Enhancing Wearable Tap Water Audio Detection Through Subclass Annotation in the HD-Epic Dataset

Robin Burchard
robin.burchard@uni-siegen.de
University of Siegen
Siegen, Germany

Kristof Van Laerhoven
kvl@eti.uni-siegen.de
University of Siegen
Siegen, Germany

Abstract

Wearable human activity recognition has been shown to benefit from the inclusion of acoustic data, as the sounds around a person often contain valuable context. However, due to privacy concerns, it is usually not ethically feasible to record and save microphone data from the device, since the audio could, for instance, also contain private conversations. Rather, the data should be processed locally, which in turn requires processing power and consumes energy on the wearable device. One special use case of contextual information that can be utilized to augment special tasks in human activity recognition is water flow detection, which can, e.g., be used to aid wearable hand washing detection. We created a new label called *tap water* for the recently released HD-Epic data set, creating 717 hand-labeled annotations of tap water flow, based on existing annotations of the *water* class. We analyzed the relation of *tap water* and *water* in the dataset and additionally trained and evaluated two lightweight classifiers to evaluate the newly added label class, showing that the new class can be learned more easily.

CCS Concepts

• **Human-centered computing** → *Ubiquitous and mobile computing*.

Keywords

Audio Activity Recognition; Wearable Computing; Dataset; Tap Water; Audio Event Detection; HD-Epic; CNN

1 Introduction

Human Activity Recognition (HAR) is an increasingly multimodal field [18, 26, 28]. In addition to the traditionally used IMUs and RGB-cameras, a multitude of other sensing modalities can be employed

to add context and additional information to the existing HAR systems. Such modalities include, but are not limited to, environmental sensors (humidity, pressure, temperature, brightness) [1, 3], biosignals (e.g., heart rate, electrodermal activity, electromyography), or proximity sensors [9].

Another promising example of an additional modality is sound, which can be recorded with commercially available microphones with small effort and low cost. Acoustic activity recognition and sound event classification are two subproblems one might want to approach when using microphones. Acoustic sensing can be employed to standalone classifiers [15, 19]. While RGB(d)-cameras perform worse when there are many occlusions, acoustic and IMU-based models are challenged by ambiguities. Huh et al. describe this fact with the words: "Some actions are audibly indistinguishable, e.g. 'wash carrot' vs 'wash tomato', as it is impossible to determine which vegetable is being washed through sound alone." [13, 14].

Rather than relying on sound alone, recent works propose using sounds as an additional modality to IMUs, RGB-cameras, or other systems [4, 6, 29]. With the added context, systems could become more reliable in picking up unusual instances of specific activities and also become more robust against false positives. One example of this is the problem of detecting hand washing among similar activities using IMU sensors: The typical pattern of hand washing often involves rapid and repetitive movements. However, these patterns are not always present for every hand washing instance, and similar patterns can also appear in other contexts, e.g., when brushing one's teeth [2]. With a microphone, the sound of the activated tap, as well as the water's "splashing" could be picked up [25, 30]. By including the information, whether a running tap could be detected nearby, both the rates of false negative and false positive detections could likely be lowered. A similar solution is allegedly used for the Apple Watch's hand-washing timer feature, but the implementation remains closed source [12, 17].

Existing sound datasets often include water-related sounds in one way or another, but the distinction of *tap water* from other water sounds is included in few. However, since many existing datasets already contain the relevant microphone recordings, just no specific annotations, we were able to use a pre-existing dataset for this work, by adding a new label class for *tap water* and annotating it accordingly.

Applications. The previously mentioned detection of hand washing is only one application of tap water detection, other applications include the monitoring of cooking (e.g., step detection, washing vegetables, rinsing dishes), cleaning, other hygiene routines (e.g., brushing teeth), bathroom usage, and resource usage monitoring (i.e. detecting water wastage). In all these fields, a wearable device could be used to support the user by tracking or habit coaching

them, or help them to conserve water. Treating *tap water* as its own separate class has multiple advantages. Tap water is strongly associated with some specific, intentional, high-relevance activities, like the ones mentioned above. In contrast, the general *water* label in audio datasets includes diverse sounds like rain, streams, splashes, gurgling water in pipes, and pouring. These water-related sounds vary greatly in frequency, duration, and intensity. *Tap water* is more homogeneous, and therefore easier to learn for machine learning models, as we show in this work.

An additional constraint of recording audio data is protecting all users' and non-users' privacy. Since microphone recordings could accidentally include private conversations, clues about the users' whereabouts, or other personal information, they have to be handled with a high level of care. Ideally, audio recordings should never leave the recording device and should be immediately removed after processing. This would ensure the highest possible protection of the users' privacy. To make this possible, classifiers should be small enough and require little complexity, as mobile devices need to be powerful enough to run the required model inference. Focusing on the task of tap water detection, we propose to train models using our new annotations, which we created specifically for this task, on the existing, but also very recently released, HD-Epic dataset.

