Out-of-distribution in Human Activity Recognition

Debaditya Roy[1],Vangjush Komini[1], Sarunas Girdzijauskas[2]

***Abstract*— With the growing interest of the research community in making deep learning (DL) robust and reliable, detecting out-of-distribution (OOD) data has become critical. Detecting OOD inputs during test/prediction allows the model to account for discriminative features unknown to the model. This capability increases the model's reliability since this model provides a class prediction solely at incoming data similar to the training one. OOD detection is well established in computer vision problems. However, it remains relatively under-explored in other domains such as time series (i.e., Human Activity Recognition (HAR)). Since uncertainty has been a critical driver for OOD in vision-based models, the same component has proven effective in time-series applications.**

We plan to address the OOD detection problem in HAR with time-series data in this work. To test the capability of the proposed method, we define different types of OOD for HAR that arise from realistic scenarios. We apply an ensemble-based temporal learning framework that incorporates uncertainty and detects OOD for the defined HAR workloads. In particular, we extract OODs from popular benchmark HAR datasets and use the framework to separate those OODs from the in-distribution (ID) data. Across all the datasets, the ensemble framework outperformed the traditional deep-learning method (our baseline) on the OOD detection task.

I. Introduction

Deep learning (DL) methods for HAR are integral to many ubiquitous applications. E.g., providing live coaching feedback to an athlete based on mobile or on-body sensor data requires an efficient HAR algorithm. The algorithm predicts if the person is running, walking, jogging, etc., and coaching feedback is generated based on that prediction. In such applications, it is common to encounter unseen out-of-distribution (OOD) activities with respect to known or in-distribution (ID) activities. E.g., taking a rest while running or performing some spontaneous activity such as taking a phone call. The model does not know the above activities (hence OOD). Therefore, it must differentiate the OOD data from ID data in those scenarios. Failing to do so leads to misclassification, affecting model reliability. However, most state-of-the-art DL models used for HAR fail to do so. The primary reason is that these models are trained to discriminate between classes with high accuracy without considering the inherent uncertainty present in the data, and the model [37].

This inability of traditional neural networks to accommodate the uncertainty compromises their robustness when encountering OOD data. In other words, traditional neural networks do not estimate the uncertainty for the prediction in the test data. Therefore, these models assert an equal confidence prediction for both OOD and ID test data.

Therefore, uncertainty estimation is an essential task for detecting OODs. Unfortunately, OOD detection is an under-explored task in the wearable sensor-based HAR domain. Therefore, in this work, we try to empower traditional neural networks at detecting OODs from time-series recordings of human activities. To properly evaluate the OOD detection capabilities, we first need to propose OOD data that tries to expose the HAR models to unforeseen discriminative features.

In addition to that, we discuss how an ensemble-based temporal learning framework provides sufficient uncertainty to separate OODs from ID data in HAR tasks.

We begin by defining activity-based OOD examples. Since there can be a wide range of different activities, we need to narrow down the space of activities such that the defined OODs possess distinguishable features to the training data. Hence we consider *walking, running, jogging, standing, sitting, walking upstairs, and walking downstairs* activities. Based on these activities, we define two categories of in-distribution (ID) versus out-of-distribution (OOD) examples, namely;

1) *Dynamic activities (ID) versus Static activities (OOD):* In this set, the dynamic activities such as *running, walking, etc.* are used to train the model (ID) and the static activities such as *standing, sitting* are used as OOD. This type of OOD can be encountered in real life. E.g., static activities can be out-of-distribution in an athletic or sporting scenario where predictions are usually dynamic sporting activities. Detecting static OODs in those scenarios can help reduce the misclassification error of the model.
2) *Known activities (ID) versus Unknown activities (OOD):* In this set, all but one activity are used to train the model, and the left out activity is used as OOD. This kind of scenario might occur in the real world, where we do not know about certain activities relevant to the HAR application. Detecting those activities as OODs allows discovering a new class that can be incorporated into the relevant activities.

To quantify the data uncertainty in this temporal setting, we have used a method from our previous work [1] called

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 813162. The content of this paper reflects the views only of their author (s). The European Commission/ Research Executive Agency are not responsible for any use that may be made of the information it contains.

[1]D. Roy and V. Komini is with Department of Electrical Engineering and Computer Science (EECS), Royal Institute of Technology (KTH), Stockholm, Sweden and, with Qamcom Research and Technology, Stockholm.

[2]S. Girdzijauskas is with Department of Electrical Engineering and Computer Science (EECS), Royal Institute of Technology (KTH), Stockholm, Sweden.

Deep time Ensembles (DTE). This method has two consecutive parts, i) extracting different temporal information by sliding window lengths of different sizes on the raw time-series input and ii) using that temporal information to train multiple models for ensembling. Initially, it was proposed to improve classification and calibration metrics for HAR tasks. Since calibration is a direct reflection of data uncertainty, in this work, we show that estimating uncertainty using DTE is also effective for OOD detection. However, traditional neural networks in HAR fail to detect OOD.

