Kunwar, Utkarsh; Borar, Sheetal; Berghofer, Moritz; Kylmälä, Julia; Aslan, Ilhan; Leiva, Luis A.; Oulasvirta, Antti

Robust and Deployable Gesture Recognition for Smartwatches

Published in:
27th International Conference on Intelligent User Interfaces, IUI 2022

DOI:
10.1145/3490099.3511125

Published: 22/03/2022

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Kunwar, U., Borar, S., Berghofer, M., Kylmälä, J., Aslan, I., Leiva, L. A., & Oulasvirta, A. (2022). Robust and Deployable Gesture Recognition for Smartwatches. In *27th International Conference on Intelligent User Interfaces, IUI 2022* (pp. 277-291). (International Conference on Intelligent User Interfaces, Proceedings IUI). ACM. https://doi.org/10.1145/3490099.3511125

Robust and Deployable Gesture Recognition for Smartwatches

Utkarsh Kunwar
Aalto University
Espoo, Finland
utkarsh.kunwar@aalto.fi

Sheetal Borar
Aalto University
Espoo, Finland
sheetal.borar@aalto.fi

Moritz Berghofer
Huawei Technologies
Munich, Germany
moritz.berghofer@uni-a.de

Julia Kylmälä
Aalto University
Espoo, Finland
julia.kylmala@aalto.fi

Ilhan Aslan
Huawei Technologies
Munich, Germany
ilhan.aslan@huawei.com

Luis A. Leiva
University of Luxembourg
Esch-sur-Alzette, Luxembourg
luis.leiva@uni.lu

Antti Oulasvirta
Aalto University
Helsinki, Finland
antti.oularsvirta@aalto.fi

ABSTRACT

Gesture recognition on smartwatches is challenging not only due to resource constraints but also due to the dynamically changing conditions of users. It is currently an open problem how to engineer gesture recognisers that are robust and yet deployable on smartwatches. Recent research has found that common everyday events, such as a user removing and wearing their smartwatch again, can deteriorate recognition accuracy significantly. In this paper, we suggest that prior understanding of causes behind everyday variability and false positives should be exploited in the development of recognisers. To this end, first, we present a data collection method that aims at diversifying gesture data in a representative way, in which users are taken through experimental conditions that resemble known causes of variability (e.g., walking while gesturing) and are asked to produce deliberately varied, but realistic gestures. Secondly, we review known approaches in machine learning for recogniser design on constrained hardware. We propose convolution-based network variations for classifying raw sensor data, achieving greater than 98% accuracy reliably under both individual and situational variations where previous approaches have reported significant performance deterioration. This performance is achieved with a model that is two orders of magnitude less complex than previous state-of-the-art models. Our work suggests that deployable and robust recognition is feasible but requires systematic efforts in data collection and network design to address known causes of gesture variability.

IUI '22, March 22–25, 2022, Helsinki, Finland

ACM ISBN 978-1-4503-9144-3/22/03…$15.00
https://doi.org/10.1145/3490099.3511125

CCS CONCEPTS

• **Human-centered computing** → **Gestural input**; *Interaction design process and methods*; • **Hardware** → *Sensor devices and platforms.*

KEYWORDS

Gestures, Sensing, Wearables, Mobile Devices, Deep Learning

ACM Reference Format:
Utkarsh Kunwar, Sheetal Borar, Moritz Berghofer, Julia Kylmälä, Ilhan Aslan, Luis A. Leiva, and Antti Oulasvirta. 2022. Robust and Deployable Gesture Recognition for Smartwatches. In *27th International Conference on Intelligent User Interfaces (IUI '22), March 22–25, 2022, Helsinki, Finland.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3490099.3511125

1 INTRODUCTION

Owing to their superior performance in classification tasks, machine learning (ML) and in particular deep learning (DL) have gained popularity as a technique for gesture recognition on mobile and wearable devices. There is interest in expanding DL also to challenging consumer devices, such as smartwatches, with first demonstrators emerging for recognition of mid-air hand and finger gestures [32, 63]. However, a significant question about real-world deployment remains: how reliably do these techniques work when users, tasks, and contexts *change*; in other words, how robust are they? Moreover, for deployment, DL models should be able to run in energy- and computation-impoverished settings, which limits model sizes and architectural solutions.

This paper looks at the problem of developing robust and deployable deep learning methods for gesture recognition. The first key challenge we address is robustness: As a data-driven technique, accuracy is highest with DL within the envelope defined by the training dataset. Noise, out-of-distribution (OOD) samples, and distributional shift pose serious issues to recognition. In a recent practical demonstration of this, Laput and Harrison [31] reported a decrease from 95% accuracy to 88% when testing a trained model with a user who was not in the training data (leave-user-out), and from 95% to 75% when users changed the watch strap from tight to loose fit. A question stands out how to improve the robustness of

DL-based methods for gesture recognition in real-world conditions. A robust algorithm would be able to support users of all walks of life, young and old, tall and short, across the conditions they want to use their devices in. The success of gestural interaction arguably lies in its flexibility and how gestures can be used to interact with digital content in varying and mobile contexts.

A second key challenge we address is deployability: Modern DL models are known to be too large for deployment on resource-constrained devices. For example, the acclaimed GPT-3 model [11] has 175 billion parameters and a file size of 350GB after some optimizations such as 16bit float precision. This is obviously not possible to deploy on commodity hardware. Smartwatches are not known for their computational prowess compared to the smartphones they are wirelessly tethered with. They are made to be energy efficient and are desirable by the average consumer if they tend to last longer on a single charge. A typical user has less use for running heavy computation tasks and would rather find a long-lasting battery life more appealing. Thus, for deployment purposes, it is necessary to consider the size of the classifier because having a large model run continuously on the smartwatch itself with copious compute operations will have a significant impact on its longevity. In this context, it is important to aim for model architectures that are not only accurate but also efficient and with a minimal footprint (i.e., small file size and a small number of FLOPs).

DL has changed the requirements for the development of robust recognition algorithms. In the past, robustness came with the price of implementing time and space agnostic gesture techniques, such as dynamic time warping algorithms trialled for contextual mid-air gesture recognition [4] and extending gesture elicitation techniques to identify contextual variations in gestures [3]. DL-based recognition techniques have transformed the gesture recognition landscape with deep recognisers seemingly able to achieve very high performance as long as the right kinds of datasets are available for training. Besides collecting better datasets, the ML community has turned to look at learning techniques that better generalize in the context of noise, OODs, and distributional shift [1, 41, 61].

We propose that prior domain knowledge on *causes* behind everyday variability and false positives should be exploited in both data collection and the design of the recogniser itself. Along these lines, we make two contributions towards training a more robust gesture recogniser. First, we present a variant of the obstacle course-based methodology that aims at collecting more representative yet diverse training datasets. Here, participants are taken through simulated conditions, such as walking, that are representative of everyday situations which are hypothesised to be challenging for recognition. The novel aspect we investigate is deliberate diversification of gesturing, with the intuition of encouraging a classifier's robustness to distribution shifts. In practice, participants are instructed to create realistic but varying levels of alterations to the gestures. Second, we review known approaches to exploit domain knowledge in recogniser design with the aim of deployability. We propose variations of convolution-based networks for classifying raw and processed sensor data which perform reliably under both individual and situational variations. We then present an implementation for the recognition of gestures on a smartwatch trained on our collected dataset. The recogniser achieves upwards from 98% accuracy on gesture recognition across conditions where previous approaches have reported failing [31]. This work highlights the need for Human-Computer Interaction (HCI) research to contribute to the training of better recognition algorithms for real-world use on commodity hardware.

2 RELATED WORK

ML is increasingly being used in everyday sensing tasks, including gesture recognition, natural language processing, computer vision, or information retrieval. Therefore, it has become important that the algorithms we develop for ML are robust to different working environments. In this section, we relate to previous work on robustness in gesture recognition, and more concretely on smartphones and smartwatches. For a general overview of robust machine learning we recommend reviews by Xu et al. [65] and Shafique et al. [53].

2.1 Robust Gesture Recognition

Robustness is synonymous with *algorithmic stability*, a property that characterises how effective an ML classifier behaves on unseen data. For example, learning in the presence of outliers [2] or adversarial examples [25]. In the domain of ML, with a wide variety of architectures and training recipes, there are many cards one can play to address and attempt to solve the issue of robustness. Adversarial robustness is an area of active research due to the pervasive use of deep models in detection and authentication [34]. Choosing between classification robustness and accuracy has been considered a trade-off given how classifiers tend to evolve decision boundaries around clusters of data points during training [42, 59, 67]. Feature redundancy and the inability of neural networks to learn high-level features have also been implicated for a decreased robustness in distribution shifts and OOD generalisation [40, 54]. Works addressing and providing solutions to these issues vary significantly – from suggesting data augmentations [22], and change in architecture [68] to change in training strategies and loss functions. For example, Zhang et al. [67] formulate a loss function that comprises the natural classification error along with a boundary smoothing term to encourage decision boundaries which allow for slight OOD samples. Generalised loss functions have also been shown to improve robustness in vision tasks for clustering and classification [7, 48]. Work towards multimodality has also demonstrated an improvement in robustness via the principle of robust overdesign through highly correlated inputs from different sources [6].

The issues of robustness in classifiers may be addressed on domain-specific applications. In the area of gesture recognition, Miao et al. [35] proposed a hand gesture recogniser that uses classic computer-vision features (Hu invariant moments and HOG features) and a sparse representation. They found that such a sparse representation improved the accuracy and robustness of gesture recognition. Schak [50] presented an analysis of the robustness of deep Long Short-Term Memory (LSTM) networks for freehand gesture recognition against temporal shifts. They concluded that including artificial gesture onset variability in the training data leads to high robustness against various tested effects.

Recent works use acoustic signals to track hand movement and recognise gestures [62], although they require dedicated, specialised hardware that might not be available in everyday conditions. For

example, Vimos et al. [60] proposed to correct the hand orientation using surface electromyography, which resulted in a boost in recognition accuracy of more than 30% over the baseline case.