The main goal of this work was to create a precisely annotated dataset for the detection of running tap water from microphone data. The dataset can be employed for automated tap water detection on wearable devices, which in turn has a multitude of applications.

Our main contributions in this paper are as follows:

(1) We motivate why on-device *tap water* detection is a crucial task in activity recognition.
(2) We hand-labeled a precise, new class in the HD-Epic dataset, resulting in 717 instances of *tap water* audio events.
(3) We show how this new *tap water* class compares to the existing *water* class, demonstrating how HD-Epic's sound annotations can be used for audio-based activity recognition
(4) We created a pre-trained CNN ExecuTorch model to detect *tap water* for deployment on wearables

Ensuring reproducibility: All code, data, annotations, and other information needed to reproduce the results of this work can be found online, in our GitHub repository[1].

2 Related Work

Several annotated sound datasets containing *water* or *tap water* have been published in the past. Table 1 allows a quick comparison between the available datasets and our work. Notably, Perett et al. have published HD-Epic, which is used as the basis of this work [20]. Before that, an earlier version called Epic Sounds [13, 14] was released as a derived work of the EPIC-KITCHENS-100 dataset [7]. These two datasets contain *water* as a class, but they avoid distinguishing it further into different sub-classes of water-related sounds. The ESC-50 dataset for environmental sound classification (ESC) [21] by Piczak et al. contains a collection of water-related sound categories like *rain*, *sea waves*, *water drops*, *pouring water*, *toilet flush*, and *thunderstorm*. However, the dataset does not contain *tap water* as a class. AudioSet by Gemmeke et al. [10] contains 10s audio

sample clips drawn from YouTube videos, which were categorized into 527 sound event classes. The dataset includes 2442 of these 10s segments, in which *tap water* is present. However, the label for each 10s audio clip only provides the information that at some point in the clip *tap water* is present, and only $1\,Hz$ embeddings of the audio data are published. Lastly, Fonseca et al. published the FSD50K dataset [8], which contains 51 thousand audio clips from Freesound, which were manually labeled into a selection of 200 classes of AudioSet's ontology. This dataset contains 458 *tap water* clips. Due to the samples being collected from Freesound, different licenses can apply, and the sounds in the FSD50K dataset are usually very specific and distinct audio clips of just one annotated class, without background noise.

To conclude, each of the listed audio datasets has important limitations. Whether due to restricted access, the absence of precise timestamp annotations, or a lack of realistic background sounds that reflect real-world conditions.

In the domain of audio event classification, many different audio event recognition systems exist. Like in many machine learning domains, feature extraction is an important part of such a system's pipeline, and features usually include Fourier transforms, cepstral features, spectrograms, wavelet transforms, and statistical properties of the audio signal [6, 23]. For training classic machine learning models, such as support vector machines or random forests, spectral and statistical features are used [5, 22]. To train deep learning models, existing research often relies on log-mel spectrograms as the input to a convolutional neural network (CNN) [16, 27].

3 Methods

In this section, we report how we created the new label class *tap water* on a pre-existing dataset. Furthermore, we describe our findings from the labeling process and the validation of the newly created annotations. Lastly, we outline the creation of an automated detection system for *tap water* detection on mobile devices using lightweight machine learning algorithms.

3.1 Dataset

We used the publicly available HD-EPIC dataset [20] with an adapted label class. The original dataset contains activities from nine participants, together with video and audio data, recorded at the participants' homes. We focus solely on the audio data, for which the dataset contains 50,968 audio annotations of 44 classes, such as *click*, *rustle*, or *water*. We chose the EPIC-HD dataset because it offers high audio quality (48 kHz sampling rate) and has a high quality of annotations. The dataset is also recorded in-the-wild, in nine different indoor environments, which helps to avoid overfitting to a specific environment. However, none of the classes in the existing annotations align perfectly with our focus on acoustic tap water flow detection. Thus, we used the *water* and *pour* classes as a starting point to annotate a new class, *tap water*.

3.2 Label creation for the new class *tap water*

A single person, the first author of this publication, manually annotated the entire dataset of nine participants. To do so, we did not go through all video recordings again at their entire length, but rather focused on improving the existing labels of *water* and *pour* for our

---

[1]https://github.com/RBurchard/hd-epic-tap

Table 1: Audio datasets containing water-related annotations. The table shows the duration of contained *tap water* annotations, the datasets' sampling frequencies, whether the dataset is open (raw data freely available), the source of the audio data, and whether the dataset is labeled accurately with start/end times for each sound (precise labels). We created 157 minutes of precise annotations on the existing HD-Epic dataset. AudioSet provides its data as 128-dimensional audio embeddings extracted at 1 Hz rather than raw audio.

dataset	tap water duration (min.)	add. water classes	freq. (kHz)	open	source	precise labels
Epic Sounds	None	water / pour	24	x	EPIC-Kitchens-100	x
HD-Epic	None (orig.) / 157 (ours)	water / pour	48	x	HD-Epic	x
AudioSet	407, in 10s segments	many, rain, streams, etc.	1 Hz (embeddings)	-	YouTube	-
FSD50K	~53, in 458 clips	waterfall, waves, etc.	variable, 44.1+	x	Freesound	-
ESC-50	None	rain, waves, etc.	44.1	x	Freesound	-

task. This was based on the idea that *tap water* should be a subset of *water* and *pour*. We include *pour*, because the word *pour* can also imply water and tap water flowing, depending on interpretations of natural language. However, we found that the original annotations did not contain any tap water under *pour* labels. During label creation, we gained some valuable insights into the audio aspect of HD-Epic, which we also describe in this subsection.