Traditional neural networks cannot internalize an entire recorded activity. Instead, they truncate the time-series recording to multiple temporal sequences extracted with the same window size. The fixed-size window will implicitly induce bias in the predictive response. In other words, the model is missing out on the data variability (uncertainty) since this requires a window size that corresponds to the designated activity's entire duration. Compounding different predictive responses conveying incoherent biases induced by the different window sizes can enhance the data's inherent (coherent) uncertainty. As a result, the prediction variance reflects the data uncertainty, and the incoherent biases are averaged out.

To attain coherent compounding within the *DTE* [1] setup, an ensemble of models is trained with sequences extracted with different window sizes. At inference time, the model's predictive responses are averaged out. Different window sizes for extracting temporal sequence induces distinct bias in each model during training. Combining the predictive output from each model increases the uncertainty in data and reduces the uncertainty from the window sizes. This coherent compounding can also be explained through the lens of softmax function. Since the learning in *DTE* is distinct across models, so are their softmax outputs as well. Eventually, these distinct outputs, when combined, convey more uncertainty through a smoother softmax. In the case of ID data, a smoother softmax increases the values for incorrect classes predictions; nevertheless, it still reaches the necessary consensus for the correct class. However, this consensus does not hold for OOD data since the averaging would bring the class predictions close to a uniform distribution. This uniformity is a direct consequence of the harnessed data uncertainty by combining ensemble models. Hence, allowing it to estimate uncertainty in the OOD inputs with success.

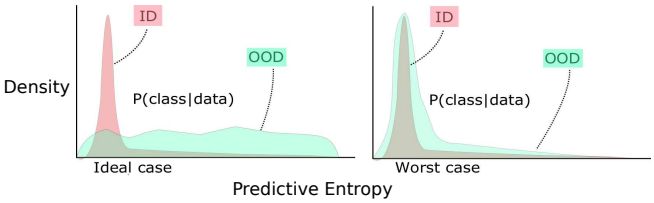

Fig. 1: Predictive Entropy Distribution for ID and OOD.

This behavior can be visualized through the distribution of the predictive entropy (c.f Figure 1). Since the training process reduces the entropy of the predictive response in the ID, the underlying assumption is that OOD samples maintain a higher entropy of their predictive distribution. In the case of ID data, the majority of the entropy coming out of the ensemble will be around zero, whereas the OOD data will push the entropy towards positive values.

Estimating the density shift of the entropy distribution for OODs serves as the primary metric for their detection. This shift can be estimated by measuring the overlapping area under the curve between ID and OOD entropy distribution. Therefore, Weitzman's Measure [32] calculating the overlapping area under two functions could serve as an elegant metric for the density shift. The lower the intersection area, the better the OOD detection for the model and vice-versa. This metric serves better than traditional divergence-based metrics that compare two different distributions. Divergence-based methods are harder to compute and can possess numerical problems when the modality of two distributions is not equal. In contrast, the Weitzman's Measure would be able to quickly tell that there is a noticeable overlap between these distributions and hence will provide the correct OOD score.

Deep time-ensemble [1] is tested on defined OODs extracted from three popular sensor-based HAR datasets, WISDM [33], UCI [34] and Motion-Sense [35]. It outperforms the baseline model [19] (a CNN architecture adopted from previous work) on the OOD detection tasks on the introduced metrics for the chosen datasets.

The main contribution of this paper is defining simple and effective OODs for HAR with wearable time-series data and using our previous ensemble-based temporal framework to detect them. We also provide reasoning to explain why *DTE* suits OOD detection tasks with temporal workloads. While a few works address uncertainty estimation in HAR, to the best of our knowledge, none of them address the OOD detection paradigm.

The paper is organized as follows: Section II discusses the *Related Work*. Following this, the *Methods* argues the rationale behind using *DTE* for OOD detection. The initial *Experiment and Results* are presented in Section IV, followed by a *Discussion* and, *Conclusion and Future Works* section.

II. Related Work

Three categories of work are discussed in this section: The first one discusses the works in *general activity recognition*, their evolution, and where this article resides in that paradigm. The second group discusses *uncertainty estimation* methods in deep-learning and compares how the method used in this paper is different. The third group reviews an *intersection of both* the above groups and compares this work with its closest alternatives. Finally, a fourth class of related work review on *anomaly detection* in the context of time-series IoT environments.

A. On Activity Recognition

Activity recognition (AR) with wearable sensors is popularly achieved using deep learning methods [7], [8], [9], [10]. Popular choices of deep-learning architectures include CNN

based on 1D convolution of time-series data [8], recurrent neural-networks [15], autoencoder-based architectures [16]. However, these methods produce classification estimates without addressing data or model uncertainty. Hence, they react inefficiently with concept drift in the dataset and fail on the OOD detection task. *Deep time-ensemble* method [1], used in this paper, can adapt to any deep-learning architecture and detect OOD successfully.

