The Data Barrier to Lightweight Drinking Detection

An Analysis of the Viability of Skeleton-Only Models on In-the-Wild Social Data.

Joelle Tijssens
Supervisor(s): Hayley Hung, Litian Li, Stephanie Tan
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: Joelle Tijssens
Final project course: CSE3000 Research Project
Thesis committee: Hayley Hung, Litian Li, Julian Urbano Merino

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

This research addresses the challenge of deploying real-time drinking gesture detection in messy, "in-the-wild" environments. We propose and evaluate two computationally inexpensive systems, one using a Random Forest classifier, another using a 1-Dimensional Convolutional Neural Net (1D-CNN) classifier. Both are trained on 2D skeleton data, or features derived from that solely that data. Tested on the Conflab social interaction dataset, our method is designed to handle sparse labels and significant data occlusion. This study reports on the performance of this light-weight, video-based approach, providing a benchmark for applicability in real-world health and human-computer-interaction applications where privacy and computational efficiency are important factors. Although we were unable to create a robust and reliable classifier (f1 of 0.07 and 0.03 respectively), this work shows that there is potential for future work to succeed (roc-auc's of 0.63 and 0.55 respectively) and provides critical insights into pitfalls to avoid when designing similar systems.

1 Introduction

Detecting when a person takes a drink in real time might sound redundant, but it's a piece of information that can be incredibly useful in a wide range of applications. In medicine, it can help monitor patient recovery or combat health issues like obesity. For human-computer interaction (HCI) and even marketing, understanding this simple action can provide valuable insights. The key to making this technology practical is finding a method that is cheap, non-intrusive, and easy to deploy. Combining simple fast-running machine learning algorithms with pose data, in this paper, we hope to contribute to exactly that goal.

Recent advancements in computer vision have made it possible to extract 2D skeletal data from videos in real-time, even with multiple people in the frame[1, 2]. This "skeleton-only" approach is promising because it respects privacy (no faces are needed) and is computationally efficient. Previous research has already shown its potential, especially when combined with multiple input modalities [1]. For instance, one study demonstrated that skeleton data is a viable alternative to more complex motion capture systems for analyzing drinking motions [3]. Another successful project used video to detect eating and drinking gestures in a cafeteria, achieving high F1 scores of 0.88 for drinking. Furthermore, research using Spatial-Temporal Graph Convolutional Networks (S-T GCNs, neural networks designed to analyze skeletal motion over time) has confirmed that skeleton data can be used to robustly detect these "intake gestures" in a way that generalizes across different datasets[4].

However, a significant gap exists between these successful studies and real-world application. A lot of the previous research was conducted in controlled environments with relatively "clean" data, using models that often require higher-end hardware to run. Study participants were often stationary, filmed from a good angle, and not blocking each other. A challenge presents itself in how well these models perform in messy, uncontrolled social settings (e.g. at a party or in a busy cafe) where people are mingling, cameras are at awkward angles, and body parts are frequently hidden from view (occluded), especially when applied to real-time detection. One study [1] has shown that drinking-specific intake detection on this type of data is possible. Yet, the model used in this study is relatively computationally expensive and might be out of scope for projects that require real-time, cheap, or offline usability. While some studies have shown that simpler intake detection models can work on imperfect footage of animals, and that data quality loss doesn't always harm gesture detection, this specific leap to complex human social scenes has not been fully explored.

Our research aims to test whether the success of real-time skeleton-based drinking detection can be extrapolated to these challenging, "in-the-wild" scenarios. To guide our investigation, we have formulated a main research question and several sub-questions.

1.1 Research Questions

Main Question: Can skeleton-only data reliably be used to detect the drinking action in uncontrolled, human social settings in real-time applications? To address this, we will investigate the following sub-questions:

1. To what extent can computationally inexpensive models, trained on sparse 2D skeleton data, achieve reliable performance for drinking detection in 'in-the-wild' social settings?
2. How well can classifiers trained on sparse pose features distinguish drinking from visually similar actions, and which actions are most commonly confused?
3. Under which specific visual conditions (e.g., occlusion, camera angle) do the trained pose-based models tend to fail?

2 Related Work

In this section, we build upon the topics introduced previously, reviewing the literature and recent developments that are foundational to our research. We will examine the progression from video to skeleton data, its application in human intake detection, and the challenges of deploying these methods in uncontrolled environments.

2.1 From Video to Skeleton Data: A Viable Alternative

The core technology enabling our research is the ability to extract skeletal keypoints from standard video footage. Recent studies have validated this "video-to-skeleton" pipeline as a low-cost, non-intrusive alternative to traditional, marker-based motion capture systems. For instance, a study on neurotypical adults performing a drinking task found that computer vision-based pose estimation yielded results comparable to those from on-body motion capture systems, effectively demonstrating its potential for objectively assessing motor function in clinical settings [3]. This underscores the potential for significant advancements in medical monitoring. Furthermore, research into detecting similarly complex social actions, such as laughter, has shown that models can be surprisingly resilient to the loss of input modality or the use of imperfect annotations during training [5]. This robustness is a promising indicator for our goal of analyzing messy, "in-the-wild" footage.

2.2 Detecting Human Intake Gestures

The use of skeleton data for detecting intake gestures is a burgeoning field. Researchers have successfully employed models designed to interpret spatial-temporal data (such as Recurrent Neural Networks (RNNs) and Spatial-Temporal Graph Convolutional Networks (ST-GCNs)) to classify actions from keypoint trajectories [2, 6, 7, 8]. For example, one notable study utilized entire-meal context from video to achieve high F1 scores of approximately 0.93 for eating and 0.88 for drinking gestures within a large cafeteria dataset. Another advanced

approach used MMPose to extract a 23-joint upper-body skeleton and fed this data into an ST-GCN combined with a BiLSTM network to effectively distinguish between eating and drinking actions.

While these results are impressive, they often rely on computationally intensive models. One of the few studies to specifically tackle drinking detection in uncontrolled social mingling situations highlighted this trade-off [1]. The authors achieved strong results using a top-down camera view combined with static social data, but noted that their approach was too computationally expensive for many real-time or embedded systems. This finding motivates our central goal: to find a computationally leaner approach, especially given that for many human-computer interaction applications, system response times are expected to be under a few seconds [9].

To achieve real-time performance, other researchers have explored alternative sensors. For example, studies have successfully detected drinking gestures using wrist-worn IMU sensors [10, 8] and even FMCW radar systems [11]. While our work focuses on video, these studies are relevant as they often employ similar temporal convolutional network architectures and share the application goal of monitoring fluid intake for the elderly and patients in daily life. Their success reinforces the feasibility of real-time gesture detection and provides insights into effective model architectures. Interestingly, several of these studies also suggest that dividing the drinking action into distinct stages (e.g., reaching, drinking, retracting) can significantly improve detection accuracy [8, 1].