2.2 Gesture Recognition on Mobile Consumer Devices

Mobile phones find increasing use in everyday tasks like navigation, shopping, and communication [5]. Proliferation in embedded electronics is enabling new consumer markets and applications for mobile technology. Smartwatches have gained attention as wearable fitness and activity trackers, including special solutions for wheelchair users [18]. There seems to be a general trend towards consumers interacting within ecosystems of devices and smart things. Thus, one of the main challenges at present is to address the increasing complexity and diversity of user interfaces to aid a consumer's seamless access to services reliably [69].

Previous research in implicit interaction techniques [9, 27, 47] and intelligent user interfaces [15, 51, 64] allows to expand into new forms of contextual interactions, including gestural interaction, and customisation and automation of services tailored to customer behaviours, preferences, and contexts [12, 43, 45].

Deep networks are fueling many recognition techniques including hand gesture recognition, which has been an active field since the introduction of smartphones. Techniques for statistical modelling [36], 3D hand recognition [13] and vision-based gesture recognition [46] have been novel milestones. Until recently, popular deep learning frameworks like Tensorflow and PyTorch were not available for mobile phones and researchers had to, for example, boot Linux-based OS on smartphones to explore deep mobile recognisers and perform on-device transfer learning on the resource-scarce device [52]. However, such an OS and solution swiftly deplete the battery. While the general field of gesture recognition is vast with applications in gaming, automotive, sign language, and so on. Inertial Measurement Units (IMUs) have been used for deep hand gesture recognition on smartwatches, although it is a rather new field with Laput and Harrison's work [31]. Previous gesture and activity recognisers on smartwatches relied on shallow classifiers, such as Bayesian networks [16] with significantly less accurate results for fine-grained gestures and activities.

In light of recent ML/DL and deployment of models on resource-constrained devices, Branco et al. [10] highlight challenges, such as memory footprint, execution time, power consumption, and scalability. Ultimately, there is a trade-off between performance and resource deprivation that we need to be especially aware of when dealing with resource-constrained devices. Direct design of smaller DL models for deployment on low-end hardware has been demonstrated to achieve similar results [26] which could be reduced further by model compression techniques. Laput et al. [32] have shown that one can achieve superior results with really high sensor sampling rates of 4 kHz. However, access to higher sampling rates impacts the number of data points that need to be stored in the buffer, computed, and potentially transferred. Moreover, a higher sampling rate directly corresponds to increased power consumption [19, 58]. This consumption spans computations in pre-processing steps to, for example, generate spectrograms for applying Convolutional Neural Networks for recognition. Laput and Harrison [31] report a drop of ≈15% when the sensor signals are downsampled from 4 kHz to 200 Hz.

3 DATA COLLECTION

A key goal in our data collection was to obtain diverse but representative gestures in conditions that resemble those real-world conditions that might pose challenges to smartwatch-based gesture recognition. With this goal in mind, we extended the familiar obstacle course-based methodology [30]. First, we identified conditions for the obstacle course that are known to be challenging, such as mobility (walking) and tightening/loosening the watch strap [31]. Second, when going through the obstacle course, participants were asked to produce diverse but realistic variants of gestures under a simulated payment scenario (see below). Third, we ensured that our sample of participants has people with different heights, ages, gender, handedness, and previous exposure to smartwatches. Finally, building on previous work, we also collected several everyday gestures that could produce false positives [31].

3.1 Participants

Twenty-four participants took part in the data collection study (7 female, 3 left-handed) with a mean age of 26.4 years and a mean height of 174.6 cm. Six of them had previously or currently owned a smartwatch, three of them for more than one year, one of them for less than a month, and the rest had experience from one month to a year. The sample was stratified using the following variables:

- Age (2 groups): young adults (18-30 years), middle-aged adults(31-45 years)
- Gender (2 groups): female, male (self-identified)
- Height (3 groups per tertiles): less than 179 cm, 179 cm or more (males); less than 166 cm, 166 cm or more (females)

The stratified sample consisted of 2 x 2 x 3 = 12 cells, with 2 participants in each cell.

Due to COVID-19 restrictions, only university employees and students were recruited through internal mailing lists. Each participant was compensated with 30€ of taxable rewards. All participants had normal or corrected-to-normal vision and hearing, with no known cognitive impairments or regular medication.

3.2 Experimental Design

The experiment was arranged into four blocks. The first three blocks had two levels of watch tightness, and two levels of instruction for the payment scenario. The payment scenario further had three levels for body posture. The last block consisted of the "false positive gestures" and was the same for all participants. The levels are described below.

- Watch tightness: The wrist band tightness levels were "tight", in which the participant was instructed to put the watch on so that it was tight but comfortable, and "loose", in which the participant was first instructed to tighten the watch as in the "tight" condition, and then loosen the wrist band by two notches. So, if the tertile was for a tight strap then it would be left as is, and if it was for the loose strap, it would be loosened by two strap notches starting from the initial setup. This helped ensure that the conditions cover the different

ways users may wear watches, however without imposing them to wear in an unrealistic or contrived way.
- Instructions for Payment: the instruction levels were either "instructed", in which the participant received both verbal and visual instructions on how to perform a payment gesture and was asked to follow the instructions precisely, or "free", in which the participant only received the verbal instructions and was asked to interpret them in a way that came naturally to them in order to execute the payment.
- Gesture: the payment gestures were either shaking or tilting. The intended motion of these gestures is shown in Figure 1
- Body posture: the payment gestures were performed first in a sitting position, then in a standing position, and finally while walking in a circle.

The gesture conditions in the first three blocks, as well as the first two gestures of the everyday block, had 10 repetitions of each gesture. The remaining everyday conditions had three repetitions of each gesture/activity. This diverse set of everyday hand gestures and activities forms a strong basis for a rejection class to enable a recogniser to reliably classify negative samples. Table 1 shows a detailed description of the everyday superclass with a schematic of the structural hierarchy in the Supplementary Material.

3.3 Apparatus

We rely on the already available and widely used Android Sensor API for the Wear OS which restricts the sampling rate of the IMU sensor data to a maximum of 100 Hz for the Ticwatch Pro 3 smartwatch. Measurements from the smartwatch's accelerometer, gravity sensor, and gyroscope were recorded. Each recorded with approximately 100 to 102 Hz sampling rate. The sampling rate fluctuated, possibly due to the delay in I/O operations while recording the data. Each sensor recorded data on three channels corresponding to the three axes in Cartesian coordinates in the device frame of reference. Additionally, a video was captured with the phone's front camera to help synchronise the sensor data during annotation.

3.4 Procedure

COVID-19 Precautions. The participant was brought into the test room while the study conductor remained in the observation room. Henceforth, the participant received instructions audibly via an online conference call, and visually by observing the study conductor through the window between the two rooms. The study conductor controlled the recording phone with `scrcpy` [17], which gives the conductor the ability to control an Android device through Android Debug Bridge. The laboratory setup for the gesture data collection is shown in Figure 2.

The first 11 participants wore the watch on their dominant hand, while the rest wore it on the hand they would normally wear a watch on ("preferred hand"). All the participants in the preferred hand condition wore the watch on their non-dominant hand, though this was by chance and not part of the procedure.

Instructions for Gesturing. The participants received visual instructions for the payment gestures and some of the everyday gestures by watching the study conductor perform the gesture. For other gestures, the participant received verbal instructions only. In the free condition, the participants only received verbal instructions and were prompted to interpret them as they "*would feel the most natural*".

In the "everyday" condition, the participants were instructed to execute the gestures only three times each, varying the gesture every repetition, and every repetition lasting for at least three seconds.

3.5 Preprocessing and Annotation

After the data was collected, we produced Comma-Separated Values (CSV) files from the processed sensors to be importable in the NOVA (Non-Verbal Annotator) software [8, 21] for human-supervised labeling. The processing script also calculated the sample rates for each sensor file, to help align the files properly in NOVA. Since the recorded rates were not constant, the script calculates an average based on the number of measurements and the duration of the recording.

In this annotation phase, the video and three sensor readings were utilised to make an annotation track for each of the data files. These were processed as described above and loaded into NOVA; the annotators then labeled each gesture in the data by visual cues. The gravity sensor was used as the main indicator for the beginning and the ending of the payment gestures. These gestures were perceived to begin when the gravity sensor displayed a rise, indicating the arm had been lifted, and the label was placed approximately to start from the "rising" slope of the gravity sensor's waveform and end on the "lowering" slope of the same waveform. This is illustrated in Figure 3. This point was chosen as the starting point because many of the participants started their gestures early, even though instructed to first lift their arm and only then begin the gesture. Because of ambiguity in human movement, it was decided safer to start the label preferably a little before the gesture starts rather than miss the start and begin the label in the middle of the gesture. In some of the "everyday" gestures, the video was utilised more, since in them a rising arm didn't necessarily signal the start of a gesture of interest.

The video recording was used as a reference but could not be used as a reliable source of gesture duration, since it did not entirely match up with the sensor data, because of the aforementioned delays. The regions of no significant activity or "rest" labels were added post-annotation by automation. The gesture labels also included a "secondary gesture" to denote any meaningful gestures the participants performed during the recording but weren't supposed to (e.g. scratching themselves or adjusting the watch).

One participant was excluded during the recording stage due to data corruption mid-experiment; this participant was replaced to bring the number of participants back to 24. Additionally, the first two participants did not have data for "everyday" gestures since the condition was added to the experiment design after these two had already been measured. In total, three unique conditions were not recorded completely or were corrupted and are therefore missing from the data set.