B. On Uncertainty Estimation

Estimating uncertainty in neural networks makes them more robust and helps detect possible domain-shift in the data. From a probabilistic viewpoint uncertainty aware neural networks can be classified into *Bayesian* [17], [19], [20] and,*Frequentist* [18], [21]. A popular application of Bayesian formalization in neural networks is to learn a probability distribution of the neural network weights that helps in uncertainty estimation [17]. However, the complexity of training a Bayesian neural network (for many applications) and complex prior assumptions motivated the researchers to explore probabilistic estimation using standard neural networks. Ensemble-based methods have generated much traction in recent years due to their ability to estimate uncertainty while utilizing standard neural network architectures [21], [23], [22]. Vyas et al. [18] formulated a loss function that, when added to standard cross-entropy loss of a neural network, increases the margin of separation between ID and OOD samples of images. They train an ensemble of neural networks in a self-supervised fashion with this composite loss function. Lakshminarayanan et al. [21] establish that through ensembling and adversarial training, deterministic neural networks can be enforced to estimate uncertainty. They show success in examples from computer-vision and standard regression datasets.

However, the ensembling method proposed in our earlier paper is designed to adapt time-series workloads that were not explored in [18], [21]. Other works that addressed OOD detection explicitly are [24], [25], [22]. A work by Hendrycks et al. [24] showcases that although viewing softmax output in isolation can be misleading to identify OOD and ID samples, collecting the global statistics about the softmax of ID samples can help differentiate ID and OOD samples. Lee et al. [25] separate OOD and ID by training a GAN network, with cross-entropy loss and a loss term based on distance from a uniform distribution. Liang et al. [22] increased the margin between ID and OOD samples by having temperature scaled softmax outputs into the cross-entropy loss and small perturbations in the training example.

Most of these methods discussed for uncertainty estimation and OOD detection have been extensively tested and benchmarked on computer vision problems and regression tasks. However, they are a bit under-explored in the context of time-series data, in particular HAR. This paper leverages the ensemble-based uncertainty estimation concept and demonstrates its capability for sensory recordings of time-series data on HAR tasks.

C. A combination of both

While activity recognition is a well-established field, with uncertainty estimation in deep learning not far behind, the combination of both is relatively new. While Nweke et al. [26] researched the usage of sensor-level ensemble stacking and improved misclassification in HAR, they did not investigate the impact of OOD on their models. Hue et al. [27] explore annotation uncertainty and mitigate the issue by a soft-labeling strategy. They tackle an uncertainty problem in activity recognition with RGB-D frames. It is comparatively more manageable for the annotator to assign soft labels to ambiguous activities with visual aid. Hence the model gets more labelling assistance for uncertainty estimation. On the other hand, *deep time-ensemble* [1] deals strictly with wearable-sensor-based activity recognition, where it is more complicated to annotate ambiguous labels from sensor data. Akbar et al. [28], [29] approaches uncertainty in activity recognition with generative modelling. Akbar et al. [29] proposes a method that allows easy integration of new sensor data with models trained with the older sensor data. The closest match to this work is with [28], where the authors propose a Bayesian CNN as a variational autoencoder [30] for estimating the density of sensor signals. Later they use the estimated density to sample and classify activities as a downstream task with a classification layer and Monte-Carlo dropout [19]. While it is an elegant way to estimate uncertainty, the authors [28] does not explicitly explore the OOD detection ability. The used *deep time-ensemble* [1] is potentially more straightforward (because of the non-Bayesian approach) and is used for detecting OODs in HAR tasks.

D. On Anomaly Detection

Anomaly detection with machine learning is a problem where the models try to identify anomalous classes, i.e. classes which are outlier with respect to the classes the model has been trained on. Intuitively, out-of-distribution data on other hand might be an outlier class or simply a new class that has not been encountered before. The requirement of an ood class (outlier or not) must be that it will be drawn from a distribution that is statistically different from the training data distribution. While, there is not a strict separation between the two, but in our understanding anomaly detection is a stricter variant of of out-of-distribution detection. Some applications of ood detection might deal with outlier classes and hence it is important to look at the existing state-of-the-art in anomaly detection.

In particular anomaly detection using machine learning has garnered enough attention in the domain of IoT and smart-home [38]. Mining time-series stream data [39], user-behaviour [40] has provided security solutions. Detecting medical events [41], [42] is also another important use-case in sensor driven anomaly detection problem. However, most of the existing anomaly detection techniques using machine learning demand exposure to few examples of anomalous data. This might be restrictive in certain context. Using our proposed out-of-distribution (OOD) in anomaly detection

task we can bypass the requirement to have some anomalous training data (see Appendix VI-A) for more insights on this.

III. Methods

A. Integrity of uncertainty in deep learning

Harvesting the necessary uncertainty presented in the data has proven beneficiary in increasing the awareness of a machine learning model toward OOD test data. In a deep learning classifier, usually, the softmax layer accommodates some stochastic behaviour of the data. The rest of the layers are dedicated to generalizing the representative features. On the one hand, the high number of such components (layers) makes a classifier more capable of achieving high predictive accuracy. On the other hand, the single last layer consolidates uncertainty from the data to a much lesser extent than needed. Based on the bias-variance trade-off, the more uncertainty incorporated in the model, the lesser unwanted bias in the final prediction. Thus, stochasticity must be included in the layers before the softmax to make the model less deterministic. An ensemble is a popular approach for delivering such behaviour. The multiple models in the ensemble simulate the behaviour of such stochastic layers.

