Multimodal Hand Washing Detection Using IMU and Humidity Data: A Comparative Study of Supervised and Self-Supervised Learning

SONALI MANOHARAN, Universität Siegen, Germany

After the huge pandemic of 2020, hand hygiene monitoring is important for reducing infectious disease transmission. Automatic hand washing detection from wearable sensors remains difficult due to inter-subject variability, class imbalance, and limited labeled data. We study multimodal hand washing detection using IMU and environmental humidity data from 20 participants, comparing supervised and self-supervised learning under Leave-One-Subject-Out (LOSO) evaluation. As a supervised baseline, we train a CNN-LSTM on IMU magnitude and humidity slope features. For self-supervised learning, we use a pretrained TS2Vec encoder (frozen) on IMU windows and fuse its embeddings with handcrafted humidity features before training an MLP classifier. We also evaluate TS2Vec without humidity to measure the effect of multimodal fusion. Because of class imbalance, overall accuracy is high across models; F1 for the minority class (hand washing) is the main performance metric. Results show how learning paradigm and sensor fusion affect cross-subject generalization and highlight the role of humidity in improving detection robustness.

Additional Key Words and Phrases: Time series representation learning, Leave-One-Subject-Out, Class imbalance, Wearable sensing, Sensor fusion

ACM Reference Format:
Sonali Manoharan. 2026. Multimodal Hand Washing Detection Using IMU and Humidity Data: A Comparative Study of Supervised and Self-Supervised Learning. 1, 1 (February 2026), 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In healthcare settings, hand hygiene prevents the transmission of harmful germs to patients and the environment [World Health Organization 2009]. But despite hand-washing protocols, compliance is inconsistent. Automated monitoring systems with wearable sensors have the potential to improve hygiene in the workplace, particularly in health care and food handling, by automatically detecting hand-washing events, thus removing the need for staff to manually report them.

We investigate supervised vs. self-supervised learning approaches for multimodal hand-washing detection, comparing their performances across sensing modalities. We evaluate a supervised CNN–LSTM baseline model, trained on IMU magnitude and humidity-slope features, and a self-supervised TS2Vec model that uses frozen feature extraction, with and without humidity fusion. We use a Leave-One-Subject-Out (LOSO) evaluation and show that the minority-class (hand washing) F1-score is a more informative metric than accuracy due to the class imbalance. We show that by adding a humidity channel to the IMU-based representations, we can improve the cross-subject generalization of our model.

In particular, detection of hand-washing events can be performed by different modalities such as cameras, microphones and wearable sensors. However, using cameras and microphones to collect such data is wrought with privacy concerns, while IMU-based approaches alone often suffer from motion ambiguity and intersubject variability [Straczkiewicz et al. 2019]. These wristbands also contain humidity sensors that register whether a person is near a source of water and can pick up on the flow of a tap during hand-washing. In addition to gathering data on fast arm and hand motions through its IMU, the system also uses humidity sensors to gather contextual data on the ambient environment. Humidity aids in the detection of hand washing, and overall performance is improved by combining the two.

Human Activity Recognition (HAR) from body-worn IMU has been approached with supervised deep models [Hammerla et al. 2016; Ordóñez and Roggen 2016], which rely on models trained on large labeled datasets and which can often perform poorly for diverse users. On the other hand, self-supervised representation learning leverages unlabeled data and produces embeddings that can transfer to new tasks. In particular, the TS2Vec framework proposed by [Yue et al. 2022] has achieved human activity recognition (HAR) using such learned representations. In using pretrained encoders as frozen feature extractors, the results demonstrate a clear trade-off.

Specifically, when labeled data is limited, lower training time is traded off with increased robustness. This paper explores the relative effectiveness of supervised vs. self-supervised methods for hand-washing detection on the LOSO condition. We additionally explore the effects on performance of including humidity data into IMU-based representations, and the impact of class imbalance on evaluation. Our supervised baselines used an LSTM classifier trained on the IMU magnitude and humidity slope features and a CNN-LSTM trained on these features. As our self-supervised baseline, we pretrained TS2Vec on segmented IMU windows and concatenated its embeddings with handcrafted humidity features to train a multilayer perceptron classifier (MLP). To evaluate the contribution of multimodal fusion methods, we ablate the TS2Vec model by removing humidity. Our main results are that (a) self-supervised TS2Vec with humidity fusion performs as well, or better, than the supervised baseline in a LOSO setting; (b) including humidity consistently improves detection ability in combination with IMU representations; (c) F1-score for the minority class is a better indicator than overall accuracy or other imbalance problems.

Author's Contact Information: Sonali Manoharan, Sonali.Manoharan@student.uni-siegen.de, Universität Siegen, Siegen, North Rhine-Westphalia, Germany.

ACM XXXX-XXXX/2026/2-ART
https://doi.org/XXXXXXX.XXXXXXX

2 Related Work

2.1 Hand Washing Detection from Wearables

Detecting handwashing is of increased interest only after COVID, given its importance for infection control and hygiene monitoring. Early studies relied on rule-based approaches, a threshold-based model for accelerometer, gyroscope, and similar signals. Later studies gathered scrub-motion data using inertial measurement units (IMUs) such as smartwatches and fitness trackers, and used traditional classifiers (e.g., Support Vector Machine, Random Forest, k-Nearest Neighbors) on hand-engineered time- and frequency-domain features to recognize these motions.