4 METHODOLOGY

With the primary goal of deploying a robust recogniser on a smartwatch, we try to design a DL model architecture that runs on the raw sensor data. Considering the meagre computational resources

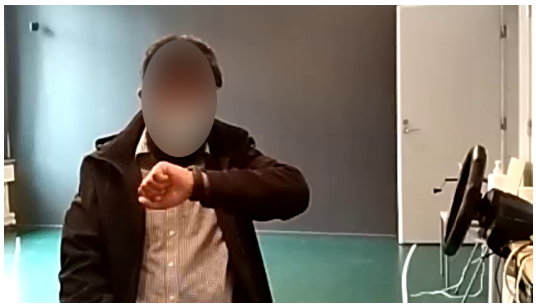
(a) Shaking gesture

(b) Tilting gesture

Figure 1: The two gesture motions to be used as triggers for the intent of payment.

Table 1: Gesture and activity classes recorded under the "everyday" superclass. The participants performed three repetitions of each gesture, varying the mannerism in each repetition. The exception were the cup conditions, where the participant did ten similar repetitions.

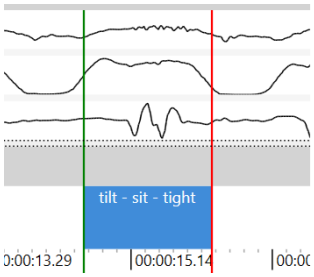
Gesture/activity code	Description
cup - shake	Performing the shaking gesture while holding a cup.
cup - tilt	Performing the tilting gesture while holding a cup.
wave	Waving.
jog	Pretending to jog on place.
comb	Combing or pretending to comb his or her hair.
phone	Using the phone.
bottle	Opening or closing the cork of a bottle.
drink	Pretending to drink from a cup.
knife	Cutting food with a knife and a fork.
fork	Pretending to eat with a fork.
spoon	Pretending to eat with a spoon.
burger	Pretending to eat a burger.
dust	Cleaning the table with a paper towel.
dishes	Pretending to wash a plate.
washer	Putting utensils into an imaginary dishwasher.
door	Opening and closing a door.



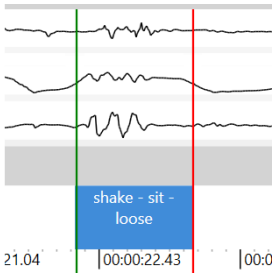
(a) Participant performing the gesture while sitting.

(b) Another participant performing the gesture while walking.

Figure 2: Frames from the video captured by the smartphone during data collection in the laboratory setup.



(a) An annotation sample for tilt gesture in the sitting posture with tight strap.



(b) An annotation sample for shake gesture in the sitting posture with loose strap.

Figure 3: Two examples showing how the gravity sensor (middle signal) is used as a guide for placing the annotations. The other two waveforms (accelerometer (top) and gyroscope (bottom)) were used as reference, when the gravity sensor's readings were unclear.

available to us on a smartwatch, we try to minimise both the pre-processing steps as well as the operations in the classifier itself. We do acknowledge the state-of-the-art results achieved by using high-resolution spectrograms by Laput and Harrison (L&H) [31]. So, we devise a similar architecture but with a smaller footprint to compare against the classifier that works on the raw sensor data and the L&H baseline.

A simpler model, with a smaller number of trainable parameters, is less prone to overfitting. This translates to smoother, less tight decision boundaries which has also been the motivation for addressing adversarial robustness [67]. Further, architectural decisions on the incorporation of certain layers have a significant impact on controlling the number of model parameters. With these motivations in mind, we prepare a pipeline and test-bench for the training of situational- and individual-invariant robust recognisers.

4.1 Data preparation

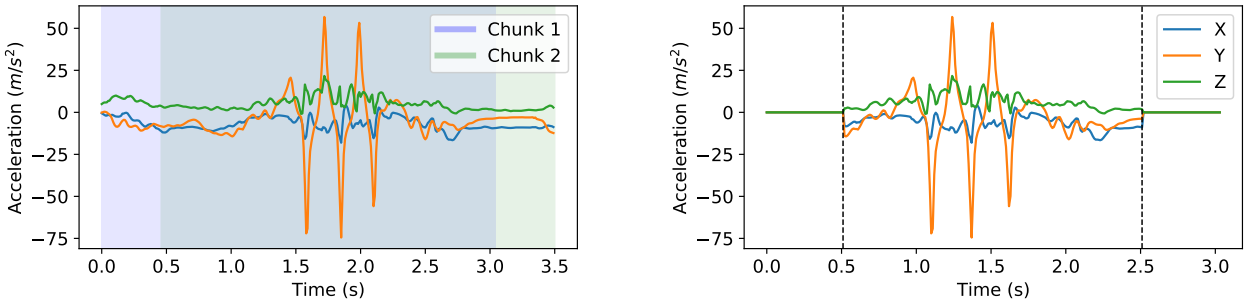
4.1.1 Time-series. For classification, approximately three seconds were considered appropriate as the maximum duration during which a gesture takes, which has also been done previously for activity recognition by L&H [31]. We consider it sufficiently long to be able to reliably capture most gestures generated with intentional hand movements. We use 304 samples at 100 Hz which represents ≈ 3.04s worth of IMU sensor data. This arbitrary length of 304 samples is chosen to be able to generate spectrograms of specific dimensions in Section 4.1.2. The raw sensor data was provided as three CSV files for each sensor along with separate files for the labelling scheme and annotations (units in seconds). For preparing the dataset for use in machine learning, the label files were parsed to obtain the row indices of the samples in the CSVs. Labelled sequences shorter than the length of 304 samples (≈ 3.04s) are zero-padded on both ends; the longer sequences are split into chunks of sample lengths of 304 according to the overlapping strategy shown in Figure 4. In summary, the number of chunks is decided based on the minimum number of 304-sized chunks that a long signal can contain when allowing for overlaps. These chunks are then fed to the network as inputs for training and inference.

4.1.2 Spectrograms. Spectrograms help in visualising the evolution of frequencies in a waveform over time. This is achieved by applying a Short-Time Fourier Transform (STFT) with a sliding window function over the waveform of interest. Spectrograms have found extensive use in the field of audio signal processing. They can be a useful tool to capture the periodic nature of human activities [37, 66].

We use the 3.04 s long time-series data sequences prepared in Section 4.1.1 to generate spectrograms. A Hanning window function was used for the STFT along with an FFT length of 64, yielding a total of 33 frequency bins for a resolvable frequency range of 0-50 Hz from a signal sampled at 100 Hz, which gives a frequency resolution of ≈ 1.5 Hz. The STFT was calculated without padding to minimise edge artefacts in the resulting spectrogram [38]. A sliding window overlap of 75% was chosen to have sufficient resolution along the temporal axis to minimise data redundancy and unnecessary additional compute STFT operations [58]. A 75% overlap is recommended for periodic window functions, like the Hanning window function, which ensures that spectral aliasing is minimised and also encourages robust signal reconstruction if desired [56, 57]. This yields 16 rolling frames for a 304 sample long signal which creates a spectrogram of dimensions 33×16. The highest frequency bin for the range of 48.5–50 Hz is discarded to conform the dimensions to powers of 2 for easier handling of max-pooling operations in the deep learning network. The spectrogram matrix consists mostly of smaller values with sparse high values in regions of interest making the pixel distribution of the spectrogram heavily skewed and similar to that of a power function. So, the computed spectrogram is scaled from the amplitude scale to the logarithmic (dB) scale with the aim of Gaussianising the pixel distribution within a single spectrogram to enable faster convergence of the network during training.

4.2 Neural network design

Recurrent networks find common use in applications of machine learning on sequential data of arbitrary lengths. Mobile SDKs provide native implementations on the smartwatch for common layers and operations, like vanilla Recurrent Neural Network (RNN), CNNs, pooling, normalisations, and activations like Rectified Linear Unit (ReLU) and TanH. However, vanilla RNNs tend to suffer from the problem of vanishing gradients over long sequences of data [23]. This issue is somewhat resolvable by using LSTMs [24] or



(a) Strategy for creating chunks from a sample longer that 3.04 s.

(b) Symmetric zero-padding for samples shorter than 3.04 s.

Figure 4: Padding and splitting strategies for fixing the input size for use in CNNs.

Gated-Recurrent Units (GRU) [14] with the additional overhead of maintaining extra internal states, but these units were not available for deployment on the smartwatch we were using. Recurrent network operations also tend to be less parallelisable because of their sequential nature, making them computationally slower for training. CNNs offer a solution around these problems with the caveat of having a predetermined fixed-sized input. This enables us to use CNNs for gestures, which tend to be of short durations, by fixing a threshold duration (in our case, approximately 3 seconds). This also has the added benefit that CNN-based classifiers are easily reusable for transfer learning as feature extractors for other domains or subsequent use with recurrent networks with the reduced latent representations.

We consider the binary classification task of whether a triggered gesture denotes payment or not. We design and propose variations of simple CNN-based neural network architectures with the aim of deployment in ready-to-use application-oriented scenarios on the low-end hardware of smartwatches. To keep the preprocessing overhead to an absolute minimum, we use the "raw" sensor data from the accelerometer, gyroscope, and gravity sensors as they come from the smartwatch API.

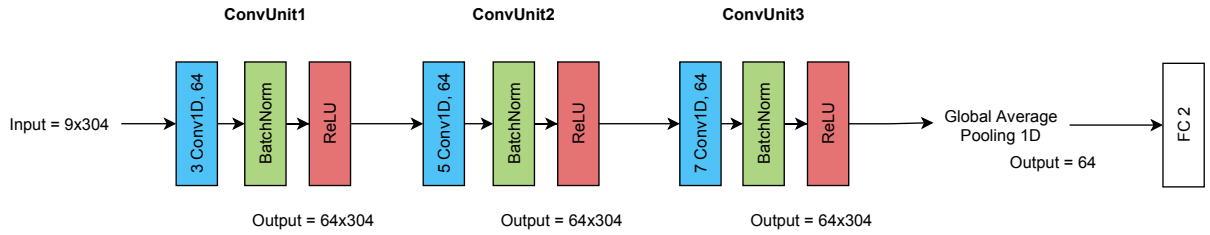
4.2.1 Network inputs and CNN features. Our base model's priority is to work on the raw sensor data for recognition. For the model input, we use 304 time-series samples representing $\approx$ 3.04s of IMU data from each of three sensors (accelerometer, gyroscope, and gravity sensor) having 3-axis measurements. These contribute 3 channels (X, Y, Z) per sensor, resulting in a total of 9 channelled input data sequences making the input shape 9×304. For the time-series model, this is passed through 3 Convolutional Units (ConvUnits) where each ConvUnit is defined as a block consisting of a 1D padded convolutional operation, a 1D batch normalisation, followed by the ReLU activation function. Normalisation techniques like batch normalisation, and variants like layer normalisation, and group normalisation are known to improve the robustness of training under varying hyperparameters. Batch normalisation achieves this by smoothing the optimisation landscape to achieve a more stable training [49]. The design for the ConvUnit is similar to that of a conventional convolutional block of the ResNet [20] but extends its 2D operations for our 1D sensor signals in the case of time-series input. The proposed architecture variation utilising the time-series data is shown in Figure 5a.