3.2.1 Description of labeling criteria and initial results. The original HD-Epic dataset contains 2974 *water* labels, and 171 instances of *pour*. After analyzing the durations of the separate labels, we found that roughly two-thirds of the labels were so short, that they only made up for around 11 % of the total *water* duration. We argue that these short *water* labels are unlikely to contain *tap water* annotations, or even if they do, they are so short that they would have little positive influence on the training and evaluation of a classification model. Therefore, we decided to only relabel labels with a duration of three seconds or more. This cut-off meant that we relabeled only slightly more than one-third of the instances of *water*, but we could still cover almost 90 % of the time that was labeled as an instance of this class. Out of the initial 2974 *water* and 171 *pour* annotations, we investigated 1058 *water* and 40 *pour* annotations that were longer than three seconds.

We created a tool to jump to *water* and *pour* labels in the recordings, and then, together with the videos, inspected and annotated all occurrences of tap water being audible in the dataset. In the HD-Epic dataset, *water* refers not only to taps running, but to all water-related sounds, including but not limited to the sound of single water drops, drain pipes gurgling, water splashing, and sometimes even the stirring of a pot. Therefore, to obtain *tap water* annotations, we hat to many *water* annotations had to be ignored, shortened, or split up into multiple parts.

After manually labeling the videos of all 9 participants, we received a total of 717 new annotations for *tap water*, with a total duration of 9595 seconds. Leaving out newly created labels that were shorter than 3 seconds, 9439 seconds of labels were created. In total, the ratio of the durations of *tap water* labels compared to *water* labels is 62 %, which means that while *tap water* is the largest subset of the *water* class, it is also significantly smaller. Thus, we argue that it should be treated as a different class than *water*, to allow for higher granularity and better usability for augmenting specific tasks, like hand washing detection.

Table 2 shows the durations and ratios of annotated times for the *water* and *tap water* label classes. In total, HD-Epic contains

16981 seconds of *water*, out of which 15182 seconds were used by us for analysis and creating the new label class *tap water*.

3.2.2 HD-Epic Dataset Findings. Overall, the quality of sound labels in the HD-Epic dataset is high. Whilst investigating and augmenting the originally contained audio labels, we obtained some learnings regarding these original audio labels, mostly related to the water class.

Miss-classifications. There are many sounds present in a kitchen environment that can sound similar to water, and that were therefore occasionally incorrectly labeled as *water* in the dataset. Some examples include stirring a pot, whisking eggs, or the sound of sizzling pans in the background. Additionally, plating non-liquid food from a pan, moving food into a container, or an instance of *pour* without water present were annotated as *water*. Additionally, we found *pour* annotations when a participant was frothing milk in one case, and handling a plastic bag in another case.

Inconsistencies. We found some inconsistencies in the labels, where similar sounds were treated differently during the label creation. As an example, single drops of water in the background were sometimes labeled and sometimes missed throughout the different recordings and participants. The same problem occurred with water in the background in general. For participant P01, water in the drain and audible pipes were annotated, but not for most of P02's recordings. In some cases, pouring water was annotated as *pour*, and in other cases it was annotated as *water*. Sometimes, multiple water labels were overlapping or contained inside another.

Diversity of water sounds. The *water* class contains a multitude of different, water-related sounds. The sound of a running tap usually comes with a distinct "noise", whereas other water-related sounds can be less specific. Single water droplets make a distinct, short sound, whilst pipes "gurgling" in the background produce a completely different sound. Between the sounds of pouring water in a cup, "splattering" water in a sink, and a pressure cooker releasing steam, there is a lot of auditory diversity. Although we fully agree that all these sounds should be grouped under the *water* label, we argue that it is naturally difficult to train a classification model for such a diverse class of many-shaped audio labels. Therefore, we argue that the less diverse and more consistent class *tap water* is a valuable addition to the HD-Epic dataset.

