Several cases where little movement is there are hard to explain, but are probably since the training and test sets being imbalanced, because of the different number of drinking moments each participant has. More data will help solve this problem. This is the same for movement in the non-drinking hand. Movement in the drinking hand is something that can be solved by adding more cases of this to the training data, or even adding it as a label itself, as in the nodding case.

Conversations lead to subconscious behavior that can look like drinking in statistical features of accelerometer data, like raising the drinking hand and nodding. Behavior like this needs to be addressed in training data. Next to these two examples, more of these cases can arise. To find these more negative samples in more data needs to be qualitatively analyzed.

In general, the model shows that on a small sample size, accelerometer data can already predict, with an accuracy of 79%, data for a participant that is already present in the dataset, and with an accuracy of 70% for a new participant. Qualitative analysis shows that more data and more accurate annotation can improve these scores. To improve these scores, more data needs to be collected for drinking in a conference setting to train the dataset. Data from more different participants will lead to more different drinking styles being recognized, which will help with a more diverse recognition of drinking behavior, and thus, a more generalizable model. In addition, a more consistent annotation can help in better evaluation. A consensus on drinking can help to become a more consistent annotation, leading to less work in preprocessing. Dividing the annotation into the three drinking phases: lifting the glass, sipping the drink, and lowering the glass, will also improve the model. This can lead to differentiating the three phases to obtain more relevant features in the different phases for accelerometer data. Now, all features can be added to the feature vector for the three different phases, leading to more features and better predictions. Annotations can also be improved by adding an annotation for nodding, movement of the drinking hand, and other similar movements to drinking that may be found in the future. Where the ceiling is for chest-mounted accelerometer data for predicting drinking behavior in conference settings, and whether it can perform as well as arm-mounted accelerometers in a controlled setting is still unanswered. What the model answers is that drinking behavior can already be predicted to some extent with scarce data, and with a lot of possibility for improvement. When the model has improved, the model can even be made into a dynamic model that is not dependent on chosen samples, but which can find drinking behavior in a video.

7 Responsible research

A research paper should consider how confidential data is handled and how research can be reproducible.

7.1 Confidential dataset

The conflab dataset used is confidential. This means that the dataset is not publicly available, but the data is not sensitive. This means that all data should be explained in a manner that does not require access to the dataset. The section dataset gives the reader a general idea of all the concepts of the dataset and how they are used. Furthermore is all modeling that is based on the dataset is explained in a general manner with meta information. The confidentiality of the data is especially important in the videos where participants are shown. How these videos are used and how somebody with access to the dataset would use them is explained clearly without showing any material where participants are visible.

7.2 Reproducibility

Section 4 explains all steps to get the results shown in Section 5 with only the dataset. All samples used are shown in Table 2. The cross-validation methods for the model are explained in Section 4.4. The parameters used are all in Section 4.5.

8 Conclusion

This study explored the feasibility of detecting drinking behavior in natural social settings using chest-mounted accelerometer data. Through careful preprocessing, feature extraction, and the use of a random forest classifier, promising results were obtained. Training and testing for a single participant indicated a clear improvement over random guessing and demonstrated that meaningful patterns related to drinking behavior can be captured. Although the model performs better when trained and tested for the same individual, the performance across participants is still moderate and reveals the challenge of generalizing drinking patterns between different individuals. Qualitative analysis proved improvements for the current models and for models that have more data available in the future.

Overall, the results indicate a detectable relationship between accelerometer signal features and drinking behavior in noisy, uncontrolled environments. However, the variability between participants and limitations of the dataset highlight the need for a larger and more diverse dataset and better annotation protocols.

References

[1] Chirag Raman, Jose Vargas-Quiros, Stephanie Tan, Ashraful Islam, Ekin Gedik, and Hayley Hung. ConfLab: A Data Collection Concept, Dataset, and Benchmark for Machine Analysis of Free-Standing Social Interactions in the Wild, October 2022. arXiv:2205.05177 [cs].

[2] Hayley Hung, Gwenn Englebienne, and Jeroen Kools. Classifying social actions with a single accelerometer. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, UbiComp '13, pages 207–210, New York, NY, USA, September 2013. Association for Computing Machinery.

[3] Abdulkerim Öztekin, Ömer Faruk Ertuğrul, Erdoğan Aldemir, and Emrullah Acar. Determining the most relevant frequency bands in motion identification by accelerometer sensors. *Multimedia Tools and Applications*, 81(8):11639–11663, March 2022.

[4] Lauren R. Williams, Steven T. Moore, Greg J. Bishop-Hurley, and Dave L. Swain. A sensor-based solution to monitor grazing cattle drinking behaviour and water intake. *Computers and Electronics in Agriculture*, 168:105141, January 2020.

[5] Lisa Benson. *Accelerometers*, volume 11. 04 2006.

[6] Accelerometer - Measure linear acceleration along X, Y, and Z axes in m/s2 - Simulink.

[7] Diana Gomes and Inês Sousa. Real-time drink trigger detection in free-living conditions using inertial sensors. *Sensors*, 19(9), 2019.

[8] Diana Gomes, João Mendes-Moreira, Inês Sousa, and Joana Silva. Eating and Drinking Recognition in Free-Living Conditions for Triggering Smart Reminders. *Sensors*, 19(12):2803, January 2019. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

[9] Seyed Mohammadreza Hosseinian. Human Activity Recognition from an Accelerometer on The Chest: Data Transformation, Feature Selection, and Classification Performance. December 2017.

[10] Aleksej Logacjov, Kerstin Bach, Atle Kongsvold, Hilde Bremseth Bårdstu, and Paul Jarle Mork. HARTH: A Human Activity Recognition Dataset for Machine Learning. *Sensors (Basel, Switzerland)*, 21(23):7853, November 2021.

[11] Niall Twomey, Tom Diethe, Xenofon Fafoutis, Atis Elsts, Ryan McConville, Peter Flach, and Ian Craddock. A comprehensive study of activity recognition using accelerometers. *Informatics*, 5(2), 2018.

[12] Sunwoo Han, Brian D. Williamson, and Youyi Fong. Improving random forest predictions in small datasets from two-phase sampling designs. *BMC Medical Informatics and Decision Making*, 21(1):322, November 2021.

[13] Rukshan Pramoditha. How to Mitigate Overfitting by Creating Ensembles, September 2021.

[14] Herbert Pang and Sin-Ho Jung. Sample Size Considerations of Prediction-Validation Methods in High-Dimensional Data for Survival Outcomes. *Genetic epidemiology*, 37(3):276–282, April 2013.

[15] Gi-Wook Cha, Hyeun Moon, Young-Min Kim, Won-Hwa Hong, Jung-Ha Hwang, Won-Jun Park, and Young-Chan Kim. Development of a prediction model for demolition waste generation using a random forest algorithm based on small datasets. *International journal of environmental research and public health*, 17, 09 2020.