2.3 "In-the-Wild" Detection

A key question is whether these successes can translate to the chaos of real-world social scenes. While research on humans in such settings is still developing, studies on animal behavior offer compelling evidence[12]. For instance, high-throughput, real-time tracking of intake and other actions has been achieved in caged tree shrews using only pose estimation data from imperfect camera angles. Similarly, the BehaviorDePOT tool demonstrates that automated behavioral detection based on markerless pose tracking can achieve excellent results in relatively uncontrolled environments with non-human subjects [13]. These successes in animal studies bolster the hypothesis that pose-only data can be a powerful tool for detecting specific actions even in complex, unpredictable settings.

Finally, a review of the literature suggests that a few key features are consistently predictive of drinking actions. The most crucial signals appear to be hand-to-face proximity, particularly the distance between wrist keypoints and the nose or mouth region. This is complemented by secondary features such as the elevation of the arm and forearm, joint angles like elbow flexion, and the overall temporal profile of the hand's trajectory. Our research will leverage these established feature sets as a starting point for our computationally inexpensive model.

3 Dataset

To effectively test our research questions, we required a dataset that captures the complexities of real-world human interaction. An ideal dataset would feature multiple subjects in an uncontrolled social setting, contain significant visual challenges like occlusion, and include annotations for both the target action (drinking) and corresponding skeletal poses. For these reasons, we selected the **Conflab** dataset.

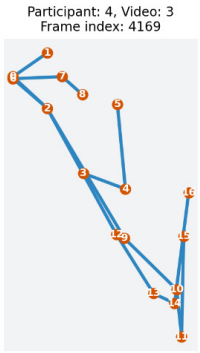


Figure 1: Example of a Skeleton (this person is viewed from their top left at a 45 degree rotational angle)

Conflab is a rich, multimodal dataset collected during a "real-life professional networking event at the international conference ACM Multimedia 2019."[14] It provides a unique opportunity to study social behaviors "in-the-wild" using footage from various cameras, capturing at 60 frames per second, and data from chest-worn accelerometers. For this study, we utilized the video footage, the publicly available drinking annotations, and the corresponding 2D COCO-like skeletal keypoint annotations (as in Figure 1) derived from the video. The dataset's camera perspectives, often from a top-down angle, capture the natural mingling of participants, providing a realistic testbed for our models.

3.1 Dataset Characteristics and Limitations

While its authenticity makes it highly suitable for our research, the Conflab dataset also presents several significant challenges that mirror the difficulties of real-world deployment.

First, the dataset contains a limited number of positive examples. We identified just under 200 instances of drinking across the entire dataset. After aligning these with the available pose data and allowing a 60-frame tolerance to capture the full motion, we were left with 109 distinct drinking sequences. This scarcity is compounded by the fact that some sequences are duplicates resulting from multiple annotators labeling the same event. This results in a severe class imbalance, with only about 1 in every 124 frames being labeled as part of a drinking action.

Second, the data is inherently messy and incomplete. Due to the nature of a crowded social event, subjects frequently occlude one another, leading to a high number of null values in the pose data where key points could not be detected. Participants also move freely between different camera views, creating discontinuities. The intersection of reliable pose data and drinking annotations is therefore relatively small, posing a significant challenge for training a robust classifier.

3.2 Exploratory Data Analysis

A preliminary analysis of the dataset revealed several key characteristics that inform our feature engineering and modeling approach.

We first analyzed the variance difference **between the two classes** of each skeletal key point's coordinates to identify which body parts exhibit the most movement during drinking

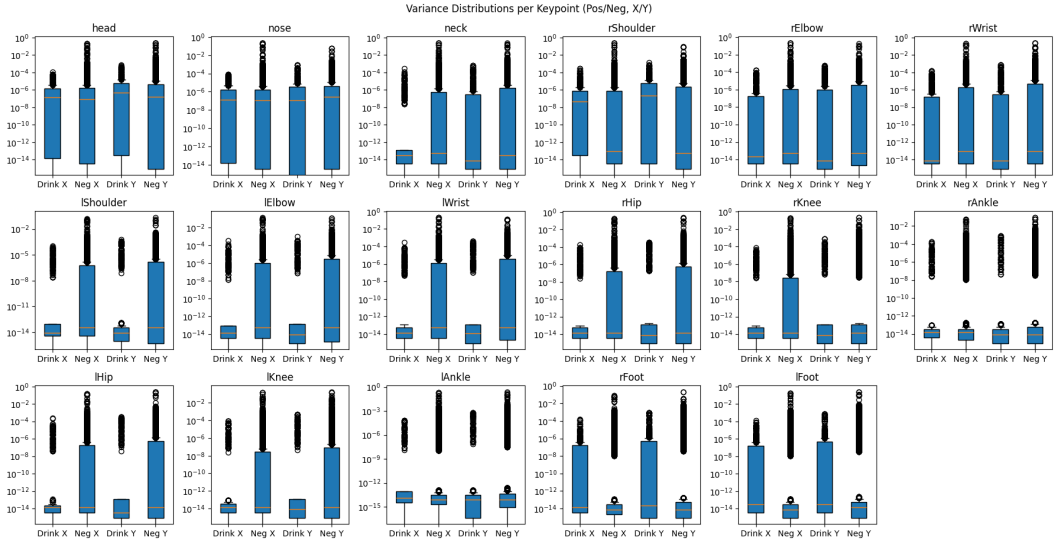


Figure 2: Temporal movement variance per keypoint, label and axis

actions (Figure 2). As expected, key points associated with the upper body, specifically the neck, shoulders, elbows, and wrists, showed high per-segment variance. This confirms their importance as predictive features. More surprisingly, our analysis also revealed significant movement in the hips, knees, and feet, suggesting that subtle shifts in posture and stance may be a secondary signature of the drinking action.

We also examined the distance between the hands and the mouth (Figure 3), using the wrist and nose keypoints as proxies. This is theoretically a primary indicator of drinking. However, a plot of the minimum wrist-to-nose distance during annotated drinking events revealed that between the two labels, the observed difference in distances was not as big as anticipated. This is further corroborated by the fact that preliminary analysis on segmented data, which shows there is not an immediately apparent correlation between the amount of frames with the positive label (y=1) in a window, and the variance of each key-point's position.

4 Methodology

This chapter details the systematic process used to investigate our research questions. We outline **two** distinct modeling approaches: a lightweight, feature-based Random Forest classifier and a heavier to train, yet light to run 1-Dimensional (the dimension being the time direction) Convolutional Neural Network (CNN) based on the architecture designed in an earlier study [5]. For both methods, we will explain the data pre-processing pipeline, feature selection and extraction, model architecture, training procedure, and evaluation protocol. By evaluating a classical feature-based model (Random Forest) against a deep learning approach (1D-CNN), we aim to compare **two distinct lightweight methodologies**. By doing so, we ensure our conclusions are not contingent on a single modeling architecture, thereby strengthening their validity.