IMU-only studies are unrepresentative of free-living, real-world hand washing if done in controlled laboratory settings, or using WHO-style hand washing protocols [World Health Organization 2009]. And other studies have demonstrated high accuracy among small numbers or in subject-dependent evaluations. We conclude that a multimodal approach is better suited to detecting free-living hand washing events.

2.2 Sensors for Multimodal Activity Recognition

Human Activity Recognition (HAR) can be accomplished using a single modality (e.g., RGB(D) cameras, IMUs). Exploiting multiple modalities often improves performance since these approaches capture richer contextual information [Aguileta et al. 2019]. These setups can be complicated, but often IMUs are used together with RGB(D) cameras, audio input, environmental sensors (temperature, humidity, barometric pressure, light) and physiological sensors (heart rate, SpO_2, ECG).

For hand-washing detection, cameras and microphones both pose privacy concerns, especially in the restroom or other public spaces. At the same time, air sensors such as humidity sensors can protect user privacy. For example, when hands are being washed, humidity is particularly informative of proximity to water, and the speed of tap flow. [Burchard and Van Laerhoven 2024] engineered a wearable device, WearPuck, which is publicly available and records accelerometer data along with gyroscopic measures of humidity, temperature, and pressure. In experiments with 10 subjects who each performed hand-washing events, they found that humidity was actually quite responsive during this activity, but simple classifiers did not reliably improve with the addition of humidity features. This has led us to more closely examine the data collection mechanisms, the data preprocessing techniques, and the learning strategies in this domain. To our knowledge, previous work has not utilized humidity measurements or focused exclusively on hand washing detection with IMU data.

2.3 Supervised Deep Learning for Time Series

Supervised deep learning has made great strides in activity recognition on wearable platforms. Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs; [Hochreiter and Schmidhuber 1997]) have proven particularly powerful for time series classification. CNNs are often used to extract localized motion features from each video sequence, while LSTMs are utilized on top of the convolutional layers to represent the sequential structure of each video. Hybrid CNN-LSTM models [Hammerla et al. 2016; Ordóñez and Roggen 2016] require little manual feature engineering but, in exchange, require large labeled datasets and significant computing resources.

While supervised models can excel in subject-dependent settings, they struggle to generalize in subject-independent settings, such as LOSO, because they often overfit to specific motion styles. This is due to the fact that supervised models are constrained by labeled datasets and the distribution shifts over test subjects that cross-user variations can cause.

To solve this problem, we use TS2Vec [Yue et al. 2022], a self-supervised learning paradigm that learns contextualized time-series embeddings with hierarchical contrastive pretext tasks along multiple temporal resolutions. TS2Vec acquires robust representations without activity labels to leverage large unlabeled datasets to improve cross-subject generalization ability while reducing the requirement of labeled annotations. For wearable HAR frameworks in which a frozen (i.e., fixed) pretrained encoder produces embeddings as input to downstream classifiers, our results show that even when this setup reduces the computational costs of training deep neural networks from scratch within each cross-validation fold, semi-supervised pretrained models—such as HARNet [Yuan et al. 2024], which is trained on large accelerometer datasets—still tend to exhibit better subject-independent performance than models trained solely on the downstream dataset. This is consistent with findings across other domains with abundant data, such as computer vision and NLP, where pretrained models can exhibit strong in-context learning performance, albeit with lower computational costs.

2.4 Multimodal Fusion for Activity Recognition

Multimodal fusion is usually described by two main strategies: early fusion, where the features of multiple sensor modalities are concatenated before any modeling is performed, and late fusion, where the decisions made by the separately processed modalities are combined. [Kasnesis et al. 2018] observed late fusion outperforming early fusion on PAMAP2; however, the performance difference is often small. Similar performance across both fusion techniques suggest that the optimal methodology may depend on the dataset and modalities.

Our goal is to detect handwashing-related events (Table 3) using early fusion of IMUs to capture distinct motion patterns alongside humidity sensor to provide key contextual information regarding water presence. We implement early fusion in both supervised CNN–LSTM architectures and self-supervised TS2Vec pipelines by concatenating IMU-derived features/embeddings with equivalent humidity indices before classification. For TS2Vec, we also evaluate the pipeline without any input from the humidity sensor to evaluate the contribution of this specific modality to the overall results.

3 Dataset and Data Processing

3.1 WearPuck Dataset

The data used in this study was provided by the authors of the previous study [Burchard and Van Laerhoven 2024], which comes from their work on hand-washing detection. The results of the initial proof-of-concept consisting of 10 subjects showed that humidity had a distinct response during hand-washing, and that humidity did

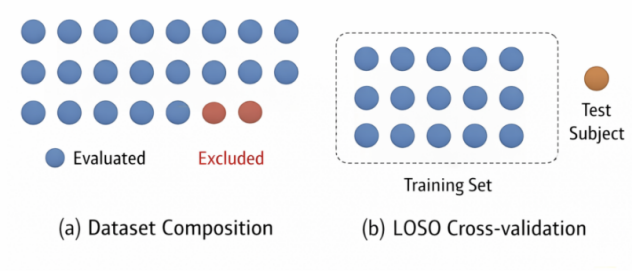

Fig. 1. Dataset composition and LOSO cross-validation. (a) 20 participants total; 18 used for evaluation, 2 excluded. (b) In each of 18 folds, one subject is held out as test; the remaining 17 form the training set.