We extend this network architecture to 2-dimensions for spectrogram inputs, in order to compare our approach to L&H's, which represents the current state-of-the-art. The 9-channel 304 sample long time-series is converted channel-wise into spectrograms. Each axis of the three sensors results in one spectrogram yielding a net input shape of $9 \times 32 \times 16$. To reduce dimensionality across the 2D ConvUnits, we incorporate 2×2 maxpooling but the overall recipe remains the same as shown in Figure 5b. We further expand this architecture by building and training end-to-end a late fusion network with inputs as both the time-series and spectrogram modalities. The motivation of using both modalities is that the resulting network would be able to benefit from the information in both the time and frequency domains as they are correlated [6]. This network, shown in Figure 5c consists of a time-series head and a spectrogram head which are the CNN feature extractors from the previously discussed architectures.

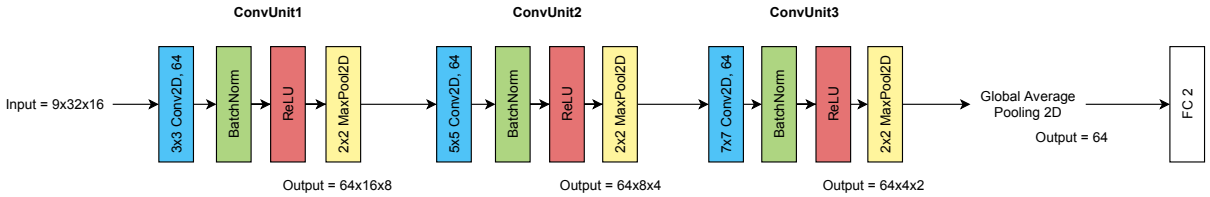
4.2.2 CNN feature classification. Each ConvUnit has 64 kernels for feature extraction and the size of kernels increases from 3 to 5 to 7 with each successive ConvUnit to view the signal at different receptive fields. The output of the final ConvUnit is then passed to a Global Average Pooling (GAP) layer which collapses information along the temporal dimension. This is done with the idea of forcing the ConvUnits to not work just as feature extractors but to output feature scores which help the final classification task in the subsequent layers [33]. GAP also reduces the number of learnable parameters in the network helping to reduce the model size substantially. This minimises the risk of overfitting when compared to the conventional VGG16-like approach of flattening the convolutional features as a vector and passing them to fully-connected layers directly [55]. We also get around the need for using dropout for regularisation because the fully-connected layers are small enough to avoid overfitting on our low-dimensional dataset. Finally, the pooled tensor is passed through a fully-connected layer which gives the class scores.

4.3 Training

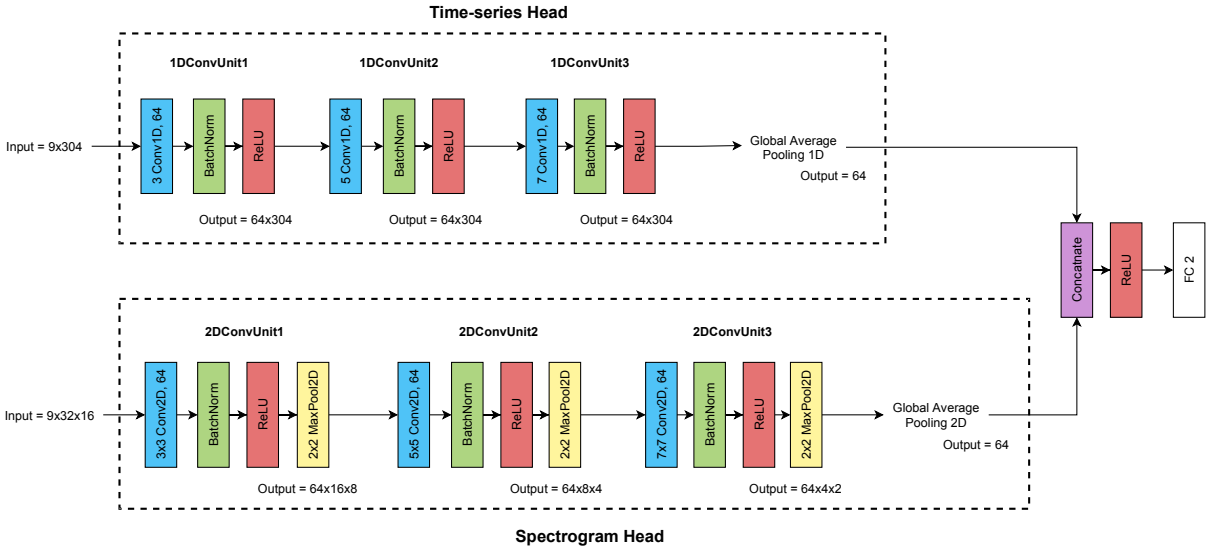
The network is trained for 60 epochs on batches of size 128 of 3.04s samples with the Adam optimiser using an initial learning rate of



(a) 1D convolution based model architecture for time-series inputs.



(b) 2D convolution based model architecture for spectrogram inputs.



(c) Modality fusion model for both time-series and spectrogram inputs.

Figure 5: Proposed network architectures for use with IMU data – both for raw time-series and for preprocessed spectrograms.

0.05 and an exponential learning rate scheduler ($\gamma = 0.9$). A random weighted sampler is used to remove bias against minority sample classes. We use the general categorical cross-entropy loss and two logits for the binary classification task instead of the conventional binary cross-entropy loss (BCELoss) with a single logit to be able to compare the effects of using different loss functions designed for multiclass classification.

4.3.1 Loss Functions. Different loss functions provide a different optimisation landscape affecting their convergence and sensitivity to activations from noisy input [7]. Using loss functions well suited to our dataset distribution can potentially help improve robustness through either ignoring noisy samples, label noise, improving regularisation, or optimising an objective function that is closer to the goal of the training task. With this intuition for improving the convergence during training for obtaining more robust decision boundaries, we compare the effect of different loss functions on the robustness of the aforementioned 1D convolutional network. We use the categorical cross-entropy loss for our models as the default loss function in all experiments unless stated otherwise.

For the loss functions presented below, y denotes the true one-hot encoded label, $\hat{y}$ is the true label in $+1/-1$ encoding, o is the output of the last layer of the network. j denotes the j^{th} dimension of a given vector, and $\sigma(.)$ denotes the probability estimate.

- Robust Adaptive Loss [7]: Uses a trainable parameter α to learn a loss function that best suits the data and the learning task, so that it does not get impacted by outliers or noisy samples. Equation 1 gives the optimisation function to be minimised.

$$\sum_j \frac{|\alpha-2|}{\alpha}\left(\left[\frac{\left(\frac{\|y^j-\sigma(o^j)\|_1}{c}\right)^2}{|\alpha-2|}+1\right]^{\frac{\alpha}{2}}-1\right) \tag{1}$$

- Expectation Losses: Expectation losses are calculated between the one-hot encoding of the label and the prediction probabilities [28]. ℓ_1 expectation loss in Equation 2 minimises the ℓ_1-norm of the misclassification probability while the ℓ_2 expectation loss, given in Equation 3, minimises the ℓ_2-norm. This is in contrast to the cross-entropy loss which aims to maximise the probability of correct labelling.

$$\sum_j \|y-\sigma(o)\|_1 \tag{2}$$

$$\sum_j \|y-\sigma(o)\|_2^2 \tag{3}$$

- Higher-Order Hinge Losses: Squared and Cube Hinge Loss described by Equation 4 and 5 respectively have been shown to report faster convergence and better performance than cross-entropy loss [28]. Higher-order hinge losses tend to penalise misclassified examples more severely. Margin-based losses have also previously outperformed other loss families in terms of generalisation ability. This is attributed to the implicit regularisation in margin losses.

$$\sum_j \max\left(0, \frac{1}{2}-\hat{y}^j\cdot\sigma(o^j)\right)^2 \tag{4}$$

$$\sum_j \max\left(0, \frac{1}{2}-\hat{y}^j\cdot\sigma(o^j)\right)^3 \tag{5}$$

4.4 Evaluating Robustness

We briefly compared several interpretations of robustness in literature in terms of boundary smoothness, optimisation landscape smoothness, and robustness to adversarial inputs. However, at a more application-oriented level, robustness in user experience can be defined and perceived as an application or service's reliability of reproducing the same result under different individuals and conditions. Our dataset divides the classes further into per-user basis, as well as under varying situations. This enables the ability to partition the data into user-specific and condition-specific divisions for applying a modified k-fold cross-validation resulting in leave-one-user-out (LOUO) and leave-one-condition-out (LOCO) cross-validations [44].

For each fold, a model is trained on $k-1$ folds and the remaining fold is used for cross-validation. The payment gesture recognition setup has three situations for strap tightness and three for body posture for 24 participants. These cross-validations help to gather model behaviours extensively for the entire combination of users and situations. The metrics for all the folds are logged and the mean metric defines the average model behaviour while the standard deviation of the metrics justifies the model's robustness to how an average model performs for an unseen user or situation not in the training set. This method of evaluation helps to narrow down the particular situations or individuals for which the models perform poorly.