B. Problem Setup

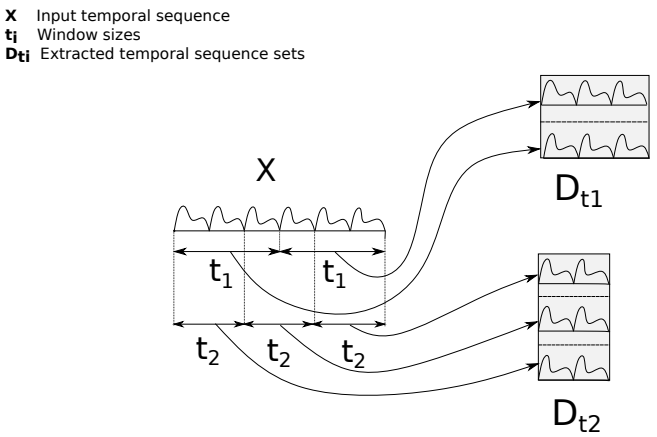

Fig. 2: Representing a temporal-sequence X as collection of different temporal sequences by using different window-sizes.

A temporal sequence X is obtained by sliding a particular window size over the raw time-series data. The goal is to predict the activity based on the observations defined by X. This temporal sequence X is fed to a neural network $p_\theta(y|X)$ parameterized by θ to produce a softmax output y. The softmax y is the probabilistic output provided by a model, and the index of the highest value in y represents the index of the predicted class.

Temporal sequence, X, can be re-represented as a combination of multiple temporal sequences extracted with different window sizes. As shown in Figure 2, X is represented as temporal sequence sets D_{t_1} and D_{t_2}, that contains multiple temporal sequences. These temporal sequence sets are extracted from the temporal sequence X by sliding window sizes t_1 and t_2. For more details of this setup, we refer to our previous work [1].

C. Deep time ensembles for uncertainty estimation

$$Y = \frac{1}{M}\sum_{i=1}^{M} p_{\theta_i}(y_i|X) \tag{1}$$

Time series recordings have the structural information encoded into their temporal order. Namely, whenever a particular trend is present in a time series, it will persist throughout consecutive recorded values. The data fed into the model is strictly ordered by the acquisition time. The scope of exploration depends explicitly on the number of consecutive values fed into the model. Traditionally, in activity recognition problems, this is achieved by extracting data (a temporal sequence) using a fixed window size and feeding it to the model. Thus, extracting the temporal information is highly dependent on the window size of sensor readings. A naive ensembling technique that trains multiple models using identical temporal sequences obtained by using the same window size is sub-optimal. The equation of output from such ensemble is given by eqn 1. Having the same window size would extract the same temporal sequence X, that is fed to the M models in the ensemble. It conveys the same information to each model in the ensemble. The only source of randomness, in this case, is obtained through converged weight values θ_i for different trained models p_{θ_i}. Even then, training from the same temporal sequences would mean that the converged weights θ_i are similar. Meaning that all the models would produce similar predictive trends, limiting the randomness. Hence, the averaging of the ensemble at the output does not harness sufficient uncertainty. Bootstrapping different data samples might mitigate this drawback, and *DTE* [1] offers a different form of bootstrapping fit for time-series data.

$$Y = \frac{1}{M}\sum_{i=1}^{M} p_{\theta_i}(y_i|D_{t_i}) \tag{2}$$

In *DTE*, bootstrapping is achieved by representing the same input differently, i.e., by extracting different temporal sequences from same data (c.f Figure 2) with different window sizes. Different temporal sequences are used to train each model of the ensemble. Similar to experience replay [36], having a varying window size increases the decorrelation between the temporal sequences and thus increases the empirical variance that is exposed to the models. Furthermore, temporal sequences of different window sizes boost the ability to explore higher-order dependencies in time series. As a result of this, it is possible to broaden the exploration capacity of the model by observing more structural information. The equation of output from *DTE* is shown in eqn 2. In the equation, the term D_{t_i} represents sets of temporal sequences that can be extracted from a single temporal sequence X. Figure 2 demonstrates a simple example showing how two set of temporal sequences (D_{t_1} and D_{t_2}) are extracted from X using windows t_1 and t_2. Thus,

without losing any information, each model of *DTE* learns a different temporal dimension of the time-series classification problem. It allows them to produce distinct outputs and model different uncertainty trends in the data. This increases the overall randomness of the process sufficiently towards efficient uncertainty estimation by averaging the softmax.

On top of the softmax averaging, in the combination step of ensembling, *DTE* also features nested averaging. This is evident from Figure 2. When temporal sequence X is represented as a collection of temporal-sequences D_{t_1} and fed to a model in the ensemble, two temporal sequences are sent to the model. Hence, the model produces two outputs for two temporal sequences. Since a single prediction is required from input X, the outputs are averaged. Similarly, for D_{t_2}, three predictions are averaged by the model to produce a single response against X. A more detailed explanation of this process can be found in [1]. The uncertainty obtained through the nested averaging adds to the uncertainty obtained in the ensemble combination step. The total accumulated uncertainty results in better OOD detection.