3.2.3 Auditory challenges with the new labels. In the cases of some of the newly created *tap water* audio annotations, issues can be

Table 2: Durations of *water* and *tap water* labels in the dataset, per participant and aggregated. The table splits the durations up into all labels and labels lasting three seconds or longer. The ratio of the remaining labels, if we remove labels shorter than three seconds, is shown in the third column from the right. On aggregate, 89% of the duration of *water* annotations are kept, and 98% of tap water annotations. The ratio between tap water and water annotations is shown in the rightmost columns.

| participant id | duration (s) | | duration (s) >= 3s | | ratio >= 3s | | ratio | ratio >= 3s |
	tap water	water	tap water	water	water	tap water	tap water / water	tap water / water
P01	895.38	2737.15	832.48	2491.81	0.91	0.93	0.33	0.33
P02	765.34	2189.15	752.25	1758.26	0.80	0.98	0.35	0.43
P03	1821.17	3261.02	1787.52	3025.74	0.93	0.98	0.56	0.59
P04	1474.32	1872.03	1457.36	1732.96	0.93	0.99	0.79	0.84
P05	297.81	541.46	294.93	427.84	0.79	0.99	0.55	0.69
P06	1432.18	1588.87	1429.38	1458.60	0.92	1.00	0.90	0.98
P07	789.91	944.50	789.91	844.36	0.89	1.00	0.84	0.94
P08	990.47	1838.34	984.14	1754.68	0.95	0.99	0.54	0.56
P09	1128.00	2008.62	1111.22	1687.98	0.84	0.99	0.56	0.66
Agg.	9594.57	16981.14	9439.19	15182.24	0.89	0.98	0.57	0.62

predicted from the characteristics of the contained sounds. Participant P02, e.g., was using a very silent tap, with water sometimes running in the background, far away from the used microphones. Additionally, P02 and P04 sometimes only turned on their taps to a slight drizzle, which was barely audible, and ran without the often-present "water flow noise".

In other cases, loud and similar-to-water noises were in the background. For P03 and P06, extractor hoods were sometimes running in the background, making the tap hard to hear.

3.3 Analysis and validation of the new labels

Intersection over Union. To better understand the similarity of the *tap water* class to the *water* class, we report the intersection over union (IoU) of the two classes, calculated and aggregated over each *water* annotation (see Eq. 1). The calculation of the IoU is done by merging all overlapping *water* and *tap water* regions and then aggregating their IoUs over the entire dataset.

$$\text{IoU} = \frac{|(\texttt{tap water} \cap \texttt{water})|}{|(\texttt{tap water} \cup \texttt{water})|} \tag{1}$$

Table 3 shows the IoU per participant and over the entire dataset. The IoU spans from 0.364 to 0.916 with a total IoU of 0.616 calculated on the dataset.

Coverage of tap water by water. We also report the coverage in Table 3, which is defined as the ratio of containment of *tap water*, i.e., the ratio of *tap water* that is contained inside *water* annotations (see Eq. 2).

$$\text{coverage} = \frac{|(\texttt{tap water} \cap \texttt{water})|}{|(\texttt{tap water})|} \tag{2}$$

The coverage is in the range of 93 % to 99.9 % across all participants, and has a value of 97.8 % for the entire dataset. The coverage value is high, as we only annotated near the *water* labels of the original dataset. However, the value also supports the initial assumption that *tap water* is a subclass of *water*, as it is mostly covered by *water* annotations.

To conclude this subsection, our initial validation of the *tap water* labels shows that we successfully annotated a subclass of *water*,

and that *tap water* makes up for around 63 % of all water-related sound in HD-Epic's *water* annotations.

3.4 Machine Learning Experiments

To validate the audio detection of *tap water*, we trained machine learning classifiers, and additionally compared the results with the same classifiers trained on *water*.

3.4.1 Classification tasks. We define the task of classifying *water* or *tap water* as the positive class against all other classes as the Null class. We report the results both for a 70 − 30 % train-test-split over all participants (Task A), as well as for a more realistic leave-one-participant-out (LOPO) paradigm (Task B). The train-test-split shows how well the classifiers perform for each of the positive classes under conditions where individual-specific patterns may be learned during training. The LOPO split yields a more realistic estimate of real-world performance: Since the participant and environment used for evaluation are not shown to the model during training, the performance is indicative of the real-world performance we could expect for unseen environments.

3.4.2 Features. Similar to previous publications, we used frequency-domain feature vectors. We focused on the Mel-Frequency Cepstral Coefficients (MFCCs), spectral features such as centroid, bandwidth, contrast, rolloff, cover ([11]), as well as the chromagram, zero crossing rate, and the root-mean-square energy (RMSE) of the audio signal. We calculated each of these for 2-second-long sliding windows without overlap. The features were calculated using the implementations contained by the "librosa" python library, and then aggregated for each window, using the mean, sd, min, max, and median as aggregation functions. In total, this left us with a 41-dimensional feature vector.

In line with the literature, we calculated log-mel spectrograms for the use in a neural network [24, 27]. Example spectrograms are shown in Figure 1, where the intervals chosen show the broad spectrum of "noise" that we can expect from tap water sounds.