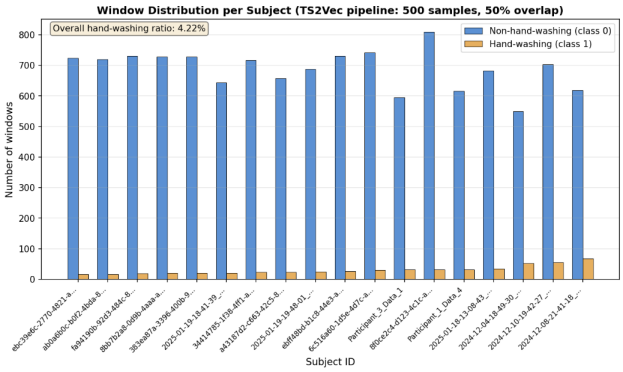

Fig. 2. Histogram of number of windows per subject for TS2Vec pipeline ($W_{\text{TS2Vec}} = 500$ samples, 50% overlap). Color annotations are self-descriptive, i.e., blue for no hand washing (class 0) and orange for hand washing (class 1). The overall hand washing rate is low, about 4.22 percent.

not improve the results of the initial machine-learning results. To increase model robustness and ecological validity, an additional set of 10 subjects was recorded using the same open-source wearable device, WearPuck, but at a later time point in the study. To facilitate comparison, the same recording setup, sensor placement, labeling approach and post-processing were applied as in the original publication.

Notation: We denote by $N = 18$ the number of subjects used for evaluation (two excluded), $S = 50$ Hz the IMU sampling rate, $W_{\text{TS2Vec}} = 500$ and $W_{\text{CNN}} = 150$ the window sizes (in samples), and $n_{\text{total}} \approx 12{,}927$ the total number of windows in the TS2Vec pipeline.

3.2 Dataset Organization

The amalgamated dataset comprises:

- `data/`—recordings from the ten original participants
- `new_data/`—recordings from the ten newly-recruited participants
- `labels.csv`—annotations for the original recordings
- `labels_new.csv`—annotations for new recordings

Despite the modest dataset size (20 subjects total, with $N = 18$ retained for analysis), we obtain comparable average classification performance to studies with larger cohorts. Two participants were excluded due to annotation inconsistencies or poor recording quality. Evaluation is performed using Leave-One-Subject-Out (LOSO) cross-validation, as illustrated in Figure 1.

3.3 Sensor Modalities

The WearPuck device used is a small form-factor wearable sensor capable of capturing motion, temperature and pressure signals. The Inertial Measurement Unit (IMU) is sampled at $S = 50$ Hz:

- `acc_x`, `acc_y`, `acc_z`
- `gyro_x`, `gyro_y`, `gyro_z`

Environmental sensors: `humid` (humidity), `temp` (temperature), `press` (pressure).

The IMU records signals at a constant rate of 50 Hz, unlike previous tests, which used variable sampling. Recording environmental sensors infrequently challenges analysts to interpolate the data, which can result in smoothing over small-scale variations. This information is logged in CSV files, with all the records for each sensor channel saved across time. The sampling rate $S = 50$ Hz means that the signal should contain frequencies lower than 25 Hz (the Nyquist limit).

Activity annotations are stored as separate JSON-formatted CSV files and indicate the `start` index, `end` index, and `timeserieslabels` (`"hw"`: hand wash, `"dry"`: dry, `"Null"`: background). Given that you need a binary classification scheme if you treat the task as distinguishing hand-washing from non-hand-washing, we labeled `Null` and `dry` as non-hand-washing (class 0) and `wet` and `hw` as hand-washing (class 1).

3.4 Dataset Expansion and Variability

The original dataset collected data from people doing a limited range of background activities—doing desk work, walking for about 10 seconds, going up stairs and washing their hands. In addition to the greatly expanded dataset, the newly acquired recordings focused on a broader range of natural movements: playing the guitar, bouncing a ball, playing video games, washing dishes, and other common manual actions. This also increased the inter-class similarity and the ecological validity of the tasks, making them more representative of real-world scenarios and therefore more applicable to the intended applications.

They then recruited 10 new subjects for the study, ages 20 to 30. Among those subjects, the average hand-washing event was about 10 seconds longer than in the original group—about 45 seconds.

3.5 Recording Duration and Class Distribution

All told, the dataset consisted of 1,240 minutes of recordings, of which only 48 minutes, or about 4 percent, was of hand-washing. After windowing, the TS2Vec pipeline yields approximately $n_{\text{total}} \approx 12{,}927$ windows: 4.22% of them (about 545) correspond to hand washing, and 95.78% (about 12,382) do not. We allow this imbalance in inputs to remain as is, though in training the model, we took explicit steps to address the imbalance. The distribution of windows across subjects is shown in Figure 2.

Property	Value
Subjects (evaluated)	18
Subjects (excluded)	2
Sensors	IMU (acc_x,, acc_y, acc_z, gyro_x, gyro_y, gyro_z), humidity, temp, pressure
Sampling rate	50 Hz
Window size (TS2Vec)	500 samples (10 s)
Window size (CNN-LSTM)	150 samples (3 s)
Step size (TS2Vec)	250
Step size (CNN-LSTM)	75
Total windows (TS2Vec)	~12,927
Hand-washing ratio	~4.22%

Fig. 3. Key characteristics of the dataset.