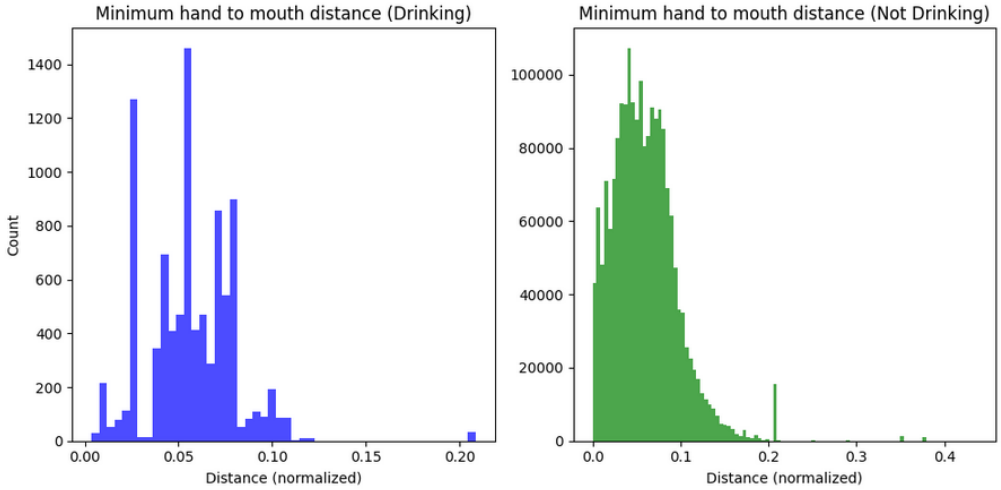


Figure 3: The Minimum hand-to-mouth Distance Per Label.

4.1 Shared Experimental Setup

Both models were trained and evaluated using a **5-fold cross-validation** scheme to ensure robust and generalizable results. To prevent data leakage between folds, the data was strictly partitioned by **video, video segment, and participant ID**. This strategy ensures that data from the same participant or continuous scene, even if captured by different cameras or labeled by different annotators, does not appear in both the training and validation set of a given fold. The core of our data handling for both approaches relied on a **sliding window** method, segmenting the continuous stream of pose data into fixed-size windows for classification. For both approaches, a window was considered part of the positive class if at least 60 percent of its frames had the positive label (drinking). Negative samples were generated as follows: For the training sets only, windows with an overall negative label (not drinking) were discarded if they contained any individual frames of the positive class. For validation, the 60 percent rule still applies, but no windows are thrown out.

On the usable data intersection (refer to section 3), drinking segments averaged 170.12 frames but were highly variable (SD=131.07). To balance capturing a large chunk of the drinking action with the need for sufficient training data, we chose a 180-frame window, as longer windows severely limited sample availability due to the 60 percent sampling rule for positives.

4.2 Approach 1: Random Forest Classifier

Our first approach prioritized computational efficiency and real-time performance by using a Random Forest (RF) classifier. This model is well-suited for learning from a curated set of higher-level features and is inherently robust to noisy data.

Data Pre-processing and Feature Extraction Before training, we worked exclusively with the intersection of data where both drinking and pose annotations were available.

The primary pre-processing step involved compressing the spatio-temporal pose data from each window into a fixed-size feature vector. Guided by our exploratory data analysis, we engineered features designed to capture the key dynamics of a drinking motion:

- **Hand-to-Face Proximity:** The maximum and minimum Euclidean distance between the wrist keypoints and the nose keypoint.
- **Hand Dynamics:** The mean speeds and total horizontal and vertical displacements of the hands.
- **Limb Variance:** The temporal variance of horizontal and vertical coordinates for the wrist, head, shoulder, and feet.
- **Arm Kinematics:** The angles formed by the shoulder-elbow-wrist keypoints.

Occluded keypoints (represented as `NaN` values) were handled during this feature extraction step. Within each sliding window, any `NaN` value was imputed using the mean of the available data points for that specific coordinate within that window.

Model Training and Implementation To address the severe class imbalance, we **downsampled the negative class** during training. Furthermore the following **hyperparameters** were tuned: window stride (5), the negative-to-positive training ratio (2), and several Random Forest parameters: number of estimators (1000), max depth (none), minimum samples for split (8), and minimum samples per leaf (4).

4.3 Approach 2: 1D-CNN Classifier

Our second approach aligns more closely with current literature, using a 1D-CNN to learn features directly from raw pose data. The architecture was inspired by a previous study[5] on laughter detection of which the source code has been published on GitHub[15].

Data Pre-processing and Input Representation Based on our analysis showing that certain joints are more informative, we selected a minimal yet effective skeleton to reduce noise and dimensionality. The input for the CNN consisted of the raw coordinate sequences of the following keypoints: **nose, neck, both shoulders, both wrists, and both hips**. This selection captures essential upper body and core movement. For this model, occluded data points within a window were imputed with zero. The raw sequences of these coordinates from a fixed-size window of **180 frames** were fed directly into the CNN.

Model Training and Implementation The sliding window approach achieved an optimal window stride at 60 frames. To manage the class imbalance, we again downsampled the negative class and used a weighted Binary Cross-Entropy (BCE) loss function with a weight factor of 8.

The following hyperparameters were optimized, with the final values used in parentheses: learning rate (3e-4), batch size (32), BCE loss weight (8), and the negative-to-positive ratio (2).

Table 1: Overall Cross-Validation Performance of the 1D-CNN Model.

Metric	**Average Score**
ROC AUC	0.5527 (+/- 0.1690)
Precision	0.0240 (+/- 0.0459)
Recall	0.0532 (+/- 0.0972)
F1-Score	0.0330 (+/- 0.0624)
Accuracy	0.9213 (+/- 0.0064)

4.4 Evaluation Protocol

Both models were evaluated using the same set of metrics: **F1-score, Precision, Recall, ROC-AUC, and overall accuracy**. To convert the probabilistic outputs of each model into binary labels for calculating F1, precision, and recall, we did not use a fixed 0.5 threshold. Instead, for each fold, we determined an **optimal threshold** by selecting the value that maximized the F1-score on the Precision-Recall curve of that fold's validation data. Finally, a manual qualitative analysis was performed to examine false positives and false negatives to identify which gestures are most commonly misclassified by each model.

For a more in-depth view of the model implementations and to ensure full reproducibility of these experiments, the source code is available on GitHub[16]. Refer to section 8 for a concrete statement on reproducibility.

5 Results

In this section, we present the performance of the two proposed classification models. The results are organized to directly address the research sub-questions outlined in section 1.1, providing a comparative analysis based on quantitative metrics and qualitative error inspection.

5.1 Comparative Model Performance

We evaluate both models according to the validation criteria discussed in section 4. To address RQ1 regarding the performance of lightweight models, the overall metrics for the 1D-CNN approach are presented in Table 1, and overall metrics for the RF approach are presented in Table 2.