3.4.3 Classification models. For our intended use of wearable *tap water* detection, models must be small and computationally simple enough to run on mobile CPUs, i.e. we have to refrain from using

Table 3: IoU: Intersection over Union between the original *water* annotations and the newly created *tap water* class. Coverage: Ratio of containment of *tap water* annotations inside *water*.

participant	P01	P02	P03	P04	P05	P06	P07	P08	P09	All
IoU	0.364	0.429	0.614	0.842	0.651	0.871	0.916	0.542	0.618	0.616
coverage	0.999	0.995	0.991	0.983	0.976	0.927	0.993	0.984	0.968	0.978
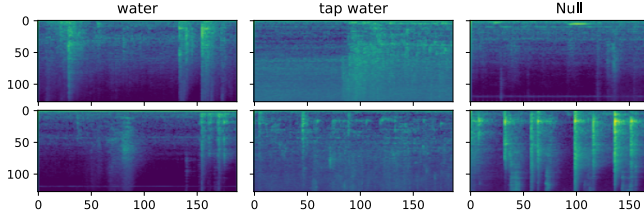
Figure 1: Two examples per class of the log-mel spectrograms, used as input for the CNNs. The spectrograms for the *water* class are chosen as examples that are not also in *tap water*.

large transformer models. Since the scope of this work is to show the usefulness of the newly added class, and not to achieve the highest possible results, we stuck to two simple but appropriate classification models.

We used a random forest (RF) classifier and trained it on the features described in 3.4.2. We used the standard settings of sklearn's implementation, except "class_weight" which we set to "balanced" to take the skewed distribution into account, and "n_estimators" which we set to 200.

Similar to the literature [27], we trained a CNN on the log-mel spectrograms as described in 3.4.2. The CNN consisted of five 2D-convolutional layers, each followed by a 2x2 max-pooling layer. Two fully connected layers form the classification head of the network architecture. We trained the CNN model using the ADAM optimizer and weighted binary cross-entropy as the loss function.

Since the dataset is highly imbalanced for our target classes, we also report a baseline performance, i.e. the best-performing "dummy classifier", w.r.t. the F1 score. In our case, the "uniform" prediction strategy yielded the best baseline performance throughout all experiments.

3.4.4 Reported metrics. We report the accuracy, as well as F1 score, precision, and recall for all tasks, models, and target classes. Since *water* and *tap water* only make up for less than 10 % of the dataset, accuracy alone is not an expressive metric. Hence, we also report the F1 score and its components, precision and recall, as they provide more information about the actual predictive results. Also, to make the comparison of the performance for the two target classes easier, we calculate the ratio between the classifier performance and the baseline performance for the F1 score metric (R. to Bln.). This performance ratio better indicates the relative performance of the classifier w.r.t. to the general difficulty of the task.

3.4.5 ExecuTorch runtime model. To share the models created in this work, we have trained one CNN model on all available data and made it available pre-trained for ExecuTorch, PyTorch's mobile deployment solution. The model can be used to detect *tap water*. As input, it requires windows of computed log-mel spectrograms, as described above. For further details, we refer to our repository.

4 Results

Table 4 and Figure 2 show the results for Task A and Task B. For Task B, the table only reports aggregated results, while Figure 2 includes the results per participant. Overall, the trained classifiers' prediction performance on the *tap water* class is higher than the performance on the *water* class. The low baseline F1 scores of 0.11 for *tap water* and 0.17 for *water* are indicative of the dataset's natural imbalance. These baseline results show that it is very difficult to achieve good F1 scores because the class distribution is strongly imbalanced.

For Task A, the Random Forest performs slightly better than the CNN, and for both classifier types, the task of classifying *tap water* against all other classes yields the highest performance, in terms of F1 score, Accuracy, and Precision. The CNN consistently outperforms the Random Forest with its Recall results. The highest F1 scores reached are 0.75 on *tap water* and 0.71 on *water* (both CNN). Due to the distribution, reaching the same F1 score is more difficult for the *tap water* class. However, the general results with a higher F1 score for *tap water* in Task A support our expectation that the *tap water* class is more homogeneous and therefore easier to learn and predict. This becomes even more apparent when we consider the ratio to the baseline (R. to Bln.) results. For the *tap water* class, the classifiers outperform the baseline by 668 % (CNN) and 656 % (RF), while for *water* the ratios are only 410 % (CNN) and 391 % (RF).

The performance for the LOSO-CV split of Task B differs strongly from participant to participant. While the baseline results are relatively similar for most participants, they are lower for P05 for both classes, (F1: *water*: 0.08, *tap water*: 0.05). P05 is also the participant for which the classifiers achieve the worst cross-validation F1 scores, all around 0.3. The generalization performance to the unseen participants' data also varies per classifier, i.e., sometimes the CNN has a slight edge over the Random Forest, and vice versa. The best performances are reached for P06, where the CNN achieves an F1 score of 0.73 for *water* and the Random Forest reaches 0.66 for *tap water*. On average, the CNN generalizes slightly better on both *tap water* (CNN: 0.48, RF: 0.47) and *water* (CNN: 0.52, RF: 0.45). The results for all participants are shown in Figure 2. While the results for the F1 score are similar across classifiers and target class, the ratio to the baseline is still better for *tap water* with a ratio of 447 % (CNN) compared to 315 % for *water* (also CNN). A Shapiro-Wilk test (stat = 0.934, p = 0.226) indicated normality. Both a paired t-test (p = 0.0005) and Wilcoxon test (p = 0.0019) showed that *tap water* significantly outperformed *water* w.r.t. the ratio to the baseline.