3.6 Data Preprocessing

The team executes a few additional preprocessing steps before training the model. The channel labeled humidity actually has one fixed garbage value, 79.1318359375, that first should be removed from the series. This value should be treated as NaN, causing it to drop out of the series; missing values need to be filled in through linear interpolation (estimating the values between adjacent valid samples via a linear function), with both forward and backward filling extending the signal at its boundaries as necessary.

Windowing: The time-series data are segmented using a sliding window approach. For the CNN-LSTM pipeline that processes windows of input data, we set the window length to $W_{\text{CNN}} = 150$ samples (~3 seconds at $S = 50$ Hz), with a step size of 75 samples (50% overlap between consecutive windows). The TS2Vec pipeline employs a window length of $W_{\text{TS2Vec}} = 500$ samples (~10 seconds) and advances with a step size of 250 samples (50% overlap). We apply a simple majority-vote rule to assign each window a binary label (positive or negative/neutral).

In particular, a StandardScaler, which standardizes the features to zero mean and unit variance, should be fit only on the training data to avoid the problem of data leakage. This ensures that subject-specific magnitude differences are removed, which helps the network learn the relative patterns, as well as provides numerical stability for your model. The scaler is fit only on the training subjects in each LOSO fold, and the test subject is transformed using that same scaler in order to ensure strictly subject-independent evaluation (and prevent data leakage).

3.7 Dataset Statistics and Imbalance Handling

Table 3 summarizes the key characteristics of the dataset. It is important to note that subjects had different numbers of windows and that hand-washing windows made up only 2 to 9 percent of each subject's data. Therefore, the class imbalance in terms of skewed percentages was maintained for each subject.

To tackle the severe class imbalance (approximately 4 percent positive class), our models use TS2Vec pipelines that augment samples with Gaussian noise when sufficient positive examples exist and employ SMOTETomek [Batista et al. 2004; Chawla et al. 2002] to create a more balanced training set; further, we apply balanced class weights to all models and use focal loss [Lin et al. 2017] to focus learning on difficult samples (we calibrated focal loss during this work to ensure that focal loss did not overfit to majority patterns). These techniques improve the recall, F1 score and the overall number of hand-washing events detected.

4 Methods

We compare supervised and self-supervised learning approaches for detecting hand-washing activity from wearable sensor data. We consider three modeling approaches: (1) a supervised CNN-LSTM model, (2) a self-supervised TS2Vec encoder followed by an MLP classifier, and (3) multimodal fusion of IMU and humidity features within both model families. All methods are evaluated using Leave-One-Subject-Out (LOSO) cross-validation.

4.1 Supervised CNN-LSTM Pipeline

With a supervised model that conditions on short temporal segments, this actually works. The input window size for the CNN is chosen as $W_{\text{CNN}} = 150$ samples (3 s at the given sampling rate of 50 Hz), with a step size of 75 samples (i.e., 50% overlap), and a minimum duration filter of 1 s is applied for the event-based post-processing. By decreasing the window length, the system increases temporal resolution.

This pipeline combines learned sequence representations with handcrafted features, and thus still relies on careful feature engineering. First, the raw IMU data are converted into magnitude sequences, which are fed to the convolutional layers. To normalize the motion measures, we compute the accelerometer and gyroscope magnitudes as follows:

$$\text{acc\_mag} = \sqrt{\text{acc\_x}^2 + \text{acc\_y}^2 + \text{acc\_z}^2}, \tag{1}$$

$$\text{gyro\_mag} = \sqrt{\text{gyro\_x}^2 + \text{gyro\_y}^2 + \text{gyro\_z}^2} \tag{2}$$

These were then fed, as time-series, into a Convolutional Neural Network (CNN). Next, a feature extraction module computes handcrafted magnitude features from the accelerometer and gyroscope signals and humidity slope features from the humidity signal. For each window, we take the mean, standard deviation, minimum and maximum of each magnitude, resulting in eight features, and the mean, standard deviation, maximum, minimum, and end−start of the humidity slope values, resulting in five features. These humidity features seemed to best capture the environmental conditions that the water-exposed backpacks were experiencing. Our feature vector, composed of only the most relevant handcrafted parts, is 13-dimensional.

The model is designed with two separate branches to keep the two parts separate. Instead of a simple stack of layers we introduce a more organized sequence branch, where we start with a stack of Conv1D layers (64 → 64 → 128 filters), followed by MaxPooling, BatchNormalization, and Dropout layers, and two LSTM layers (128 → 64 units). For the handcrafted feature branch, we take in the aforementioned 13-dimensional vector and pass it through Dense(64), BatchNorm, and Dropout layers. The outputs of the two branches are concatenated and run through Dense(128→64→sigmoid) layers.

In this way, we give our neural network a hybrid design that combines both learned temporal patterns and domain-specific statistical descriptors.