5.1.1 1D-CNN-Based Approach

The 1D-CNN was trained using the optimal hyperparameters found during tuning, which included a training stride of 60, a learning rate of 3e-4, a batch size of 32, a BCE loss weight of 8, and a negative-to-positive training ratio of 2. The average performance across all five folds, using a dynamically optimized classification threshold for each fold, is summarized in Table 1.

While the overall scores are low, the model's performance was consistently better than random chance. Figure 4 displays the Receiver Operating Characteristic (ROC) curves for each fold and their mean, plotting the true positive rate against the false positive rate. As

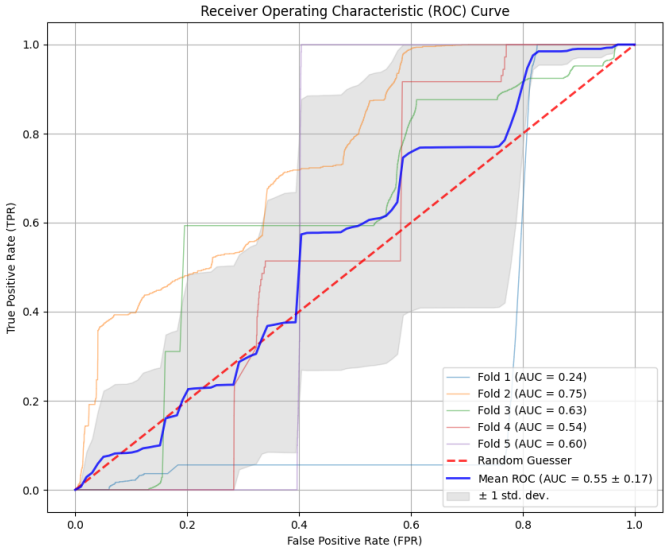


Figure 4: ROC curves for the 1D-CNN model. The plot shows the performance of each of the 5 folds, their mean (in blue), and the performance of a random guesser (dashed line).

Table 2: Overall (mean) Performance of the Random Forest Model.

Metric	**Score**
ROC AUC	0.6391
Accuracy	0.9427
Precision	0.0512
Recall	0.1730
F1-Score	0.0790

shown, the mean curve and several individual folds remain above the diagonal line representing a random guesser, indicating that the model learned a discernible, albeit weak, signal from the data.

The Precision-Recall curves, shown in Figure 5, further illustrate the model's behavior. This plot of precision versus recall highlights that Fold 2 demonstrated the most potential in distinguishing positive instances.

To provide a qualitative view, Figures 6a through 6d show examples of the model's predictions plotted against the ground truth on a timeline.

5.1.2 Random Forest-Based Approach

The Random Forest (RF) model, which operated on engineered features, yielded the performance metrics (mean across folds) shown in Table 2. Interestingly, the best results were achieved by training the model to predict the negative class (class 0) and taking the probability for the positive class as 1 - P(0).

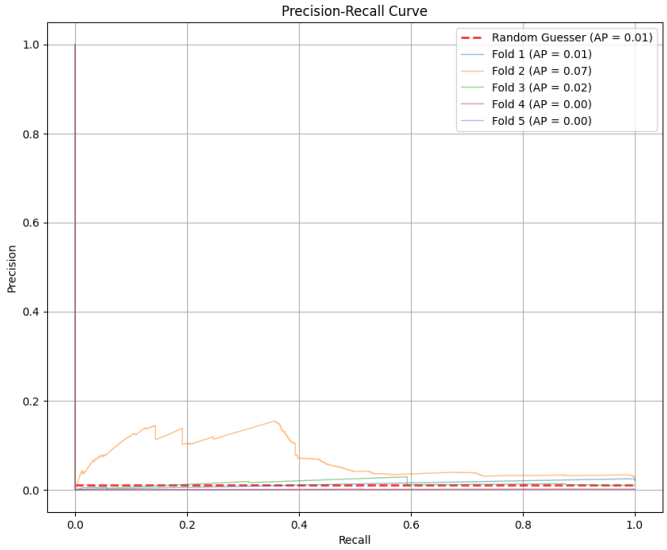


Figure 5: Precision-Recall curves for the 1D-CNN model for each fold against a random guesser baseline.

The ROC curve for the RF model, seen in Figure 7, shows that the mean performance is notably above the random baseline, with Fold 4 performing particularly well. However, the Precision-Recall curves in Figure 8 are mostly flat and close to the baseline, with a sharp spike in precision only at a near-zero recall. This indicates the model struggled to confidently identify positive samples without creating many false positives. This behavior is further clarified by observing that the optimal threshold for achieving a realistic F1-score was very high (approximately 0.98 , suggesting the model learned to default a single class.

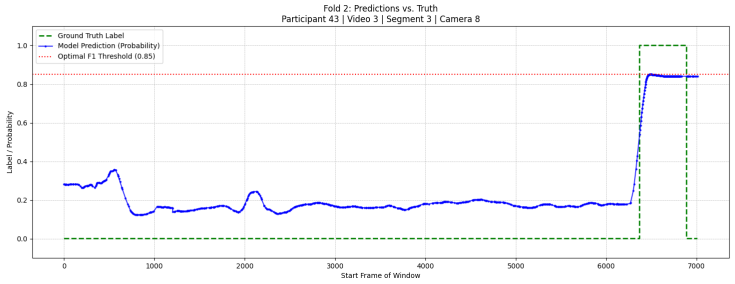
5.2 False Positives: Misclassifications on Similar Actions

Since the CNN showed slightly more potential for nuanced classification despite its low scores, we focused the detailed qualitative analysis on its errors.

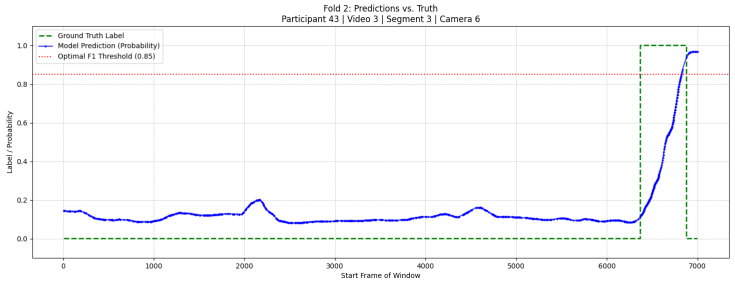
In total there were 43 continuous sequences of frames (not individual windows, but subsequences of the total data) either classified as false positive or false negative. These were gathered from the validation set predictions of the associated folds. The full list of these identified sequences can be found in the Appendix B.

To understand which actions are most commonly confused with drinking, we manually inspected the false positives generated by the models.