Notably, the RF achieves higher accuracy and precision results throughout the experiments, and, in turn, the CNN achieves higher recall scores.

5 Discussion and Future Work

We discussed that *tap water* detection has direct applications in smart home systems, hygiene monitoring (e.g., handwashing), and

Table 4: Results of Task A: Random Train-Test-Split and Task B: Leave-One-Participant-Out Cross-Validation.

Target Class	Classifier	Task A: Random Train-Test-Split					Task B: LOSO-CV				
		F1 Score	Acc.	Precision	Recall	R. to Bln.	F1 Score	Acc.	Precision	Recall	R. to Bln.
Tap Water	Baseline	0.11	0.50	0.06	0.49	–	0.11	0.50	0.07	0.51	–
	CNN	**0.75**	0.97	0.75	**0.76**	**668.29 %**	**0.48**	0.92	0.48	**0.54**	**446.83 %**
	Random Forest	0.74	**0.97**	**0.92**	0.62	655.68 %	0.47	**0.95**	**0.76**	0.35	427.39 %
Water	Baseline	0.17	0.50	0.10	0.50	–	0.17	0.50	0.10	0.51	–
	CNN	**0.71**	0.94	0.73	**0.69**	**409.63 %**	**0.52**	0.89	0.51	**0.60**	**314.56 %**
	Random Forest	0.68	**0.95**	**0.92**	0.53	391.29 %	0.45	**0.92**	**0.78**	0.33	280.98 %
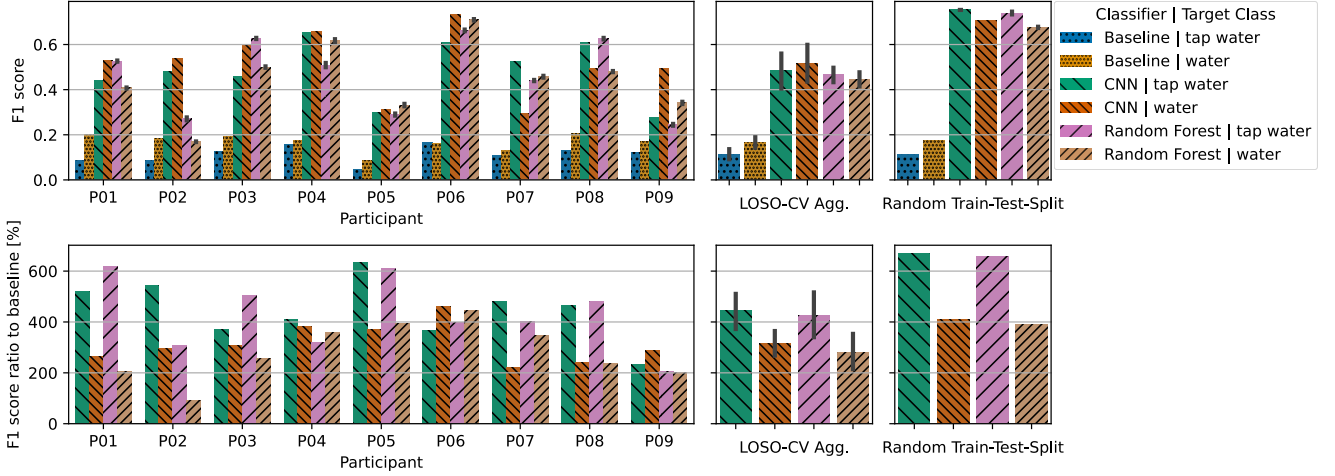
Figure 2: Results of classifier training and evaluation for Task A and Task B. F1 score results (upper row) are displayed for the leave-one-participant-out cross-validation, per participant (and thus per environment, left) and aggregated (center), as well as for Task A (random train-test-split, right). The baselines are consistently outperformed by both CNN and Random Forest, for both target classes. We also show the ratio of the F1 score to the respective baseline F1 score (lower row), for the same splitting modalities (left: Task B and right: Task A). Classifiers for *tap water* mostly achieve higher ratios than classifiers for *water*.

resource conservation. Especially for HAR tasks, the general presence of *water* sounds contains less contextual information than *tap water* being audible.