We train separately within each LOSO fold, using a combination of focal loss [Lin et al. 2017] ($\alpha = 0.25$, $\gamma = 2.0$) and Adam [Kingma and Ba 2015], with per-fold class weights. The training data is split 80/20 into training and validation, and we monitor loss on the validation set for early stopping. The decision threshold was set either to a fixed 0.5 level or by optimizing the F1 score on the validation set. This setup both corrects for class imbalance and also avoids overfitting. After inference, we apply a median filter (kernel size = 5) and remove short spurious detections by enforcing a minimum duration of 1 s.

4.2 Swapped-Labels Sanity Check

To ensure that our performance was not artificially inflated by class imbalance, we conduct an experiment in which we swap the labels so the positive class is the behaviors that do not correspond to hand washing with soap (Null or Dry), and the negative class is the hand washing classes. While this experiment uses the same architectural implementations and training processes, a model that meaningfully learns discriminative patterns should achieve F1 scores on the real task (with handwashing as the positive class) that are not artificially inflated by the predominance of majority-class events or labeling biases.

4.3 Self-Supervised TS2Vec Pipeline

The self-supervised approach [Yue et al. 2022], which decouples the representation learning from the classification task, can be used to learn useful features before learning to classify them. We use a window size of $W_{\text{TS2Vec}} = 500$ samples (10 seconds) and only the IMU channels `acc_x`, `acc_y` and `acc_z`, with a step size of 250 samples (50% overlap). Systems that use larger windows acquisition of richer contextual information, facilitating more effective representation learning.

A dilated Temporal Convolutional Network (TCN; [Bai et al. 2018]), TS2Vec uses hierarchical contrastive learning while being trained, self-supervised, on all the unlabeled IMU windows, to learn rich time-series representations. In its default configuration, the encoder generates 128-dimensional embeddings for each input. After pretraining the encoder, we fix all of its parameters, and use only the `encode()` function for each fold of leave-one-subject-out (LOSO) validation, without any retraining. This reduces computational cost and ensures that embeddings are consistent across folds.

In the multimodal version, the code calculates 11 other humidity features for each window: mean, std, maximum, minimum, median, count of values greater than 50, end-minus-start, range from 90th to 10th percentile, the mean and standard deviation of the first derivative and count of peaks. What we have done is to duplicate the humidity features and concatenate them to the TS2Vec embedding as the final feature vector, so the baseline serves only as a magnifier of the humidity's influence on the final prediction results. For class imbalance problems we use SMOTETomek [Batista et al. 2004; Chawla et al. 2002] and augment class one with Gaussian noise, with class weights. The multi-layer perceptron classifier (Dense 128 $\rightarrow$ 64 $\rightarrow$ Sigmoid) is trained with focal loss [Lin et al. 2017]. Earlier predictions are smoothed with a median filter (kernel size 3).

The TS2Vec IMU-only variant, excluding humidity features, is the ablation study that quantifies the contribution of the humidity modality. In all other respects, including the underlying architecture and training protocol, the system was left unchanged.

4.4 Evaluation Protocol

To evaluate the model, we use Leave-One-Subject-Out (LOSO) cross-validation with $N = 18$ folds. The evaluation split assigns 17 subjects to the training set and 1 subject to the test set, in each fold. The fitting of our model and the choosing of thresholds was done entirely on the training subjects' data; no data from the test subject was used during the training or threshold-choosing process.

For each subject, we report the F1 score and accuracy. Final results: the average F1 score and average accuracy are reported across all subjects. Due to the class imbalance in the dataset, the F1 score of the minority class (hand washing) remains the main metric for performance. To ensure a fair and unbiased comparison, no parameters are tuned on the test set: StandardScaler is fit only on the training data and, where applicable, early stopping is performed based on validation metrics alone. This approach avoids cross-fold leakage and provides good estimates of the performance on unseen subjects under the same data assumptions.

5 Results

We conducted several experiments with different model settings and report our main results in Table 8. To summarize overall performance, this table reports the mean F1 score and mean accuracy across subjects under Leave-One-Subject-Out (LOSO) cross-validation. Due to the strong class imbalance of the dataset with only about 4% of the data consisting of hand-washing, all models achieve similarly high accuracy scores (0.94–0.97). Thus, we treat F1 as our primary evaluation metric and report additional metrics in Table 8. The average F1 and accuracy across only the subjects that were part of the LOSO evaluation are shown; we report metrics only for those 18 subjects.

All models achieve high accuracy (~94–97%), due to the extreme class imbalance. Although accuracy is often reported, the F1 score is a more informative metric for minority-class detection. The TS2Vec + Humidity model performs best among the non-swapped configurations but can still be improved upon. The supervised CNN-LSTM performs the worst for minority-class F1 under fixed thresholding and improves from 0.47 to 0.55 when using validation-based threshold selection. We compare the models by presenting their results in Figure 4.

Per-Subject Performance Variation: When looking at the range of F1 scores over subjects (Figure 5) reveals considerable inter-subject variability. While some participants achieve F1 scores greater than 0.8, for others F1 is less than 0.3–0.4 in the supervised CNN-LSTM and TS2Vec IMU-only models. The differences typically have been attributed to various washing methods (amount of time and vigor), different amounts of exposure to humidity, different movement amplitude and different sensor positioning.