5 RESULTS

We compare the proposed model variations with an extensive testbench reflective of real-life scenarios when the model is subject to deployment. We consider the binary classification task of tilt and shake gestures for payment authentication. The positive class for payment consists of both the tilt and shake gestures, while the negative class contains the everyday gestures and activities from Table 1 along with the intermediate activities of no significance (labelled "rest"). This results in a total of 4688 positive inputs and 10045 negative inputs for the 24 participants with each input having a sample length of 304. Since the classes are imbalanced, a weighted random sampler is used. All networks were trained on a Dell Precision 5820 Ubuntu 20.04 workstation with Intel Xeon W-2133 (12-core 1.2-3.9GHz) and an NVIDIA TU106 (GeForce RTX 2070).

Accuracy has been used as a standard metric for almost all classification tasks. But for imbalanced datasets or multiclass problems, accuracy can be a misleading metric. To address this, we report metrics — precision, recall, the macro-F1 score, and the balanced accuracy which take into account these known issues. We consider both the mean of the metrics and their standard deviations to be important indicators during our evaluation. We prioritise a high mean metric value but if two approaches happen to be tied for the same place then we consider the one with the lower standard deviation to be the better approach.

For a state-of-the-art baseline classifier, we use the model proposed by L&H [31] for comparison. The L&H model takes a 3-channel spectrogram as input of dimensions 256×48 from the accelerometer sensor. As mentioned in the previous section, we change the final layer to output for 2 classes instead of their 25 class for our gesture recognition task.

The L&H dataset is not directly usable on our proposed architectures. This is because the L&H dataset is provided in the form of 3-channel spectrograms without the original raw time-series data. Naive downsampling of their spectrograms for use with our model would not be representative of real-world testing since spectrograms have both temporal and frequency axes which cannot be scaled independently. Access to the raw time-series data would have enabled us to generate spectrograms with our parameters (Section 4.1.2). However, we can infer the parameters they used in their study and try to simulate their procedure.

We simulate the L&H spectrogram generation process by upsampling our time-series sequences from 100 Hz to 4 kHz. We perform a cubic interpolation of our time-series signal and evaluate the signal at an effective sampling rate of 4 kHz. We then recreate their steps for spectrogram generation with an FFT window size of 4096, a 2.998 s sample interpolated at 4 kHz, and a hop size of 168, yielding a 2049×48 sized spectrogram. Only the bottom 256 bins are kept representing frequencies from 0–128 Hz at a resolution of

0.5 Hz. We did not have access to the high sampling rate of 4 kHz for our data collection. Note that due to the property of the Nyquist sampling rate, it is fundamentally impossible to resample a 100 Hz signal to 4 kHz by the Fourier method [39]. However, we have tried to circumvent this problem with interpolation. The spectrograms obtained with this method were visually consistent with those in the original study as can be seen in Figure 6.

5.1 User-independent experiments

Following the conventional approach for training a deep learning classifier for unpartitioned data, we pool together all participant (or user) data and collapse all levels of situations (strap tightness, body posture) under their respective superclasses of payment and rejection. The resulting singular dataset contains samples from all users and all situations. We randomly shuffle the dataset and split it into 75% training, 12.5% validation, and 12.5% testing. This experiment aims to demonstrate the model's capability to classify new samples from a user who is already present in the training data. Table 2 shows the results of the experiment for the different model architectures.

5.2 Leave-one-user-out experiments

A classifier, after deployment, seldom is retrained on fresh data. For a classification task involving participants, it is not always possible to train on a sample size representative of the target population. Hence, ensuring robustness across users becomes paramount to mitigating variability that may arise due to individual differences in performing gestures. To demonstrate how a model (pretrained on a set of users) performs when used by a new unseen user (not present in the training data), we perform the leave-one-user-out cross-validation. For each cross-validation fold, data from one user is held out for testing and the remaining participants are used for training the classifier. The modality fusion network performs the best on all fronts for an unseen user as shown in Table 3 with the extra information it gets from both the temporal and frequency domains. The effect of the other loss functions can be seen clearly in this experiment, with ℓ_2 expectation loss and the robust adaptive loss functions improving the baseline CCELoss, with higher metric scores across most users. Robust adaptive loss learns the optimisation surface to minimise the effect of outliers, which would explain the reduction in variance compared to that of the CCELoss.

5.3 Situational robustness experiments

We design an experiment similar to the leave-one-user-out where the test dataset for the left-out user is split into multiple constituent situations. For the payment gesture, those situations are for watch tightness (tight, loose, free), and body stature (sit, stand, walk). The pre-trained network from the leave-one-user-out experiment is then evaluated against these split datasets and the results are totalled across all user folds for these six situations. Table 4 shows the modality fusion model is in the lead again with higher mean metrics and smaller standard deviations in most situations. Our spectrogram model ranks next at similar but marginally better performance to the baseline L&H. Note that only the CCELoss trained time-series-based model is shown in Table 4 because CCELoss performed the best for situational robustness tests out of all the other loss functions.

5.4 Leave-one-condition-out experiments

Finally, we study the scenario of having a missing situation in the training set. To demonstrate the robustness of the models against unseen conditions, we design a leave-group-out experiment for the different situations in the dataset. We refer to this as the leave-condition-out experiment. For the payment scenario, we have six situations but they are not fully independent. The first level of the variation in the payment dataset is the watch strap tightness while the second level is the body posture. However, variation in watch tightness and body posture occurs together. So, the experiment is split into two parts for a total of six leave-condition-out folds as follows.

5.4.1 Watch tightness. The dataset has the body posture situations collapsed and merged giving a three-situation dataset of watch strap tight, loose, and free. A leave-group-out cross-validation is performed on these three situations, e.g., train on tight and loose and validate on free. This part yields three folds for the experiment.

5.4.2 Body posture. Similarly, the watch strap tightness situations are collapsed which gives a dataset for sitting, standing, and walking. This, again, yields three leave-group-out folds for the experiment. As shown in Table 5, the modality fusion model outperforms all the other approaches.

Augmentation strategies and other loss functions tested on the following experiments did not offer significant improvements over the already mentioned results so they have been left out in the final evaluation.

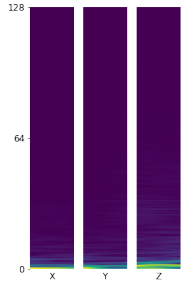
5.5 Model comparison

In Table 6, we compare the model sizes, their footprint, and their number of parameters to gauge the deployability of these architectures on the low-end hardware of wearable devices. It can be seen that our proposed models for the specified modalities not only are highly accurate, as demonstrated in the previous experiments, but also have minimal computational impact, which is desirable for deployment on smartwatches and resource-constrained devices.

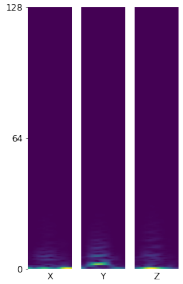
6 DISCUSSION

We have shown that robust and deployable DL models for gesture sensing on low-resource devices are within reach, assuming an appropriate approach towards data collection and neural architecture design. In particular, we suggest a variation of an obstacle-course-like environment for in-lab data collection to encourage diversified data to train robust classifiers and to be able to quantify robustness using suitable stratification methods. We suggest variations of common neural network architectures for gesture recognition that lead to highly efficient models.

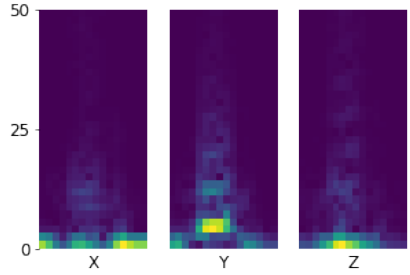
Our model achieves an upwards of 98% across challenging recognition tasks, owing to both the data collection method as well as the network design, each contributing towards its robustness to situational and individual variations. Our collected dataset spans a multitude of conditions that occur in our everyday lives, guaranteeing a rich dataset distribution. On the other hand, the low



(a) Activity spectrogram from the L&H dataset [31].



(b) Spectrogram from 4 kHz interpolated gesture.



(c) 32×16 spectrogram from 100 Hz signal for our 2DConv-based network architecture.

Figure 6: Accelerometer spectrograms for the L&H baselines and for our 2DConv-based architecture. (b) and (c) were generated from the same gesture signal. The artificially generated spectrograms using interpolation to 4 kHz in (b) have a visually similar looking frequency distributions as in the original 4 kHz spectrograms from the L&H dataset. By using the 100 Hz signal without interpolation, only smaller and relatively low-resolution spectrograms can be generated as shown in (c) but they provide sufficient frequency and temporal features.

Table 2: Binary classification performance on the test set for payment gesture detection on the user-independent experiments (averaged over 5 independent training runs).

Model Architecture	Accuracy	Balanced Accuracy	Macro-F1	Precision	Recall
Baseline L&H ($n_{\text{class}} = 2$)	0.994 ± 0.002	0.993 ± 0.002	0.993 ± 0.002	0.992 ± 0.003	0.993 ± 0.001
Time-series (1DConv, CCELoss)	0.992 ± 0.001	0.991 ± 0.001	0.990 ± 0.001	0.991 ± 0.001	0.990 ± 0.002
Time-series (1DConv, ℓ_1 Expectation)	0.987 ± 0.003	0.984 ± 0.003	0.984 ± 0.003	0.983 ± 0.004	0.985 ± 0.002
Time-series (1DConv, ℓ_2 Expectation)	0.990 ± 0.003	0.988 ± 0.004	0.977 ± 0.002	0.989 ± 0.002	0.999 ± 0.006
Time-series (1DConv, Squared Hinge)	0.991 ± 0.002	0.990 ± 0.002	0.990 ± 0.002	0.990 ± 0.002	0.989 ± 0.002
Time-series (1DConv, Cube Hinge)	0.991 ± 0.002	0.990 ± 0.002	0.988 ± 0.004	0.990 ± 0.001	0.990 ± 0.003
Time-series (1DConv, Robust Adaptive)	0.978 ± 0.007	0.974 ± 0.008	0.974 ± 0.012	0.978 ± 0.006	0.974 ± 0.009
Spectrogram (2DConv)	0.992 ± 0.001	0.991 ± 0.001	0.991 ± 0.001	0.990 ± 0.002	0.992 ± 0.001
Modality fusion (Time-series + Spectrogram)	$\mathbf{0.995 \pm 0.001}$	$\mathbf{0.994 \pm 0.001}$	$\mathbf{0.994 \pm 0.001}$	$\mathbf{0.994 \pm 0.001}$	$\mathbf{0.995 \pm 0.001}$

Table 3: Classification performance on unseen users for payment gesture detection in the leave-user-out experiments ($n_{\text{fold}} = 24$).