Through the statistics and validation analysis presented in 3.2.1 and 3.3, it becomes apparent that *tap water* is mostly a subclass of *water* in the HD-Epic dataset. Its duration in the dataset makes up for around 60 % of the duration of all *water*. While *tap water* can also represent diverse sounds, depending on the tap used and the water flow intensity, it appears to be more homogeneous than *water*. *Water* represents a superclass of multiple liquid-related sounds that are related, but oftentimes not specific enough for possible use cases, especially in HAR.

The machine learning results presented in Section 4 also hint at *tap water* being easier to learn by classifiers. Especially when compared to the baseline performance, the resulting F1 scores for *tap water* are consistently higher than the F1 scores for *water*. We used lightweight classifiers and did not extensively optimize the model training. In future work, we want to analyze failure cases and relate them to the challenges listed in Section 3.2.3, searching for potential solutions. A higher raw performance could likely be achieved using state-of-the-art edge-optimized classifiers. In a future application, prediction post-processing, e.g., smoothing, could be used to improve the detection results, as the smoothing would likely be able to filter out outliers in the prediction.

6 Conclusions

This work describes how we created and evaluated a publicly available dataset of precise *tap water* audio annotations in real-world kitchen environments. We based our work on the HD-Epic dataset, which provided a well-annotated basis with a generic *water* class.

We systematically evaluated how the *tap water* audio annotation class is mainly a subclass of *water*, and argued why it is useful to consider it separately. Applications for the newly created labels can be found in a multitude of tasks, such as hand washing detection, hygiene monitoring and compliance, or resource usage surveillance.

By reporting initial classification performance results, we show that despite it representing a less frequent class in the dataset, training classifiers on the more uniform *tap water* is not more difficult than on the *water* class. Especially when compared to the baseline performance, models trained to detect *tap water* achieve higher results.

In this work, we additionally discussed multiple applications of *tap water* detection in acoustic sensing.

We conclude that the created 717 labels provide a valuable contribution to the audio labels of HD-Epic and can, in future work, be used to train off-line classifiers for *tap water* detection. All annotations, code for generating the results, and the pre-trained model are available in our GitHub repository[1].

References

[1] Arnab Barna, Abdul Kadar Muhammad Masum, Mohammad Emdad Hossain, Erfanul Hoque Bahadur, and Mohammed Shamsul Alam. 2019. A Study on Human Activity Recognition Using Gyroscope, Accelerometer, Temperature and Humidity Data. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (Cox'sBazar, Bangladesh, 2019-02). IEEE, 1–6. doi:10.1109/ECACE.2019.8679226

[2] Robin Burchard, Philipp M. Scholl, Roselind Lieb, Kristof Van Laerhoven, and Karina Wahl. 2022. WashSpot: Real-Time Spotting and Detection of Enacted Compulsive Hand Washing with Wearable Devices. In *Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Cambridge United Kingdom, 2022-09-11). ACM, 483–487. doi:10.1145/3544793.3563428

[3] Robin Burchard and Kristof Van Laerhoven. 2025. Multi-Modal Atmospheric Sensing to Augment Wearable IMU-Based Hand Washing Detection. In *Sensor-Based Activity Recognition and Artificial Intelligence* (Cham, 2025), Orhan Konak, Bert Arnrich, Gerald Bieber, Arjan Kuijper, and Sebastian Fudickar (Eds.). Springer Nature Switzerland, 55–68. doi:10.1007/978-3-031-80856-2_4

[4] Chieh-Feng Cheng, Abbas Rashidi, Mark Davenport, and David Anderson. 2016. Audio Signal Processing for Activity Recognition of Construction Heavy Equipment. doi:10.22260/ISARC2016/0078

[5] Selina Chu, Shrikanth Narayanan, and C.-C. Jay Kuo. 2009. Environmental Sound Recognition With Time–Frequency Audio Features. 17, 6 (2009), 1142–1158. doi:10.1109/TASL.2009.2017438

[6] Stefania Cristina, Vladimir Despotovic, Rodrigo Pérez-Rodríguez, and Slavisa Aleksic. 2024. Audio- and Video-Based Human Activity Recognition Systems in Healthcare. 12 (2024), 8230–8245. doi:10.1109/ACCESS.2024.3353138

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2022. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. 130, 1 (2022), 33–55. doi:10.1007/s11263-021-01531-2

[8] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2022. FSD50K: An Open Dataset of Human-Labeled Sound Events. 30 (2022), 829–852. doi:10.1109/TASLP.2021.3133208

[9] Biying Fu, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. 2020. Sensing Technology for Human Activity Recognition: A Comprehensive Survey. 8 (2020), 83791–83820. doi:10.1109/ACCESS.2020.2991891

[10] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017-03). 776–780. doi:10.1109/ICASSP.2017.7952261

[11] Patrice Guyot, Julien Pinquier, and Régine André-Obrecht. 2012. Water Flow Detection from a Wearable Device with a New Feature, the Spectral Cover. In *2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI)* (2012-06). 1–6. doi:10.1109/CBMI.2012.6269814

[12] Dade Hayes. 2020. *Apple Watch Can Now Tell If You're Washing Your Hands.* Deadline. https://deadline.com/2020/06/apple-watch-can-tell-if-you-are-washing-your-hands-coronavirus-1202966667/

[13] Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. 2023. Epic-Sounds: A Large-Scale Dataset of Actions That Sound. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2023-06). 1–5. doi:10.1109/ICASSP49357.2023.10096198

[14] Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. 2025. EPIC-SOUNDS: A Large-Scale Dataset of Actions that Sound. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2025).