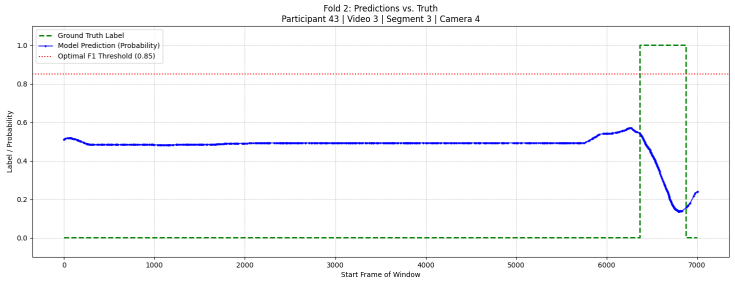
What resulted was the following list of actions with the associated amount of occurrences (this list does not remove duplicates among camera angles since the model classified both differently on some occasions): **Wrist clutching**: 1, **General talking gestures**: 5, **Scratch**: 2, **Adjust hair**: 2, **Fully occluded**: 1, **partially occluded**: 4, **mostly occluded**: 6, hands and mouth in close 2D proximity **due to camera angle**: 2, **head bowing down**: 2, **no specific associated action** (likely due to the decision threshold



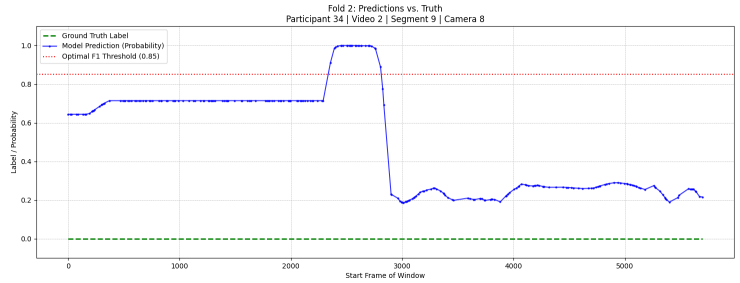
(a) A correctly identified drinking sequence.



(b) A delayed positive classification.



(c) An incorrect negative classification.



(d) An incorrect positive classification (person was gesturing while speaking).

Figure 6: Qualitative examples of the 1D-CNN model's performance. (a) A correct detection. (b) A correct but delayed detection. (c) A false negative (missed event). (d) A false positive (incorrect event).

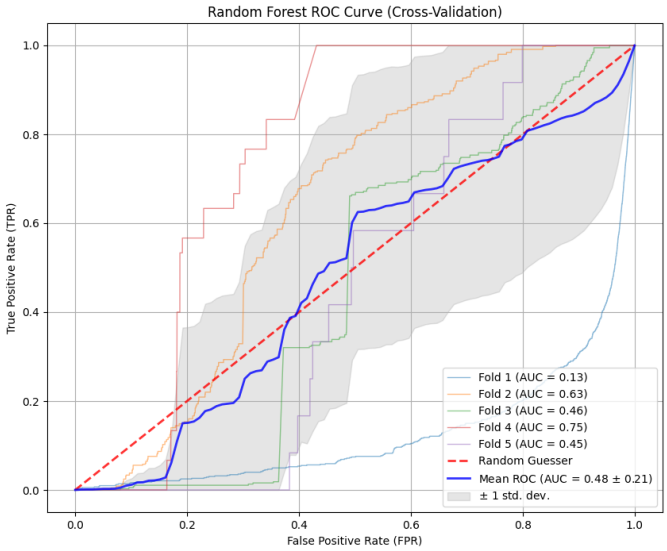


Figure 7: ROC curves for the Random Forest model, showing each fold, their mean, and a random guesser.

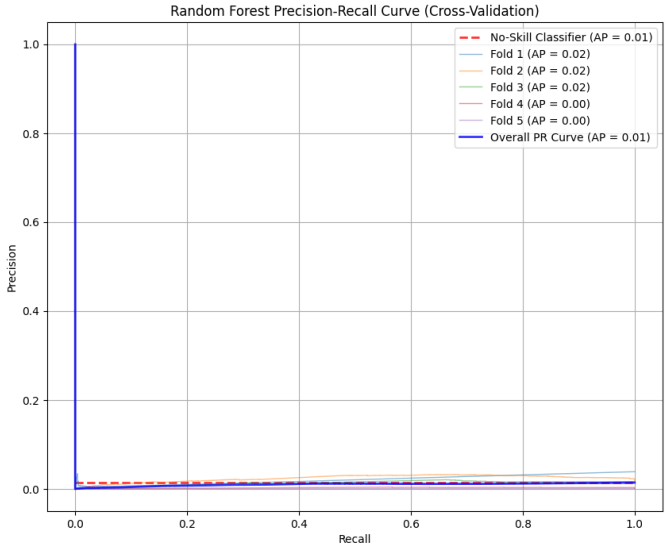


Figure 8: Precision-Recall curves for the Random Forest model.

being sub-optimal): 7.

5.3 False Negatives: Analysis of Failure Conditions

Finally, we investigated the conditions under which the CNN model failed to detect true drinking events. A primary hypothesis was that performance would degrade with poorer data quality. Inspection of the drinking instances missed by the model confirmed that the model performed worse on segments with a higher number of occluded keypoints. An investigation into the model's false negatives confirmed our hypothesis that performance degrades with poorer data quality. Of the 11 drinking events missed by the model, the distribution was as follows: **Partially occluded**: 3, **Mostly occluded**: 2, **fully occluded**: 2, **annotator mistakes**: 2, **similarity** (between drinking and not drinking pose due to camera angle): 1, **not occluded**: 1.

6 Discussion

The performance of both the 1D-CNN and Random Forest models was unmistakably poor, with F1-scores only slightly better than random chance. This leaves the answer to **RQ1**, which questioned if lightweight models could achieve reliable performance, largely inconclusive, save for the observation that they performed slightly better than random guessing.

In response to **RQ3**, which investigated specific failure conditions, this poor performance is largely attributable to severe data scarcity (and to a lesser extent, inherently messy data); with fewer than 40 positive training instances, the models were data-starved, making the lack of varied data the most severe bottleneck. However, the limited results that were garnered do suggest that occlusion of the limbs and the accidental 2D proximity of hands and limbs due to camera angles both contribute to a model's predictive uncertainty.

Furthermore, regarding **RQ2**, which focused on confusion with visually similar actions, the dataset's top-down perspective created ambiguous cues where common non-drinking movements, such as general hand gesturing or adjusting the hair, were easily mistaken for the target action due to similar hand-to-head proximity. The models also seemed to perform roughly as bad on ambiguous actions as on occluded actions. Despite these fundamental challenges, the ROC curves for both models remained consistently above the random-chance baseline, indicating that a weak, learnable signal does exist, even if it could not be effectively captured by models trained from scratch on this limited dataset.

6.0.1 Limitations and Future Directions

The primary limitation of this study was the attempt to train models on an insufficient dataset, a decision constrained by the project's 10-week timeline. This timeframe precluded more effective strategies such as acquiring a richer dataset, performing additional annotation, or correctly implementing a transfer learning approach with a pre-trained model like MMAction2[17]. Additionally, the Random Forest model was hampered by a sub-optimal feature set. This could be mitigated by utilizing features known to have a higher rate of success. The features outlined by EatSense[18] in their study could lead to improved results, but were not implemented due to time constraints.