D. Density of entropies

The coherent uncertainty of the model is mainly a consequence of the incompatible prior assumption made for the model. This type of uncertainty persists through the training process for the given data. The incoherent uncertainty of the model is then a direct consequence of the imperfect calibration of the hyperparameter of the model. This type of uncertainty is generally misleading as it is just a subjective reflection of the hyperparameters and clutters the coherent uncertainty. An ensemble can suppress the incoherent uncertainty, eventually decluttering the model uncertainty. The predictive output comes as a normalized distribution degree of belief, also known as softmax output (cf. Eqn 3).

$$P(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}, \forall i \in \{1, ..., N\} \tag{3}$$

The overall uncertainty represented by the softmax is then compressed into a single empirical value using the entropy. The entropy weights the softmax output by the amount of information that a particular output possesses (cf. Eqn 4). The more uniform the predictive output is, the more uncertain the model is, resulting in a higher entropy value.

$$H(P(x_i)) = \sum_i P(x_i) * log\{P(x_i)\} \tag{4}$$

Using a collection of entropies that fully characterizes the ensemble represents the total amount of uncertainty produced by the model for a given test data. The model is highly confident whenever entropies are close to zero, and the uncertainty is relatively low. These cases are mostly related to ID data, where the model has been heavily trained. Whenever there are OOD data, the ensemble produces a frequent amount of high entropy values, given that predictive outputs are, on average, distributed at equal probabilities. Furthermore, having a density estimation from the collection of entropy values is then computationally attractive to judge the overall behaviour in an ensemble.

E. Density-shift estimation

$$wm = \int min\{Density_1(x), Density_2(x)\}dx \tag{5}$$

The final output of an ensemble is a distribution of entropy values, and assessing the discrepancy between two different distributions in the context of density shift is not as trivial. OOD distribution entropies are expected to shift their density mass towards high positive values, whereas ID entropies remain around zero. A measure that fits the need the most should target the density shift and put lower importance on the shape of distributions. Weitzman measure (cf. Eqn 5) quantifies this shift quite elegantly by measuring the amount of intersecting area between two distributions. The further apart from one another two distributions are, the lower the Weitzman measure (wm) is and the better the OOD detection capability of the model.

IV. Experiment and Results

We have used a *CNN* architecture adapted from Ignatov et al. [8] as the baseline model and ensembled the same architecture using *DTE*. Since, *DTE* [1] also compared with Ignatov et al. [8] using the same *CNN* architecture, the configurations of the window-sizes and hyper-parameters were directly adopted from [1]. The goal of the experiments was to detect OODs extracted from standard HAR datasets. For classification results, we would refer to our previous work [1].

A. Datasets

We have primarily defined OODs based on activities from WISDM [33], UCI [34] and Motion-Sense [35] dataset. Below we discuss the datasets and how they are used to formulate ID and OOD data.

1) WISDM dataset: : The WISDM dataset consists of accelerometer recordings from six activities, namely, *walking, jogging, upstairs, downstairs, sitting, standing* obtained from 36 subjects. The *baseline model* is trained on 200 timesteps indicating 10 seconds of time-series data [8], [1]. The *DTE* is trained on time-steps ranging from 200 to 100 [1]. The WISDM dataset is divided based on the activities. Dynamic activities *walking, jogging, upstairs, downstairs* form the *Dynamic WISDM* dataset, and static activities, i.e. *sitting and standing* forms the *Static WISDM*. *Dynamic WISDM* is used for training the models. *Static WISDM* is used as an OOD set.

2) UCI Dataset:: The UCI dataset consists of 6 activities *lying, standing, sitting, downstairs, upstairs, and walking* recorded from 30 subjects. The modality of the sensor is a triaxial accelerometer and gyroscope resulting in 6 dimensions. It is used as ID data for our experiments. The baseline model is trained on 256 steps of UCI dataset ([8], [1]. *DTE* trains 5 models between the range of 128 to 256 timesteps [1].

3) Motion-sense OOD dataset[35]: : This dataset consists of accelerometer and gyroscope readings from six activities, *downstairs, upstairs, sitting, standing, jogging, walking*. Since the UCI dataset does not have an instance of jogging, the input signal(accelerometer and gyroscope) for jogging is used as an OOD input to the models trained on the UCI dataset.

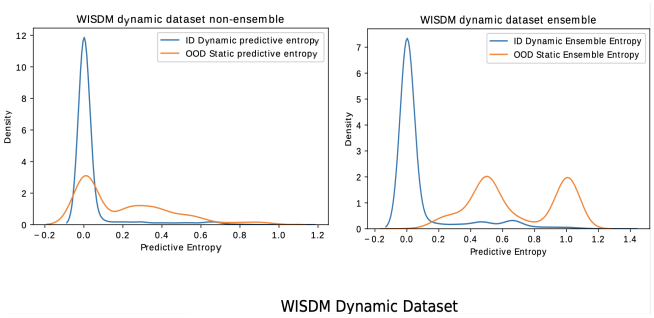

Fig. 3: Comparing OOD detection of a single model against Deep Time Ensembles. The training dataset consist dynamic activities from WISDM and the OODs are the static activities.