[15] Kyuwoong Hwang and Soo-Young Lee. 2012. Environmental Audio Scene and Activity Recognition through Mobile-Based Crowdsourcing. 58, 2 (2012), 700–705. doi:10.1109/TCE.2012.6227479

[16] Seung Ho Hyun. 2023. Sound-Event Detection of Water-Usage Activities Using Transfer Learning. 24, 1 (2023), 22. doi:10.3390/s24010022

[17] Apple Inc. 2024. *Set up Handwashing on Apple Watch.* Apple Support. https://support.apple.com/en-my/guide/watch/apdc9b9f04a8/watchos

[18] Misha Karim, Shah Khalid, Aliya Aleryani, Jawad Khan, Irfan Ullah, and Zafar Ali. 2024. Human Action Recognition Systems: A Review of the Trends and State-of-the-Art. 12 (2024), 36372–36390. doi:10.1109/ACCESS.2024.3373199

[19] Dawei Liang and Edison Thomaz. 2019. Audio-Based Activities of Daily Living (ADL) Recognition with Large-Scale Acoustic Embeddings from Online Videos. 3, 1 (2019), 1–18. arXiv:1810.08691 [cs] doi:10.1145/3314404

[20] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. 2025. *HD-EPIC: A Highly-Detailed Egocentric Video Dataset.* arXiv:2502.04144 [cs] doi:10.48550/arXiv.2502.04144

[21] Karol J. Piczak. 2015. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd ACM International Conference on Multimedia* (Brisbane Australia, 2015-10-13). ACM, 1015–1018. doi:10.1145/2733373.2806390

[22] Ivan Miguel Pires, Nuno M. Garcia, Nuno Pombo, and Francisco Flórez-Revuelta. 2017. *User Environment Detection with Acoustic Sensors Embedded on Mobile Devices for the Recognition of Activities of Daily Living.* arXiv:1711.00124 [cs] doi:10.48550/arXiv.1711.00124

[23] Arjun Prashanth, S. L. Jayalakshmi, and R. Vedhapriyavadhana. 2024. A Review of Deep Learning Techniques in Audio Event Recognition (AER) Applications. 83, 3 (2024), 8129–8143. doi:10.1007/s11042-023-15891-z

[24] Soonshin Seo, Changmin Kim, and Ji-Hwan Kim. 2022. Convolutional Neural Networks Using Log Mel-Spectrogram Separation for Audio Event Classification with Unknown Devices. 21, 2 (2022), 497–522. doi:10.13052/jwe1540-9589.21216

[25] Babak Taati, Jasper Snoek, David Giesbrecht, and Alex Mihailidis. 2010. Water Flow Detection in a Handwashing Task. In *2010 Canadian Conference on Computer and Robot Vision* (2010-05). 175–182. doi:10.1109/CRV.2010.30

[26] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. 2019. Deep Learning for Sensor-Based Activity Recognition: A Survey. 119, C (2019), 3–11. doi:10.1016/j.patrec.2018.02.010

[27] Ah Hyun Yuh and Soon Ju Kang. 2021. Real-Time Sound Event Classification for Human Activity of Daily Living Using Deep Neural Network. In *2021 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)* (2021-12). 83–88. doi:10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics53846.2021.00027

[28] Shibo Zhang, Yaxuan Li, Shen Zhang, Farzad Shahabi, Stephen Xia, Yu Deng, and Nabil Alshurafa. 2022. Deep Learning in Human Activity Recognition with Wearable Sensors: A Review on Advances. 22, 4 (2022), 1476. Issue 4. doi:10.3390/s22041476

[29] Yunhua Zhang, Hazel Doughty, Ling Shao, and Cees G. M. Snoek. 2022. Audio-Adaptive Activity Recognition Across Video Domains. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA, USA, 2022-06). IEEE, 13781–13790. doi:10.1109/CVPR52688.2022.01342

[30] Haoyu Zhuang, Liqiang Xu, Yuuki Nishiyama, and Kaoru Sezaki. 2023. Detecting Hand Hygienic Behaviors In-the-Wild Using a Microphone and Motion Sensor on a Smartwatch. In *Distributed, Ambient and Pervasive Interactions* (Cham, 2023) *(Lecture Notes in Computer Science)*, Norbert A. Streitz and Shin'ichi Konomi (Eds.). Springer Nature Switzerland, 470–483. doi:10.1007/978-3-031-34609-5_34