Model Architecture	Accuracy	Balanced Accuracy	Macro-F1	Precision	Recall
Baseline L&H ($n_{\text{class}} = 2$)	0.979 ± 0.016	0.976 ± 0.018	0.975 ± 0.019	0.972 ± 0.024	0.980 ± 0.014
Time-series (1DConv, CCELoss)	0.976 ± 0.018	0.974 ± 0.018	0.973 ± 0.019	0.973 ± 0.021	0.974 ± 0.019
Time-series (1DConv, ℓ_1 Expectation)	0.974 ± 0.019	0.969 ± 0.021	0.969 ± 0.023	0.965 ± 0.023	0.975 ± 0.016
Time-series (1DConv, ℓ_2 Expectation)	0.981 ± 0.014	0.978 ± 0.015	0.977 ± 0.015	0.976 ± 0.018	0.979 ± 0.013
Time-series (1DConv, Squared Hinge)	0.975 ± 0.022	0.972 ± 0.023	0.971 ± 0.024	0.970 ± 0.024	0.974 ± 0.023
Time-series (1DConv, Cube Hinge)	0.975 ± 0.019	0.971 ± 0.022	0.971 ± 0.023	0.971 ± 0.021	0.971 ± 0.025
Time-series (1DConv, Robust Adaptive)	0.940 ± 0.045	0.934 ± 0.048	0.927 ± 0.058	0.927 ± 0.063	0.941 ± 0.042
Spectrogram (2DConv)	0.976 ± 0.022	0.973 ± 0.023	0.972 ± 0.025	0.969 ± 0.024	0.977 ± 0.024
Modality fusion (Time-series + Spectrogram)	$\mathbf{0.982 \pm 0.016}$	$\mathbf{0.979 \pm 0.018}$	$\mathbf{0.979 \pm 0.019}$	$\mathbf{0.977 \pm 0.024}$	$\mathbf{0.982 \pm 0.016}$

complexity of the DL model we propose ensures that its generalisability is not harmed due to the common problem of overfitting, which is particularly pressing on smaller and low-dimensional datasets. Further, we exhibit the impact of supplementing the input with correlated features, thereby increasing the performance of the classifier.

The proposed model is shown to be almost two orders of magnitude less complex than a previous state-of-the-art approach while ensuring no loss in classification accuracy. This opens the door for this model to be readily deployable without the need for further modifications or fine-tuning such as model pruning and compression. The small size and minimal preprocessing requirements of

Table 4: Comparison of the classification performance of the models for specific situations for an unseen user. LOUO evaluation on conditions taken individually for situational robustness experiments ($n_{\text{fold}} = 24$).

Model Architecture	Metric	Tight strap	Loose strap	Free condition	Sitting	Standing	Walking
Baseline L&H ($n_{\text{class}} = 2$)	Accuracy	0.989 ± 0.013	0.985 ± 0.015	0.973 ± 0.028	0.974 ± 0.028	0.980 ± 0.021	0.986 ± 0.013
	Balanced Acc.	0.982 ± 0.019	0.976 ± 0.022	0.956 ± 0.038	0.965 ± 0.031	0.972 ± 0.023	0.980 ± 0.018
	Macro-F1	0.981 ± 0.021	0.976 ± 0.023	0.953 ± 0.045	0.963 ± 0.034	0.971 ± 0.025	0.979 ± 0.019
Time-series (1DConv, CCELoss)	Accuracy	0.973 ± 0.019	0.980 ± 0.022	**0.976 ± 0.021**	0.980 ± 0.013	**0.987 ± 0.011**	0.985 ± 0.011
	Balanced Acc.	0.957 ± 0.028	0.969 ± 0.034	**0.967 ± 0.021**	0.971 ± 0.019	0.980 ± 0.016	0.978 ± 0.016
	Macro-F1	0.955 ± 0.032	0.968 ± 0.036	**0.965 ± 0.023**	0.971 ± 0.020	0.979 ± 0.017	0.978 ± 0.017
Spectrogram (2DConv)	Accuracy	0.990 ± 0.013	0.985 ± 0.018	0.972 ± 0.029	0.978 ± 0.022	0.982 ± 0.017	0.987 ± 0.015
	Balanced Acc.	0.984 ± 0.020	0.976 ± 0.027	0.956 ± 0.042	0.970 ± 0.026	0.975 ± 0.022	0.980 ± 0.021
	Macro-F1	0.983 ± 0.022	0.975 ± 0.029	0.951 ± 0.050	0.968 ± 0.029	0.974 ± 0.024	0.979 ± 0.023
Modality fusion (Time -series + Spectrogram)	Accuracy	**0.994 ± 0.006**	**0.989 ± 0.009**	0.976 ± 0.033	**0.981 ± 0.025**	0.987 ± 0.023	**0.991 ± 0.007**
	Balanced Acc.	**0.990 ± 0.010**	**0.982 ± 0.014**	0.963 ± 0.042	**0.974 ± 0.026**	**0.982 ± 0.022**	**0.987 ± 0.011**
	Macro-F1	**0.990 ± 0.011**	**0.982 ± 0.014**	0.958 ± 0.053	**0.973 ± 0.029**	**0.981 ± 0.025**	**0.986 ± 0.011**

Table 5: All-user classification performance on unseen situations in the leave-condition-out experiments ($n_{\text{fold}} = 6$).

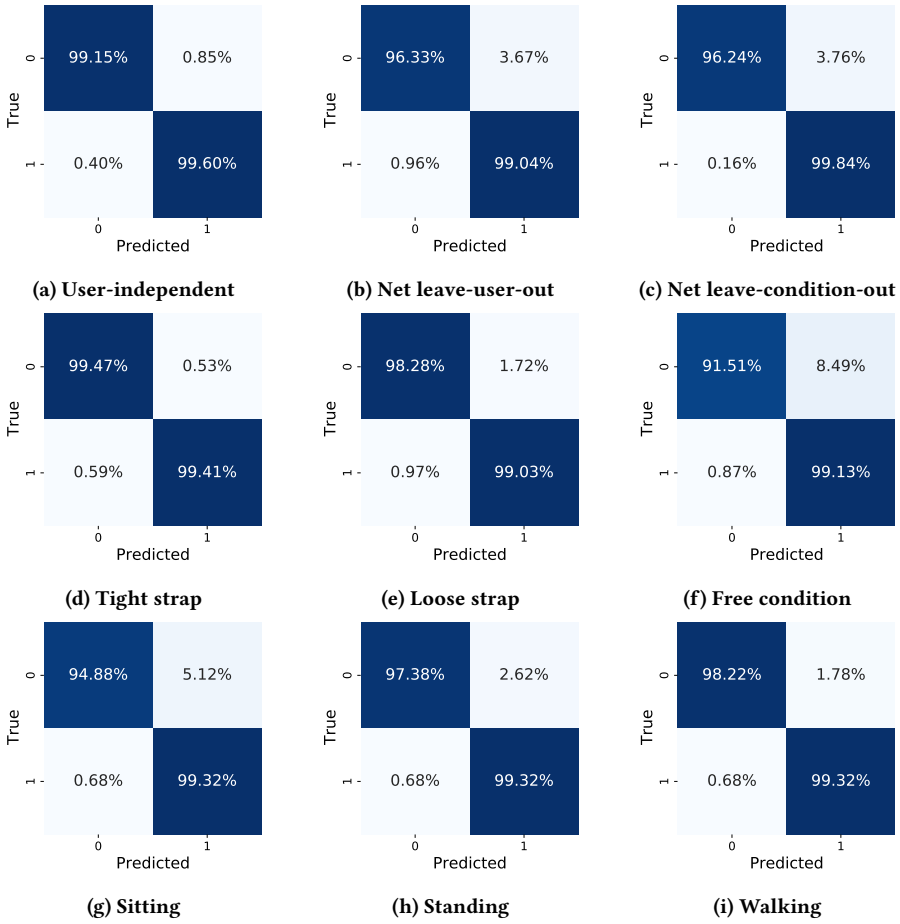
Model Architecture	Accuracy	Balanced Accuracy	Macro-F1	Precision	Recall
Baseline L&H ($n_{\text{class}} = 2$)	0.991 ± 0.011	0.985 ± 0.019	0.984 ± 0.020	0.976 ± 0.030	0.994 ± 0.008
Time-series (1DConv, CCELoss)	0.989 ± 0.010	0.982 ± 0.017	0.981 ± 0.018	0.976 ± 0.029	0.987 ± 0.007
Time-series (1DConv, ℓ_1 Expectation)	0.985 ± 0.014	0.975 ± 0.024	0.974 ± 0.026	0.969 ± 0.040	0.981 ± 0.008
Time-series (1DConv, ℓ_2 Expectation)	0.991 ± 0.010	0.985 ± 0.017	0.984 ± 0.019	0.979 ± 0.029	0.991 ± 0.006
Time-series (1DConv, Squared Hinge)	0.988 ± 0.008	0.981 ± 0.013	0.981 ± 0.014	0.974 ± 0.021	0.988 ± 0.007
Time-series (1DConv, Cube Hinge)	0.987 ± 0.009	0.979 ± 0.015	0.978 ± 0.016	0.974 ± 0.024	0.983 ± 0.024
Time-series (1DConv, Robust Adaptive)	0.956 ± 0.027	0.938 ± 0.033	0.934 ± 0.037	0.946 ± 0.040	0.931 ± 0.044
Spectrogram (2DConv)	0.990 ± 0.014	0.983 ± 0.024	0.982 ± 0.027	0.975 ± 0.039	0.991 ± 0.010
Modality fusion (Time-series + Spectrogram)	**0.992 ± 0.011**	**0.987 ± 0.018**	**0.986 ± 0.020**	**0.981 ± 0.030**	**0.992 ± 0.030**

Table 6: Comparison of the properties of network architectures and their inputs.