Consequently, future work should directly address these limitations. The most critical next step is to establish a robust baseline on this dataset by fine-tuning a large, pre-trained action recognition model. This would provide a much clearer answer to the core research

question. Subsequent efforts could then focus on engineering more discriminative features for classical models and developing methods to account for potential annotator lag in the dataset.

6.0.2 Broader Implications

While the direct results were not successful, they carry several important implications for research in this area:

An "In-the-Wild" lesson: Our findings serve an advisory role for future work, highlighting the significant performance gap between models trained on controlled data versus those deployed in messy, real-world scenarios.

The Necessity of Transfer Learning: This work strongly suggests that for sparse data problems like this one, transfer learning with large, pre-trained models should be the default methodology, not an alternative to training from scratch.

A Focus on Data-Centric AI: The challenges posed by camera angles, occlusion, and potential annotation errors indicate that progress may depend more on curating higher-quality datasets than on simply designing novel model architectures.

7 Conclusion and Future Work

This research investigated the viability of using lightweight, skeleton-only models for real-time drinking detection in challenging, "in-the-wild" social environments. We evaluated two distinct approaches, a feature-based Random Forest classifier and a 1D-CNN, on the **Conflab** dataset, which is characterized by sparse labels and significant visual noise.

Our findings demonstrate that training these models from scratch on such data, limited in size as the dataset turned out to be, is insufficient to achieve reliable performance. Both models performed only marginally better than a random baseline, failing to produce a robust classifier suitable for real-world application. The primary obstacles identified were the severe scarcity of positive training examples, occlusion of the limbs, and the inherent visual ambiguity of the top-down camera angles, which together made it difficult to distinguish drinking from other common hand-to-head movements.

Despite these low performance scores, our analysis suggests a weak, learnable signal is present in the data. Therefore, future work should not abandon this problem but rather pivot to more suitable methods. The most critical next step is to leverage transfer learning by fine-tuning large, pretrained action recognition models, which we hypothesize are necessary to overcome the data-related challenges encountered in this study. Securing richer datasets for training and engineering more sophisticated features for classical models remain vital secondary avenues for exploration.

Ultimately, while this study did not produce a deployable model, it successfully underscores the significant gap between controlled and real-world scenarios and provides a clear, data-centric path forward for achieving practical, real-time behavioral analysis.

8 Responsible Research and Privacy

This research complies with the 'TU Delft Code of Conduct', as well as the 'Netherlands code of conduct for research integrity'[19].

We acknowledge the ethical responsibilities inherent in conducting research involving human data and developing technologies for behavioral analysis. This work was conducted in adherence with the principles outlined in the TU Delft Code of Conduct for Research Integrity. Our ethical considerations focused on three key areas: study reproducibility, the privacy of the dataset participants and the broader societal implications of this research.

8.1 Reproducibility

Standard libraries have been used in the implementation of the models and evaluation criteria mentioned in the article. The source code has also been released[16]. The main hurdle to reproducibility is access to the database, which can be requested via university institutions. The dataset was not used in it's entirety, seeing as not all the data was useful, the code to attain the same intersection of used data has also been released in the same repository as the rest of the source code. Most of the results are numerical and can be verified. A 'random state' set at an unchanging 42 was used to prevent random results after running the training loop. These facts together should make the results of this paper easily verifiable and reproducible on hardware with at least 16GB of RAM.

To ensure full transparency and to allow for the verification of our findings, the source code[16] for this project has been released publicly in an open-source repository. This approach is in line with our commitment to reproducibility, allowing other researchers to build upon and critique this work.

Finally, all software libraries used in our implementation are **open-source** with open licenses, ensuring there are no implementation related licensing barriers to reproducing this work and replicating our results.

8.2 Bias

We acknowledge that several forms of bias could influence this research.

The **Conflab** dataset was collected at an international academic conference. The demographics of such an event may not be representative of the general population, potentially skewing towards specific age groups and ethnicities, and will especially be biased in terms of levels of education. A model trained on this data may therefore perform differently and less equitably when applied to other demographic groups.

With a very small number of positive examples (around 40 training instances), there is a high risk that a model could learn spurious correlations. For example, if a specific gesture or camera angle coincidentally appeared in several drinking sequences, the model might incorrectly learn that this feature is predictive of drinking.

The poor performance of our models inherently mitigates the risk of deploying a biased system, as the models are not reliable enough for any application. However, we acknowledge these biases are present in the dataset. We strongly recommend that any future, more successful research using this or similar datasets to include a rigorous evaluation of the model's performance across diverse demographics, ensuring fairness and preventing the amplification of existing societal biases.

8.3 Participant Privacy and Data Handling

The primary dataset used in this study, Conflab, consists of multimodal recordings of individuals at a real-life event. Our first step was to examine the ethical protocol of the original data collection. The authors of the Conflab dataset state that participants provided

informed consent for their data to be used in further research studies, which provides the ethical foundation for our work.

To build upon this foundation and to ensure the privacy of the participants, our methodology was designed to be inherently privacy-preserving. To train the models, we exclusively used binary drinking annotations, and 2D skeleton data derived from the video footage. This process abstracts the visual information into a set of anonymized keypoints and their trajectories, removing identifiable features such as faces and making the re-identification of an individual from the processed data extremely difficult.

8.4 Societal Implications and Mitigation

We recognize that any technology capable of detecting human actions from video footage has the potential for misuse. Such tools could be applied to applications of surveillance, leading to potential abuses by those in positions of power. While this technology is already widely and openly available, a successful outcome of this study could contribute to its ability to be deployed more covertly or on a wider scale by enabling it to run on smaller, local hardware.

Several factors mitigate this risk in the context of our project. First, this study does not develop a new, more powerful detection paradigm; it merely examines the performance of existing techniques on challenging, "in-the-wild" data. Our work serves as an analysis of what is currently possible rather than a tool created for public release.

Due to the potential for misuse, we recommend future successful works to only deploy the resulting technologies in situations where it is a best fit solutions, and where subject have provided **explicit consent** for the technology to observe them, thus preventing (unethical) surveillance.

8.5 Use of GenAI

A full statement on the use of Generative AI can be found in the appendices. In short Generative AI was used in the making of this paper in two main ways:

1. Writing help: An LLM was asked to improve the use of language and sentence structure of student-written text. In the process it made suggestions on parts of the student explanation that were still unclear and needed improvements.

2. Implementation help: An LLM was prompted to provide the boilerplate code for the DataSet/Loader and the main training and evaluation loops.

A Statement on the Use of Generative AI

Generative AI was used in the making of this paper and the source code. The following strategy was employed

- Writing help: An LLM was asked to improve the use of language and sentence structure of student-written text. In the process it made suggestions on parts of the student explanation that were still unclear and needed improvements.