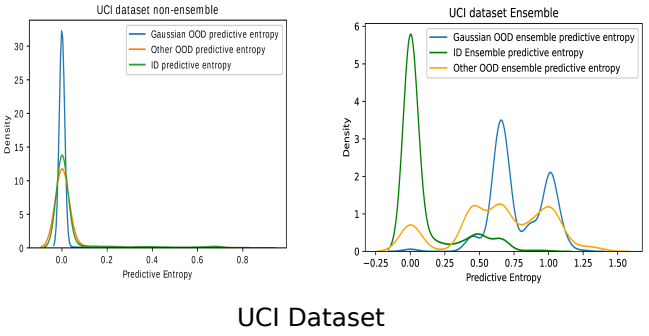

Fig. 4: Comparing OOD detection of a single model against Deep Time Ensembles. The training dataset consist dynamic activities from UCI and the OOD is Jogging from Motion-Sense dataset (other OOD in image) and, random Gaussian noise.

B. Uncertainty estimation: OOD vs ID inputs

Based on the definitions of OOD in the Introduction, two experiments were formulated.

- Dynamic WISDM as training and Static WISDM as OOD.
- Training on full UCI dataset and using *jogging* from Motion-Sense dataset as OOD.

The models are evaluated for uncertainty estimation by probing them with OOD inputs. In addition to probing the UCI dataset with OODs from the Motion Sense dataset, we also probe it with a random Gaussian noise drawn far from the original training data distribution ($\mu = 5, \sigma^2 = 6$).

Predictive entropy distribution of the output is a central concept for visualizing uncertainty estimation in classification tasks. When tested with OOD examples, a well-behaved model provides uncertain or low-confidence outputs. This essentially translates to higher predictive entropy for the outputs. However, for the ID samples, the confidence is higher, and hence predictive entropy of the output is lower. Thus an ideal model gives *low predictive entropy* for ID samples and *higher predictive entropy* for OOD samples. In terms of *Weitzman measure* used to evaluate the model, a lower score indicates better OOD detection.

TABLE I: Weitzman measure for different types of OOD - WISDM Dynamic

Model-type	WISDM static OOD
Ensemble	**0.16**
Non-ensemble	0.38

TABLE II: Weitzman measure for different types of OOD - UCI Dataset

Model-type	Gaussian-OOD	Motion-sense OOD
Ensemble	**0.10**	**0.29**
Non-ensemble	0.43	0.82

The hyperparameters and model architecture are presented in Appendix B. Next, we discuss the results of the experiments.

1) Probing Dynamic WISDM dataset with OODs: : The models trained on *Dynamic WISDM* dataset are probed with *Static WISDM* dataset as OOD. For the baseline model, the predictive entropy distribution resides in a low-value region with a mean around 0 for both ID and OOD sets 3. *DTE* reacts similarly for the ID test data. However, for the OOD, i.e., static data, the predictive entropy distribution of the outputs shifts further to the right. The clear margin of separation in entropy distribution between ID and OOD data for *DTE* reflects in the *Weitzman measure* metric as well. The *Weitzman measure* is **2.3** times lower for the *Deep time-ensemble* compared to the baseline model.

2) Probing UCI dataset with OODs: : The expected behaviour is well-captured in the experiments, as showcased by Figure 4. The *ID* test inputs to the *baseline* model produces low entropy values as expected. However, even for OOD datasets(*Gaussian* and *Motion-sense OOD*) the predictive entropy resides in the low-value region. It signifies overconfident misclassified outputs by the baseline.

DTE, on the other hand, produces high entropy values for OOD inputs and low-entropy values for the ID inputs. There exist a clear margin of separation between the ID and OOD datasets for *DTE* trained models. The above result is also quantified by the *Weitzman measure*, as shown in Table II. The *Weitzman measure* shows a **4-fold** decrease for the *DTE* compared to the *baseline*, for Gaussian OOD and **2.5 fold** decrease for *Motion-sense OOD dataset*. Lower *Weitzman measure* indicates lesser overlap between ID and OOD samples and hence a better margin of separation.

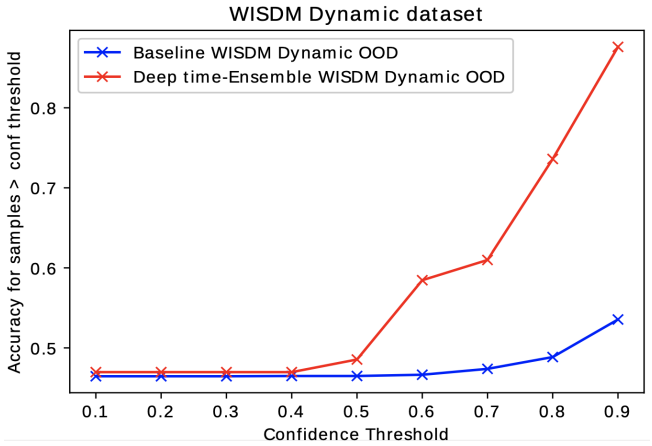

Fig. 5: Confidence versus Accuracy curve: Comparison among baseline model and DTE on WISDM dataset. OOD in this experiment is Static WISDM.