Model Architecture	Sensor sample rate (Hz)	Preprocessing MFLOPs [29]	Input dimensions	Number of parameters	Size (MB)	Training time (s)	Inference MFLOPs
Baseline L&H ($n_{\text{class}} = 2$)	4000	12.38	[3, 256, 8]	23,667,062	91	825	1895.28
Time-series (1DConv)	**100**	–	**[9, 304]**	**51,394**	**0.2**	**130**	**31.16**
Spectrogram (2DConv)	**100**	0.02	[9, 32, 16]	308,802	1.2	178	44.54
Modality fusion (Time-series + Spectrogram)	**100**	0.02	[9, 304], [9, 32, 16]	360,194	1.2	197	59.88

our model together offer huge savings in terms of computational resources, but more importantly in battery life for use in wearable devices. The models make use of the OS-exposed APIs for the sensor data which makes it developer-friendly, while the previous state-of-the-art required low-level kernel modifications to obtain access to a higher sampling rate for classification resulting in more power consumption. This is a positive step as it allows the application developers the ability to integrate these models in their applications without the worry of having to do device-specific modifications. The significant size improvement also enables the recogniser to be able to run entirely "offline" on, for example, smartwatches eliminating the need for distributed recognition on both smartphones and smartwatches.

Our work, along the lines of applied research, provides a minimum viable solution for gesture detection. We see several exciting opportunities to improve and build upon our research. First, we tackled the task of binary classification, so the next step would be to focus on extending it to the multiclass problem. The ability of our method to be robust and resource-efficient prospects its use as a one-vs-all (OVA) classifier unit which can then be easily extended for multiclass classification. Second, the use of multiple onboard sensors could potentially improve robustness further. Our modality-fusion approach can benefit from multi-modal input, by utilising multiple sensing sources. To this end, we have performed the actual deployment on the wearable itself, avoiding previous workarounds that relied on a paired smartphone to perform the



(a) User-independent (b) Net leave-user-out (c) Net leave-condition-out

(d) Tight strap (e) Loose strap (f) Free condition

(g) Sitting (h) Standing (i) Walking

Figure 7: Confusion matrices for the binary classification experiments for the modality fusion network architecture. If there are multiple folds in the experiment, then the matrices are averaged across all folds to generate a "net" confusion matrix.

sensing computations. Another interesting aspect to address individual robustness could be the on-device personalisation of the classifier for the intended owner of the device. Lastly, exploring the feasibility of transfer learning with our CNN-based architectures could substantially increase the reach of the approach in other domains such as instrumentation and controls.

7 CONCLUSION

This paper has made two contributions toward the engineering of gesture recognisers for challenging conditions. First, we show the advantage of data collection with intentional variability to train more robust recognisers. Second, we used this dataset to train CNN-based architectures which perform at par with the state-of-the-art models on our dataset. We further demonstrate that a small enough architecture can accurately recognise the target gestures under individual variations as well as situational variations and reliably reject false positives. We believe that for practical deployment, the pros of having a small model outweighs that of a larger network by being more power-efficient as well as space-efficient, making it ideal for deployment in wearable devices. We also explored the possible effects of the choice of loss functions on the model training process as a first step towards its impact on classification robustness.

Our future work aims towards the direction of using the time-series model as an OVA unit for extending our approach to multi-class classification, the effects of data augmentations of the different modalities on recognition robustness, and artificial data synthesis to compensate for limited participant's training data.

OPEN SCIENCE

The gesture dataset for the 24 participants, along with additional data for steering activity is available at the URL `https://userinterfaces.aalto.fi/robustgestures`. The dataset can be either used directly with the raw CSV files or loaded from the preprocessed Pandas dataframes included for convenience.

ACKNOWLEDGMENTS

This work was supported by the Department of Communications and Networking – Aalto University, Finnish Center for Artificial Intelligence (FCAI) and the Academy of Finland projects Human

Automata (Project ID: 328813), BAD (Project ID: 318559), Huawei Technologies, and the Horizon 2020 FET program of the European Union (grant CHIST-ERA-20-BCI-001).

SOCIETAL IMPACT

Improvements in on-device sensing have extended the uses of mobile devices, enabling people to better stay connected, enhancing safety, and helping them navigate. Mobile sensing has also been a driver of digital transformation at large, for example supporting more active life via services like health monitoring and features such as fall detection. At the same time, more robust discriminative ability, like the one presented here, could be used for biometric identification; that is, for determining the identity of the person performing the gestures. This may have a societal impact that is either positive or negative. On the one hand, the system can ensure that it is the actual owner of the device the one performing the gestures, thereby a positive impact. On the other hand, the collected gesture data could be used to profile the user, thereby a negative impact. Finally, methods that can significantly decrease the model size, like one presented here, should not be neglected, since even small improvements will contribute to sustainability especially when we have to consider that the number of devices that are in use and need replacement regularly is very large and increasing.

REFERENCES

[1] Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H. Falk, and Ioannis Mitliagkas. 2020. Generalizing to Unseen Domains via Distribution Matching. http://arxiv.org/abs/1911.00804. arXiv:1911.00804 [cs, stat]

[2] Frank J. Anscombe. 1960. Rejection of Outliers. *Technometrics : a journal of statistics for the physical, chemical, and engineering sciences* 2, 2 (1960), 123–146.

[3] Ilhan Aslan, Tabea Schmidt, Jens Woehrle, Lukas Vogel, and Elisabeth André. 2018. Pen+ Mid-Air Gestures: Eliciting Contextual Gestures. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 135–144.

[4] Ilhan Aslan, Andreas Uhl, Alexander Meschtscherjakov, and Manfred Tscheligi. 2016. Design and Exploration of Mid-Air Authentication Gestures. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 6, 3 (2016), 1–22.

[5] Ilhan Aslan, Feiyu Xu, Hans Uszkoreit, Antonio Krüger, and Jörg Steffen. 2005. COMPASS2008: Multimodal, Multilingual and Crosslingual Interaction for Mobile Tourist Guide Applications. In *International Conference on Intelligent Technologies for Interactive Entertainment*. Springer, 3–12.

[6] Nihat Ay, Jessica Flack, and David C Krakauer. 2007. Robustness and Complexity Co-Constructed in Multimodal Signalling Networks. *Philosophical Transactions of the Royal Society B: Biological Sciences* 362, 1479 (March 2007), 441–447. https://doi.org/10.1098/rstb.2006.1971

[7] Jonathan T. Barron. 2019. A General and Adaptive Robust Loss Function. arXiv:1701.03077 [cs.CV]

[8] Tobias Baur, Alexander Heimerl, Florian Lingenfelser, Johannes Wagner, Michel F. Valstar, Björn Schuller, and Elisabeth André. 2020. eXplainable Cooperative Machine Learning with NOVA. *KI - Künstliche Intelligenz* (Jan. 2020). https://doi.org/10.1007/s13218-020-00632-3

[9] Björn Bittner, Ilhan Aslan, Chi Tai Dang, and Elisabeth André. 2019. Of Smarthomes, IoT Plants, and Implicit Interaction Design. (2019).

[10] Sérgio Branco, André G Ferreira, and Jorge Cabral. 2019. Machine Learning in Resource-Scarce Embedded Systems, FPGAs, and End-Devices: A Survey. *Electronicsweek* 8, 11 (2019), 1289.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. 1877–1901.

[12] Nicholas Capurso, Bo Mei, Tianyi Song, Xiuzhen Cheng, and Jiguo Yu. 2018. A Survey on Key Fields of Context Awareness for Mobile Devices. *Journal of Network and Computer Applications* 118 (2018), 44–60.

[13] Hong Cheng, Lu Yang, and Zicheng Liu. 2015. Survey on 3D Hand Gesture Recognition. *IEEE transactions on circuits and systems for video technology* 26, 9 (2015), 1659–1673.

[14] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1724–1734. https://doi.org/10.3115/v1/D14-1179

[15] Keum San Chun, Ashley B Sanders, Rebecca Adaimi, Necole Streeper, David E Conroy, and Edison Thomaz. 2019. Towards a Generalizable Method for Detecting Fluid Intake with Wrist-Mounted Sensors and Adaptive Segmentation. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 80–85.

[16] Simon Flutura, Andreas Seiderer, Ilhan Aslan, Chi-Tai Dang, Raphael Schwarz, Dominik Schiller, and Elisabeth André. 2018. Drinkwatch: A Mobile Wellbeing Application Based on Interactive and Cooperative Machine Learning. In *Proceedings of the 2018 International Conference on Digital Health*. 65–74.

[17] Genymobile. 2021. Scrcpy.

[18] Evan Glasheen, Antoinette Domingo, and Jochen Kressler. 2021. Accuracy of Apple Watch Fitness Tracker for Wheelchair Use Varies According to Movement Frequency and Task. *Annals of physical and rehabilitation medicine* 64, 1 (2021), 101382.

[19] Francisco Javier González-Cañete and Eduardo Casilari. 2020. Consumption Analysis of Smartphone Based Fall Detection Systems with Multiple External Wireless Sensors. *Sensors* 20, 3 (Jan. 2020), 622. https://doi.org/10.3390/s20030622

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]

[21] Alexander Heimerl, Tobias Baur, Florian Lingenfelser, Johannes Wagner, and Elisabeth André. 2019. NOVA - a Tool for eXplainable Cooperative Machine Learning. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 109–115. https://doi.org/10.1109/ACII.2019.8925519

[22] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. 2020. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. http://arxiv.org/abs/2006.16241. arXiv:2006.16241 [cs, stat]

[23] Sepp Hochreiter. 1998. The Vanishing Gradient Problem during Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6, 02 (1998), 107–116.