- Implementation help: An LLM was prompted to provide the boilerplate code for the DataSet/Loader and the main training and evaluation loops.

The following Prompts were used in the writing process:

```
Hi! Could you please improve the flow, sentence structure and word usage of the following, wh

[insert paragraph or section]
```

The following prompts were used in the coding process:

```
Hi! Could you please show me how to implement the general structure/architecture of an ML pip
```

Debugging prompts:

```
Hi! Could you please help me figure out why the follwing code:

[...]

Produces the following error:

[...]
```

The above was used when a google search did not yield sufficient results.

```
Hi Gem, could you please turn this into a tabel for me? Please leave the text as is:

p5 v3 s5 c6: FP 2900-3000 frames - walking clutching wrist with other hand
p5 v3 s5 c8: FN 6700 frames - has walked out of frame
p6 v3 s2 c4: FP on the entire thing - starts segment with hands on mouth but then goes to ges
p6 v3 s5 c4: FP 1700-2000 and 4500 - talking hand gestures
p12 v2 s9 c4: FP after 1000 til the end - goes to scratch ear, then keeps gesturing
p36 v2 s9 c2: FP 1450 frame - adjusting know/ponytail/hair
p36 v2 s9 c4: FP 500-2000 frame  - adjusting hair and walks out of frame
p36 v2 s9 c6: FP after frame 500 til the end - not in frame except legs at frame 500
p44 v3 s2 c6: FP 2700 - (walks into frame and) scratches nose
p46 v3 s5 c6: FP f 5000-6300 - walks into frame, hands behind back
p25 v3 s2 c8: FP frame 0-10 - camera angle occludes hands, when they show they are close to h
p34 v2 s9 c8: FP 2400-2700 - walks into frame to shake hands (counted as hand gesturing)
p43 v3 s2 c6: multiple FN on 600-900, 2700-3200, 4200-4500, 6600-6900 - only one keypoint occ
p43 v3 s3 c2: FN frame 6400-6900 - not in frame
p43 v3 s3 c4: FN frame 6400-6900 - not occluded
p43 v3 s4 c2: entire segment FP - scratches a few times but not occluded, could be explained
p1 v3 s6 c4: entire thing FP - due to camera angle hands are behind/on the head
p10 v2 s9 c8: FP 4600-end - walks into frame
p20 v3 s2 c2: FN frame 5900-6700 - only leg visible and turns around
p20 v3 s2 c4: FN frame 5900-6700 - positioned in the middle of the screen, almost viewed from
p20 v3 s2 c6: FP frames 2200-2400 - almost entirely occluded
p23 v2 s9 c6: FN frames 2100-2400 - annotator mistake, cup empty
p23 v2 s9 c8: FN frames 2100-2400 - annotator mistake, cup is empty
p24 v3 s5 c4: FP entire segment - hands behind back occluded
p26 v3 s2 c2: FP entire segment - just standing: threshold failure
p29 v3 s5 c4: FP entire segment - regular movement, hands only sometimes occluded, mostly mod
p40 v3 s5 c4: FP entire segment - mostly occluded, hands moving rapidly
p41 v3 s6 c4: FP after frame 5000  - hands become non-occluded and starts gesturing
```

p41 v3 s6 c6: FN frames 1500-1800 - almost entirely occluded
p3 v2 s8 c4: entire segment - just standing, threshold failure
p3 v2 s8 c6: FP frames 5500-5600 - circular hand gestures while talking
p3 v3 s2 c4: entire segment - mostly standing, threshold failure
p5 v3 s6 c2: FN frames 1000-1300 and frame 6900-end - camera angle makes cup look close to fa
p13 v3 s5 c4: FP entire segment - hands mostly occluded turned away from camera
p16 v3 s5 c4: FP entire segment - hands behind back mostly occluded
p41 v3 s5 c4: FP entire segment - regular pose variation, threshold failure
p44 v3 s5 c4: entire segment - both gesturing and scratching continually, camera angle makes
p1 v3 s4 c4: FN frames 1100-3600 - scratches head (turned away from camera)
p6 v2 s9 c4: FP frames 5500-end - makes multiple circular talking hand gestures
p7 v3 s2 c2: FP entire segment
p10 v3 s2 c8: FP 3600 - bows head closer to hands
p25 v2 s9 c8: FP begin-4100 - mostly out of frame
p41 v2 s9 c8: FP 4000-end - entirely occluded

a lot of times on the worse folds, just walking into frame seems to be enough for the mislcas

model seems to falsely attribute drinking label to bows??

B Identified False Positives and False Negatives

In the following tables B.2 and B.1, the identifier key is specified as follows:

p_: Participant number
v_: Video Number
s_: Segment Number
c_: Camera Number

B.1 False Positives (FP)

Identifier	Frames	Description
p5 v3 s5 c6	2900-3000 frames	walking clutching wrist with other hand
p6 v3 s2 c4	on the entire thing	starts segment with hands on mouth but then goes to gesturing while talking the entire segment
p6 v3 s5 c4	1700-2000 and 4500	talking hand gestures
p12 v2 s9 c4	after 1000 til the end	goes to scratch ear, then keeps gesturing
p36 v2 s9 c2	1450 frame	adjusting know/ponytail/hair
p36 v2 s9 c4	500-2000 frame	adjusting hair and walks out of frame
p36 v2 s9 c6	after frame 500 til the end	not in frame except legs at frame 500
p44 v3 s2 c6	2700	(walks into frame and) scratches nose
p46 v3 s5 c6	f 5000-6300	walks into frame, hands behind back
p25 v3 s2 c8	frame 0-10	camera angle occludes hands, when they show they are close to head due to camera angle
p34 v2 s9 c8	2400-2700	walks into frame to shake hands (counted as hand gesturing)
p43 v3 s4 c2	entire segment	scratches a few times but not occluded, could be explained by *bowing* action which not many participants exhibit
p1 v3 s6 c4	entire thing	due to camera angle hands are behind/on the head
p10 v2 s9 c8	4600-end	walks into frame
p20 v3 s2 c6	frames 2200-2400	almost entirely occluded
p24 v3 s5 c4	entire segment	hands behind back occluded
p26 v3 s2 c2	entire segment	just standing: threshold failure
p29 v3 s5 c4	entire segment	regular movement, hands only sometimes occluded, mostly model threshold failure
p40 v3 s5 c4	entire segment	mostly occluded, hands moving rapidly
p41 v3 s6 c4	after frame 5000	hands become non-occluded and starts gesturing
p3 v2 s8 c4	entire segment	just standing, threshold failure
p3 v2 s8 c6	frames 5500-5600	circular hand gestures while talking
p3 v3 s2 c4	entire segment	mostly standing, threshold failure
p13 v3 s5 c4	entire segment	hands mostly occluded turned away from camera
p16 v3 s5 c4	entire segment	hands behind back mostly occluded
p41 v3 s5 c4	entire segment	regular pose variation, threshold failure
p44 v3 s5 c4	entire segment	both gesturing and scratching continually, camera angle makes hand close to head
p6 v2 s9 c4	frames 5500-end	makes multiple circular talking hand gestures
p7 v3 s2 c2	entire segment	
p10 v3 s2 c8	3600	bows head closer to hands
p25 v2 s9 c8	begin-4100	mostly out of frame
p41 v2 s9 c8	4000-end	entirely occluded