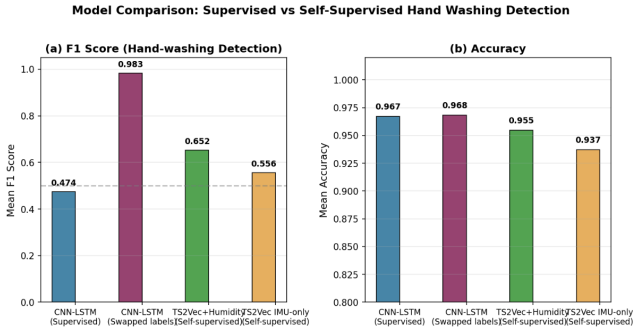

Fig. 4. Model comparison. (a) and (b) Summaries of hand-washing detection performance. Our trained TS2Vec + Humidity model performs better than the supervised baseline, and flipping the labels causes the model to obtain trivially high F1 scores by simply predicting the majority class.

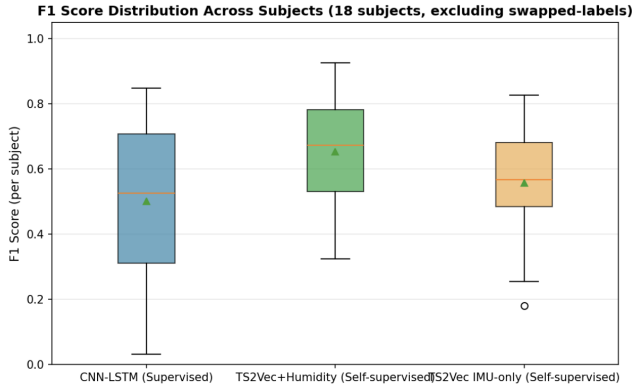

Fig. 5. Distribution of F1 scores for all 18 subjects, excluding others with switched labels. Boxplots for three methods: CNN-LSTM (supervised), TS2Vec+Humidity (self-supervised) and TS2Vec IMU-only (self-supervised).

Self-supervised models (TS2Vec + Humidity) also tend to achieve the highest median F1 score with the lowest variance compared with the supervised baseline. The fact that we see this across the profiles of participants gives credibility to the idea that the model is generalizing better to new users as well. Looking at the per-subject results, CNN-LSTM generally performs worse than TS2Vec + Humidity. Moreover, TS2Vec + Humidity outperforms the supervised model the most for participants for whom CNN-LSTM had the worst performance.

Effect of Humidity: To isolate the effect of humidity, we perform a direct comparison in the humidity ablation experiment, where we compare TS2Vec IMU only with TS2Vec + Humidity. Adding humidity produces an average mean F1 improvement of ∼0.096 (∼10 percentage points). Most subjects benefit from the addition of humidity, but low-F1 individuals benefit more, as humidity provides additional discriminative information for detecting water exposure. Accuracy remains steady at ∼96%. Multimodal fusion of IMU with others allows for better minority-class detection compared to a

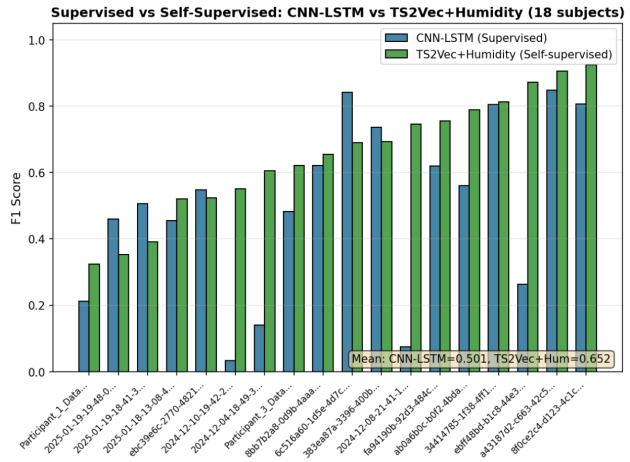

Fig. 6. Per-subject F1 scores: CNN-LSTM (supervised) and TS2Vec + Humidity (self-supervised).

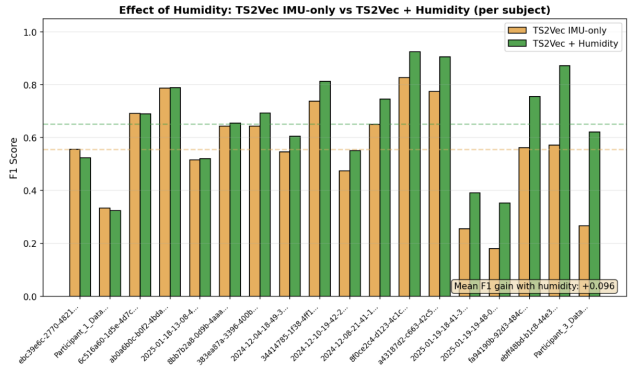

Fig. 7. Impact of humidity: comparison of TS2Vec IMU-only with TS2Vec + Humidity per subject (mean F1 gain with humidity: +0.096).

model that relies only on IMU (which more than adequately captures motion related patterns even on its own). Figure 7 shows the impact of humidity by comparing TS2Vec IMU-only with TS2Vec + Humidity per subject, where we obtain a mean F1 gain with humidity of +0.096.