[24] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural computation* 9, 8 (1997), 1735–1780.

[25] Ling Huang, Anthony D. Joseph, Blaine Nelson, Benjamin I.P. Rubinstein, and J. D. Tygar. 2011. Adversarial Machine Learning. In *Proc. AISec*. 43–58.

[26] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. 2016. SqueezeNet: AlexNet-level Accuracy with 50x Fewer Parameters and ¡0.5MB Model Size. arXiv:1602.07360 [cs.CV]

[27] Abhinandan Jain, Adam Haar Horowitz, Felix Schoeller, Sang-won Leigh, Pattie Maes, and Misha Sra. 2020. Designing Interactions beyond Conscious Control: A New Model for Wearable Interfaces. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–23.

[28] Katarzyna Janocha and Wojciech Marian Czarnecki. 2017. On Loss Functions for Deep Neural Networks in Classification. *CoRR* abs/1702.05659 (2017). arXiv:1702.05659

[29] Steven G. Johnson and Matteo Frigo. 2007. A Modified Split-Radix FFT with Fewer Arithmetic Operations. *IEEE Transactions on Signal Processing* 55, 1 (2007), 111–119. https://doi.org/10.1109/TSP.2006.882087

[30] Jesper Kjeldskov and Mikael B Skov. 2014. Was It Worth the Hassle? Ten Years of Mobile HCI Research Discussions on Lab and Field Evaluations. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services*. 43–52.

[31] Gierad Laput and Chris Harrison. 2019. Sensing Fine-Grained Hand Activity with Smartwatches. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300568

[32] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. Viband: High-fidelity Bio-Acoustic Sensing Using Commodity Smartwatch Accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 321–333.

[33] Min Lin, Qiang Chen, and Shuicheng Yan. 2014. Network in Network. arXiv:1312.4400 [cs.NE]

[34] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2019. Towards Deep Learning Models Resistant to Adversarial Attacks. http://arxiv.org/abs/1706.06083. arXiv:1706.06083 [cs, stat]

[35] Wei Miao, Gongfa Li, Ying Sun, Guozhang Jiang, Jianyi Kong, and Honghai Liu. 2016. Gesture Recognition Based on Sparse Representation. 11, 4 (2016).

[36] Sushmita Mitra and Tinku Acharya. 2007. Gesture Recognition: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 37, 3 (2007), 311–324.

[37] Yasser Mohammad, Kazunori Matsumoto, and Keiichiro Hoashi. 2018. Primitive Activity Recognition from Short Sequences of Sensory Data. *Applied Intelligence* 48, 10 (Oct. 2018), 3748–3761. https://doi.org/10.1007/s10489-018-1166-6

[38] Meinard Müller. 2015. *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications* (1st ed. 2015 ed.). Springer International Publishing : Imprint: Springer, Cham. https://doi.org/10.1007/978-3-319-21945-5

[39] Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck. 1999. *Discrete-Time Signal Processing* (2nd ed ed.). Prentice Hall, Upper Saddle River, N.J.

[40] Guillermo Ortiz-Jiménez, Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. 2020. Redundant Features Can Hurt Robustness to Distribution Shift. (2020), 8.

[41] Shreyas Padhy, Zachary Nado, Jie Ren, Jeremiah Liu, Jasper Snoek, and Balaji Lakshminarayanan. 2020. Revisiting One-vs-All Classifiers for Predictive Uncertainty and Out-of-Distribution Detection in Neural Networks. (2020), 10.

[42] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. 2020. Understanding and Mitigating the Tradeoff Between Robustness and Accuracy. http://arxiv.org/abs/2002.10716. arXiv:2002.10716 [cs, stat]

[43] Shriti Raj, Kelsey Toporski, Ashley Garrity, Joyce M Lee, and Mark W Newman. 2019. " My Blood Sugar Is Higher on the Weekends" Finding a Role for Context and Context-Awareness in the Design of Health Self-Management Technology. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[44] Sebastian Raschka. 2020. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. *arXiv:1811.12808 [cs, stat]* (Nov. 2020). arXiv:1811.12808 [cs, stat]

[45] Joachim Rathmann, Christoph Beck, Simon Flutura, Andreas Seiderer, Ilhan Aslan, and Elisabeth André. 2020. Towards Quantifying Forest Recreation: Exploring Outdoor Thermal Physiology and Human Well-Being along Exemplary Pathways in a Central European Urban Forest (Augsburg, SE-Germany). *Urban Forestry & Urban Greening* 49 (2020), 126622.

[46] Siddharth S Rautaray and Anupam Agrawal. 2015. Vision Based Hand Gesture Recognition for Human Computer Interaction: A Survey. *Artificial intelligence review* 43, 1 (2015), 1–54.

[47] Marco C Rozendaal, Boudewijn Boon, and Victor Kaptelinin. 2019. Objects with Intent: Designing Everyday Things as Collaborative Partners. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 4 (2019), 1–33.

[48] Lamyaa Sadouk, Taoufiq Gadi, and El Hassan Essoufi. 2020. Robust Loss Function for Deep Learning Regression with Outliers. In *Embedded Systems and Artificial Intelligence*, Vikrant Bhateja, Suresh Chandra Satapathy, and Hassan Satori (Eds.). Advances in Intelligent Systems and Computing, Vol. 1076. Springer Singapore, Singapore, 359–368. https://doi.org/10.1007/978-981-15-0947-6_34

[49] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Mądry. 2018. How Does Batch Normalization Help Optimization?. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 2488–2498.

[50] Monika Schak. 2019. Robustness of Deep LSTM Networks in Freehand Gesture Recognition. In *Proc. ICANN*. 330–343.

[51] Gianluca Schiavo, Alessandro Cappelletti, Eleonora Mencarini, Oliviero Stock, and Massimo Zancanaro. 2014. Overt or Subtle? Supporting Group Conversations with Automatically Targeted Directives. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*. 225–234.

[52] Andreas Seiderer, Michael Dietz, Ilhan Aslan, and Elisabeth André. 2018. Enabling Privacy with Transfer Learning for Image Classification DNNs on Mobile Devices. In *Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good*. 25–30.

[53] Muhammad Shafique, Mahum Naseer, Theocharis Theocharides, Christos Kyrkou, Onur Mutlu, Lois Orosa, and Jungwook Choi. 2020. Robust Machine Learning Systems: Challenges, Current Trends, Perspectives, and the Road Ahead. *IEEE Design Test*. 37, 2 (2020).

[54] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. 2020. The Pitfalls of Simplicity Bias in Neural Networks. http://arxiv.org/abs/2006.07710. arXiv:2006.07710 [cs, stat]

[55] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV]

[56] Julius O. Smith. 2011. *Spectral Audio Signal Processing*.

[57] Nicolas Sturmel, Laurent Daudet, et al. 2011. Signal Reconstruction from STFT Magnitude: A State of the Art. In *International Conference on Digital Audio Effects (DAFx)*. 375–386.

[58] Andreas Tobola, Franz J. Streit, Chris Espig, Oliver Korpok, Christian Sauter, Nadine Lang, Björn Schmitz, Christian Hofmann, Matthias Struck, Christian Weigand, Heike Leutheuser, Björn M. Eskofier, and Georg Fischer. 2015. Sampling Rate Impact on Energy Consumption of Biomedical Signal Processing Systems. In *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. 1–6. https://doi.org/10.1109/BSN.2015.7299392

[59] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2019. Robustness May Be at Odds with Accuracy. http://arxiv.org/abs/1805.12152. arXiv:1805.12152 [cs, stat]

[60] Victor H. Vimos, Marco Benalcázar, Alex F. Oña, and Patricio J. Cruz. 2019. A Novel Technique for Improving the Robustness to Sensor Rotation in Hand Gesture Recognition Using sEMG. In *Proc. CSEI*. 226–243.

[61] Haoxuan Wang, Anqi Liu, Yisong Yue, and Anima Anandkumar. 2020. Deep Robust Classification under Domain Shift with Conservative Uncertainty Estimation. (2020), 5.

[62] Yanwen Wang, Jiaxing Shen, and Yuanqing Zheng. 2020. Push the Limit of Acoustic Gesture Recognition. *IEEE Transactions on Mobile Computing* (2020).

[63] Hongyi Wen, Julian Ramos Rojas, and Anind K Dey. 2016. Serendipity: Finger Gesture Recognition Using an off-the-Shelf Smartwatch. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3847–3851.

[64] Mathias Wilhelm, Jan-Peter Lechler, Daniel Krakowczyk, and Sahin Albayrak. 2020. Ring-Based Finger Tracking Using Capacitive Sensors and Long Short-Term Memory. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 551–555.

[65] Huan Xu and Shie Mannor. 2012. Robustness and Generalization. 86 (2012).

[66] Kei Yaguchi, Kazukiyo Ikarigawa, Ryo Kawasaki, Wataru Miyazaki, Yuki Morikawa, Chihiro Ito, Masaki Shuzo, and Eisaku Maeda. 2020. Human Activity Recognition Using Multi-Input CNN Model with FFT Spectrograms. (2020), 4.

[67] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. http://arxiv.org/abs/1901.08573. arXiv:1901.08573 [cs, stat]

[68] Jingfeng Zhang, Bo Han, Laura Wynter, Kian Hsiang Low, and Mohan Kankanhalli. 2019. Towards Robust ResNet: A Small Step but A Giant Leap. (Feb. 2019).

[69] Ru Zhang, Yuanchun Shi, Björn Schuller, Elisabeth André, Sharon Oviatt, Aaron Quigley, Nicolai Marquardt, Ilhan Aslan, and Ran Ju. 2021. User Experience for Multi-Device Ecosystems: Challenges and Opportunities. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–5.