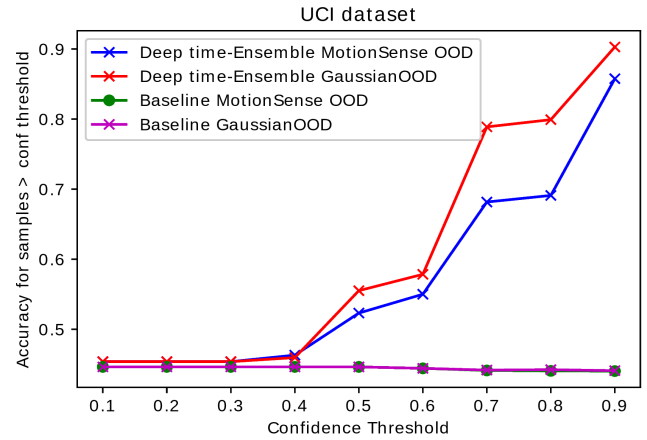

Fig. 6: Confidence versus Accuracy curve: Comparison among baseline model and DTE on UCI dataset. OOD in this experiment is Jogging from Motion-Sense and random Gaussian noise.

V. Discussion

A. On OOD detection

A couple of interesting observations can be made from the experiments.

- Firstly, it is clear from the *Weitzman measure* of different experiments that the random Gaussian noise is more out-of-distribution compared to the real datasets (see Figure 6). This behavior is expected because the random Gaussian noise was deliberately drawn away from the training distribution. It served as a first step to show that *DTE* could pull apart the Gaussian noise successfully as OOD from the ID samples. However, the single model failed to do so even though the Gaussian noise values do not contain any discriminative features that the model is trained on. As a result, the single model misclassified the random data and assigned a high score to one of the classes in the softmax output. In *DTE* however, even though individual models assign a high score (in softmax) to one of the classes, not all of them assign a high score to the same class. Thus, through averaging, the overall softmax output of the ensemble is smoothed to be more uniform. This uniformity translates to a higher entropy value for the predictive response, differentiating the OOD from ID.
- When *jogging* activity from *Motion-sense* dataset is used as an OOD, there is a small overlap in predictive entropy distribution even for *DTE* (see Figure 4). The similarity of jogging with some ID activities such as running and walking might be why. The discriminative features that are extracted by each model in *DTE* are not sophisticated enough to differentiate between jogging and some of the ID activities. There is a consensus among the models on the predicted class for some OOD data. Hence low uncertainty is obtained through model averaging. Thus, some portion of the realistic OOD dataset may be detected as ID, which explains the overlap. One possible way to mitigate this is by adding more layers to each model for improved feature extraction for each class. However, the models risk overfitting in the classification task. Since *DTE* is also optimized to obtain better predictions, overfitting each model would oppose that.
- All the images (Figure 3 and Figure 4) with *DTE* shows an interesting feature. There is a bump in the predictive entropy distribution on the right-hand side, even for ID samples. Since each model in *DTE* produces a distinct output, there are some cases where the uncertainty in ID data does not allow for a consensus for the majority class in the softmax output. In those cases, the randomness of the model increases during the averaging process. Intuitively, these can be some borderline examples of ID data that are corrupted by noise or incorrect annotation.

This experiment validates two important hypotheses that were proposed earlier.

1) Single deep learning models for the HAR tasks produce misclassifications in the face of OOD data because of their inability to capture uncertainty. This property may guide users towards wrong interpretation and make the models unreliable.
2) *Deep time-ensemble* can estimate the uncertainty associated with the OOD samples and establish a clear margin of separation. A property desired for robustness and reliability.

B. Accuracy vs Confidence

One way of analyzing the uncertainty estimation ability of a model is by looking at the accuracy vs. confidence curve. Given a model prediction $p(y|x)$ for k classes, the predicted label is given as $y_{true} = \arg\max(p(y|x))$ out of k classes. For such an output the *confidence* is defined as $conf = max(p(y|x))$. In this experiment, for all the test samples (both ID and OOD), the *confidence* values are extracted. Then based on a selected confidence threshold,

examples having a value greater than the threshold are retained. The model accuracy is calculated for the retained samples. A general intuition suggests that for an ideal model, lower confidence should also indicate low accuracy and vice-versa. Because for such a model, the under-confident samples bring down the model accuracy through misclassifications. Once those samples are rejected, the left-outs represent the confident predictions. These predictions contribute positively to the accuracy. For this experiment, OOD and ID examples are combined in an equal proportion. In Figure 5 and Figure 6, the images indicates the *accuracy vs confidence* curve for all the models on all the datasets and respective OOD sets. As seen from the figure, the *baseline model* fails to produce high accuracy even at high confidence thresholds for all the datasets. The *baseline model* makes high-confidence predictions even for the OOD, so at higher confidence thresholds, they contribute negatively to the accuracy through misclassifications. On the other hand, *Deep time-ensembles* are more robust in this aspect. The majority of the OOD samples are rejected at low confidence thresholds. Hence, correctly classified samples are retained at high-confidence points, improving the accuracy.