Swapped-Labels Sanity Check: Even when the labels are switched so that the positive class corresponds to not handwashing—the majority class—the CNN-LSTM still achieves F1 of ≈0.98 and accuracy of ≈0.97, as the expected probability of the majority-class dominating is ≈0.96. This suggests that the high F1 scores in the label-switched case are due to trivial predictions based on majority-class labelling; the actual task of recognizing handwashing as a minority class is much harder, with the moderate F1 scores (0.47–0.65) seen there representing real task difficulty rather than deficiencies of evaluation protocol. All models have similarly high accuracy (0.94–0.97), but the F1 score for the minority class gives a more realistic

Summary of Model Performance (LOSO Cross-Validation)

Model	Approach	N	F1 (fixed)	F1 (val τ)	Acc (fixed)	Acc (val τ)
CNN-LSTM (Supervised)	Supervised	20	0.474	0.570	0.967	0.962
CNN-LSTM (Swapped labels)	Supervised	20	0.983	0.983	0.968	0.967
TS2Vec+Humidity (Self-supervised)	Self-supervised	18	0.652	0.652	0.955	0.955
TS2Vec IMU-only (Self-supervised)	Self-supervised	18	0.556	0.556	0.937	0.937

Fig. 8. Mean F1 score and mean accuracy across subjects under LOSO cross-validation.

measure of hand-washing detection performance. Table 8 focuses on this metric, treating F1 as the metric of primary concern.

6 Discussion

The results of machine learning analyses on the combined dataset show that humidity is an important modality for detecting handwashing. In contrast to prior studies using the WearPuck, this suggests that by leveraging the distinctly observable humidity patterns that occur in handwashing events, our approach that combines humidity and IMU data is able to provide contextual information that is distinctive and improves performance, particularly when combined with a representation-learning approach that is able to generalize across subjects. We further analyze whether we can learn informative vector embeddings of activity data in order to perform zero-shot learning from a small number of labeled samples. Using humidity alone does not solve this problem—if users have unique hand-washing behavior, then identifying them, even with humidity as an added loss, still results in large inter-subject F1 score variability.

The CNN-LSTM model uses labeled data to learn features, and it does so with short temporal windows (3 s). However, while the overall accuracy is high (~97%), the minority-class F1 is quite modest when using fixed thresholds (0.47) and only improves when using threshold selection based on the validation set (0.55). The problem described above suggests that supervised learning is extremely sensitive to class imbalance and may not generalize well when positive samples are rare and background activities are highly diverse.

Our model, based on TS2Vec, separates representation learning from classification. During training, TS2Vec first uses contrastive learning to pretrain on unlabeled IMU windows, allowing the model to learn generalized temporal representations of motion that are not label-dependent. The latter uses a lightweight MLP classifier with handcrafted humidity features to obtain a higher minority-class F1 score (0.65 with humidity), suggesting that incorporating these humidity features increases between-subject robustness. Indeed, we did not finetune the TS2Vec encoder and trained only the classification head. Instead, it may be possible that finetuning all layers, which has been shown to be beneficial in other domains [Yosinski et al. 2014], could achieve better performance.

An ablation study including humidity features shows that the addition of these variables greatly increases detection performance. We also found that appending 11 handcrafted humidity-related features (mean, standard deviation, max, min, slope, peak count, etc.) to the TS2Vec embeddings improved mean F1 scores by about 10 percentage points compared with models that relied solely on the IMUs. The improvement was seen across all subjects but was most pronounced among subjects who had the lowest baseline performance.

This makes sense intuitively—IMU data capture motion patterns associated with hand movements, while humidity is a signal of the contextual environment of being exposed to water during a handwashing event. Handwashing behaviors are not only repeated, but also involve sustained contact with a water source. For example, a model relying only on IMU data might conflate two similar hand movements, such as rubbing hands without water. Our dataset was inherently challenging due to high subject-to-subject variability, but in general, models with additional humidity modalities outperformed those using IMU alone. Therefore, we argue for the continued collection of humidity measurements in future work using similar datasets.

Our dataset is also highly imbalanced, containing hand-washing events in only about 4 percent of all observation windows. These imbalances explain why, although all our models were extremely accurate (about 94–97 percent), their F1 scores were more variable. Running a sanity check with inverted labels reveals that a naive majority class classifier can still achieve extraordinarily high F1 scores.

To address class imbalance, we used focal loss [Lin et al. 2017] ($\alpha = 0.25$; $\gamma = 2$) and class-weighting with SMOTETomek resampling [Batista et al. 2004; Chawla et al. 2002] for TS2Vec variations with additional humidity data, as well as minimum-duration filtering. We did not tune on the test set; threshold selections and early stopping were limited to the validation splits in each leave-one-subject-out fold. Methodological rigor holds evaluations leak-bias free. While mean F1 scores are the most significant metric, all of the measures we evaluated are informative across the tasks.

There was a wide range of F1 scores between different subjects, probably caused by the idiosyncratic nature of their hygienic routines. Most scores higher than 0.8 showed that some people were already doing well, while scores between about 0.3 and 0.4 suggested that other subjects might not have been washing every spot on their hands completely. What accounts for the difference in results could be the difference in the duration and intensity of the sequences, or the greater variation in movement amplitude or positioning accuracy that occurs when sensors are attached to extremities of the body in the scenarios. Or they could be related to the unique ways that our participants perform these gestures, even after correcting for ambient moisture.