B.2 False Negatives (FN)

Identifier	Frames	Description
p5 v3 s5 c8	6700 frames	has walked out of frame
p43 v3 s2 c6	600-900, 2700-3200, 4200-4500, 6600-6900	multiple FN on - only one keypoint occluded
p43 v3 s3 c2	frame 6400-6900	not in frame
p43 v3 s3 c4	frame 6400-6900	not occluded
p20 v3 s2 c2	frame 5900-6700	only leg visible and turns around
p20 v3 s2 c4	frame 5900-6700	positioned in the middle of the screen, almost viewed from the top
p23 v2 s9 c6	frames 2100-2400	annotator mistake, cup empty
p23 v2 s9 c8	frames 2100-2400	annotator mistake, cup is empty
p41 v3 s6 c6	frames 1500-1800	almost entirely occluded
p5 v3 s6 c2	frames 1000-1300 and frame 6900-end	camera angle makes cup look close to face, then lifts other hand to gesture while talking
p1 v3 s4 c4	frames 1100-3600	scratches head (turned away from camera)

References

[1] X. Teng, "Drinking Behaviour Detection."

[2] Z. Tang and A. Hoover, "Video-based Intake Gesture Recognition Using Meal-length Context," *ACM Trans. Comput. Healthcare*, vol. 6, no. 2, pp. 15:1–15:24, Feb. 2025. [Online]. Available: https://dl.acm.org/doi/10.1145/3709151

[3] J. Huber, S. Slone, and J. Bae, "Computer vision for kinematic metrics of the drinking task in a pilot study of neurotypical participants," *Scientific Reports*, vol. 14, no. 1, 2024.

[4] Q. Wang, K. Zhang, and M. A. Asghar, "Skeleton-Based ST-GCN for Human Action Recognition With Extended Skeleton Graph and Partitioning Strategy," *IEEE Access*, vol. 10, pp. 41 403–41 410, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9749063/

[5] J. D. V. Quiros, L. Cabrera-Quiros, C. Oertel, and H. Hung, "Impact of Annotation Modality on Label Quality and Model Performance in the Automatic Assessment of Laughter In-the-Wild," *IEEE Transactions on Affective Computing*, vol. 15, no. 2, pp. 519–534, Apr. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10136533/

[6] R. Niewiadomski, M. Mancini, G. Varni, G. Volpe, and A. Camurri, "Automated Laughter Detection From Full-Body Movements," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 113–123, Feb. 2016. [Online]. Available: https://ieeexplore.ieee.org/document/7298420

[7] Z. Yu, C. Crane, L. Chen, M. Testa, and Z. Zheng, "A Computation Model to Estimate Interaction Intensity through Non-Verbal Behavioral Cues: A Case Study of Intimate Couples under the Impact of Acute Alcohol Consumption," *ACM Trans. Comput. Healthcare*, vol. 5, no. 3, pp. 13:1–13:23, Sep. 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/3664826

[8] D. Gomes and I. Sousa, "Real-Time Drink Trigger Detection in Free-living Conditions Using Inertial Sensors," *Sensors*, vol. 19, no. 9, p. 2145, Jan. 2019, number: 9 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/19/9/2145

[9] B. Shneiderman, "Response time and display rate in human performance with computers," *ACM Computing Surveys*, vol. 16, no. 3, pp. 265–285, Sep. 1984. [Online]. Available: https://dl.acm.org/doi/10.1145/2514.2517

[10] C. Wang, T. S. Kumar, W. De Raedt, G. Camps, H. Hallez, and B. Vanrumste, "Drinking Gesture Detection Using Wrist-Worn IMU Sensors with Multi-Stage Temporal Convolutional Network in Free-Living Environments," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2022, pp. 1778–1782, Jul. 2022.

[11] C. Wang, T. S. Kumar, W. D. Raedt, G. Camps, H. Hallez, and B. Vanrumste, "Eat-Radar: Continuous Fine-Grained Intake Gesture Detection Using FMCW Radar

and 3D Temporal Convolutional Network with Attention," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 2, pp. 1000–1011, Feb. 2024, arXiv:2211.04253 [cs]. [Online]. Available: http://arxiv.org/abs/2211.04253

[12] A. Alameer, I. Kyriazakis, and J. Bacardit, "Automated recognition of postures and drinking behaviour for the detection of compromised health in pigs," *Scientific Reports*, vol. 10, no. 1, p. 13665, Aug. 2020, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-020-70688-6

[13] C. J. Gabriel, Z. Zeidler, B. Jin, C. Guo, C. M. Goodpaster, A. Q. Kashay, A. Wu, M. Delaney, J. Cheung, L. E. DiFazio, M. J. Sharpe, D. Aharoni, S. A. Wilke, and L. A. DeNardo, "BehaviorDEPOT is a simple, flexible tool for automated behavioral detection based on markerless pose tracking," *eLife*, vol. 11, p. e74314, Aug. 2022, publisher: eLife Sciences Publications, Ltd. [Online]. Available: https://doi.org/10.7554/eLife.74314

[14] C. Raman, J. Vargas Quiros, S. Tan, A. Islam, E. Gedik, and H. Hung, "ConfLab: A Data Collection Concept, Dataset, and Benchmark for Machine Analysis of Free-Standing Social Interactions in the Wild," Oct. 2022, version Number: 4. [Online]. Available: https://data.4tu.nl/collections/_/6034313/4

[15] josedvq, "lared-laughter," https://github.com/josedvq/lared-laughter, 2022.

[16] Joelle225, "data-expl," https://github.com/Joelle225/data-expl, 2025.

[17] MMAction2 Contributors, "OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark," Jul. 2020. [Online]. Available: https://github.com/open-mmlab/mmaction2

[18] M. A. Raza, L. Chen, L. Nanbo, and R. B. Fisher, "EatSense: Human centric, action recognition and localization dataset for understanding eating behaviors and quality of motion assessment," *Image and Vision Computing*, vol. 137, p. 104762, Sep. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885623001361

[19] KNAW, NFU, NWO, TO2-Federatie, Vereniging Hogescholen, and VSNU, "Nederlandse gedragscode wetenschappelijke integriteit," 2018. [Online]. Available: https://phys-techsciences.datastations.nl/dataset.xhtml?persistentId=doi:10.17026/dans-2cj-nvwu