Our results indicate that user calibration steps (i.e., retraining with small amounts of subject-specific data) may be helpful for real-world deployment of systems that detect whether someone is practicing the required routines. This likely increases the overall efficiency scores of users who have difficulty working under the generalized models used here.

Despite these results, we note several limitations to address in future iterations of the proposed work. (1) TS2Vec is only given the training subjects of each LOSO fold and therefore no potential leakage can occur. This was followed by finetuning full architectures with encoding frameworks but freezing everywhere except the classification heads. (2) While highly experimental, it may be

that adding more modalities—sound, say, in the form of microphone channels—could improve performance by picking up the sound of tap water, which tends to be audible in surrounding areas where these interactions typically occur. (3) Our first priority was developing efficient computer architectures for "online" classification to run on wearable devices during daily use. (4) We also aim to collect larger and more diverse datasets in the real world to increase the ecological validity of our results.

By designing a full wearable system incorporating IMUs, moisture sensors and microphones, reliable, environment-agnostic detection of routine hygienic practices can become feasible, as well as verification of proper execution.

7 Conclusion

In this study, we studied the problem of multimodal hand-washing detection using IMU and humidity data collected from 20 subjects. Compared to supervised approaches, self-supervised learning results in more reliable detection of hand washing. Previous work was unable to consistently exploit humidity patterns as our approach, which used information from both humidity and IMU recordings, improved performance dramatically in a self-supervised representation learning framework. In particular, we found that the TS2Vec model using just IMU data achieved a mean F1 score of 0.56, while the TS2Vec model including hand-crafted humidity features reached a mean F1 score of 0.65, exceeding both the supervised CNN-LSTM baseline (0.55), and the IMU-only TS2Vec variant (0.56). Additionally, we show that, given the large inter-subject variability in F1 scores, personalized retraining using small quantities of subject-specific data is likely to be more suitable for use in practice.

We presented an evaluation framework where we addressed the large class imbalance by using focal loss [Lin et al. 2017], SMOTE-Tomek resampling [Batista et al. 2004; Chawla et al. 2002] and a fair Leave-One-Subject-Out (LOSO) scheme, without any test set tuning. Furthermore, the swapped-label sanity check allowed us to confirm that if we only see moderate F1 scores on the real task, this is due to the difficulty of the task, and not due to an erroneous evaluation methodology. This indicates that with self-supervised representation learning combined with multimodal sensing, robust, subject-independent hand-washing detection is feasible. Third, we expect humidity sensors to be useful in hand-washing-related tasks as well as other water- and humidity-related activities in HAR (human activity recognition) applications; these use cases remain to be explored. Fourth, future work will involve integrating additional wearable sensors to produce a multimodal detection system.

References

A. A. Aguileta, R. F. Brena, O. Mayora, E. Molino-Minero-Re, and L. A. Trejo. 2019. Multi-sensor fusion for activity recognition—A survey. *Sensors* 19, 17 (2019), 3808. doi:10.3390/s19173808

S. Bai, J. Z. Kolter, and V. Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* (2018).

G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter* 6, 1 (2004), 20–29. doi:10.1145/1007730.1007735

R. Burchard and K. Van Laerhoven. 2024. Multi-modal atmospheric sensing to augment wearable IMU-based hand washing detection. In *Proceedings of the 9th International Workshop on Sensor-Based Activity Recognition and Artificial Intelligence (iWOAR 2024)*. Springer, 45–62. doi:10.1007/978-3-031-80856-2_4 arXiv:2410.03549.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357. doi:10.1613/jair.953

N. Y. Hammerla, S. Halloran, and T. Plötz. 2016. Deep, convolutional, and recurrent models for human activity recognition using wearables. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*. 1533–1540. https://www.ijcai.org/Proceedings/16/Papers/220.pdf

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780. doi:10.1162/neco.1997.9.8.1735

P. Kasnesis, C. Z. Patrikakis, and I. S. Venieris. 2018. Perception net: A multimodal deep neural network for machine perception. In *Proceedings of the Intelligent Systems Conference (IntelliSys)*. IEEE, 1–10. doi:10.1109/IntelliSys.2018.8628711

D. P. Kingma and J. Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2980–2988. https://arxiv.org/abs/1708.02002

F. J. Ordóñez and D. Roggen. 2016. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16, 1 (2016), 115. doi:10.3390/s16010115

M. Straczkiewicz, P. James, and J.-P. Onnela. 2019. On placement, location and orientation of wrist-worn tri-axial accelerometers during free-living measurements. *Sensors* 19, 9 (2019), 2095. doi:10.3390/s19092095

World Health Organization. 2009. WHO guidelines on hand hygiene in health care: A summary. https://www.who.int/publications/i/item/WHO-IER-PSP-2009.07

J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. 2014. How transferable are features in deep neural networks?. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 27. 3320–3328.

H. Yuan, S. Chan, A. P. Creagh, C. Tong, A. Acquah, D. A. Clifton, and A. Doherty. 2024. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *npj Digital Medicine* 7 (2024), 91. doi:10.1038/s41746-024-01062-3

Z. Yue, Y. Wang, J. Duan, T. Yang, C. Huang, Y. Tong, and B. Xu. 2022. TS2Vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 8980–8987. https://ojs.aaai.org/index.php/AAAI/article/view/20881

Received 27 February 2026