VI. Conclusion and Future Work

In this work, we set out to define and detect simple yet effective out-of-distribution (OOD) data for Human Activity Recognition (HAR) tasks. In particular, HAR problems with time-series data originating from wearable sensors. The defined OODs are usually encountered in realistic scenarios where HAR models are deployed. Although OOD detection is explored in the domain of computer vision and specific regression tasks, to the best of our knowledge, there is no previous work that defines OODs for time-series workloads originating in the HAR domain. To detect the proposed OODs, we have adopted an ensemble-based framework called *Deep Time Ensembles (DTE)* that takes the temporality of the workload into account. In particular, we have used a baseline convolutional neural architecture that yielded promising results in HAR tasks from [8] and ensembled it using *DTE*. Our experiments on OOD data extracted from popular HAR datasets indicate that *DTE* outperforms the baseline model in detecting OODs. This nature will allow *DTE* to be incorporated for modelling robust and reliable solutions in HAR. It opens up exciting research avenues to be explored in the future, e.g., incorporating OOD detection with *DTE* in safety-critical HAR applications. Initial results and a possible architecture of one such application, *Elderly Fall Detection* is presented in Appendix A. In our use-case with Fall Detection, we tried to show how uncertainty estimation could detect falls with fewer data.

While this work delves into some initial and simple OOD definitions for the HAR task, more interesting OODs for time-series data could be defined in the future. E.g., In-distribution (ID) data originating from periodic activities versus OOD originating from aperiodic activities. Also, using OOD metrics to measure uncertainty arising from IoT devices could lead to better fine-tuning and calibration. Although ensembling captures stochasticity, the inherent complexity of training an ensemble can still be an issue. However, progress in distilling the ensemble models [23] is a direction that could be explored for the sensor data domain. Apart from extensive testing on more fall detection datasets, OOD detection and the proposed metrics can find applicability in other domains such as sports, mobile sensing, etc., where time-series data from sensors play a vital role. The research presented in the paper defines interesting OODs and successfully demonstrates that OOD detection can be used for HAR tasks with a simple yet effective method.

References

[1] Roy, D., Girdzijauskas, S. and Socolovschi, S., 2021. Confidence-calibrated human activity recognition. Sensors, 21(19), p.6566.

[2] Bao, L. and Intille, S.S., 2004, April. Activity recognition from user-annotated acceleration data. In International conference on pervasive computing (pp. 1-17). Springer, Berlin, Heidelberg.

[3] Kwapisz, J.R., Weiss, G.M. and Moore, S.A., 2011. Activity recognition using cell phone accelerometers. ACM SigKDD Explorations Newsletter, 12(2), pp.74-82.

[4] Maurer, U., Smailagic, A., Siewiorek, D.P. and Deisher, M., 2006, April. Activity recognition and monitoring using multiple sensors on different body positions. In International Workshop on Wearable and Implantable Body Sensor Networks (BSN'06) (pp. 4-pp). IEEE.

[5] Vail, D.L., Veloso, M.M. and Lafferty, J.D., 2007, May. Conditional random fields for activity recognition. In Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems (pp. 1-8).

[6] Gary M Weiss, Jessica L Timko, Catherine M Gallagher, Kenichi Yoneda, and Andrew J Schreiber. 2016. Smartwatch-based activity recognition: A machine learning approach. In 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE, 426–429

[7] Hammerla, N.Y., Halloran, S. and Plötz, T., 2016. Deep, convolutional, and recurrent models for human activity recognition using wearables. arXiv preprint arXiv:1604.08880.

[8] Ignatov, A., 2018. Real-time human activity recognition from accelerometer data using Convolutional Neural Networks. Applied Soft Computing, 62, pp.915-922.

[9] Song-Mi Lee, Sang Min Yoon, and Heeryon Cho. 2017. Human activity recog- nition from accelerometer data using Convolutional Neural Network. In 2017 ieee international conference on big data and smart computing (bigcomp). IEEE, 131–134.

[10] Ordóñez, F.J. and Roggen, D., 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors, 16(1), p.115.

[11] Casale, P., Pujol, O. and Radeva, P., 2011, June. Human activity recognition from accelerometer data using a wearable device. In Iberian conference on pattern recognition and image analysis (pp. 289-296). Springer, Berlin, Heidelberg.

[12] Zengtao Feng, Lingfei Mo, and Meng Li. 2015. A Random Forest-based ensemble method for activity recognition. In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 5074–5077.

[13] He, Z. and Jin, L., 2009, October. Activity recognition from acceleration data based on discrete consine transform and SVM. In 2009 IEEE International Conference on Systems, Man and Cybernetics (pp. 5041-5044). IEEE.

[14] He, Z.Y. and Jin, L.W., 2008, July. Activity recognition from acceleration data using AR model representation and SVM. In 2008 international conference on machine learning and cybernetics (Vol. 4, pp. 2245-2250). IEEE.

[15] Abdulmajid Murad and Jae-Young Pyun. 2017. Deep recurrent neural networks for human activity recognition. Sensors 17, 11 (2017), 2556.

[16] Wang, L., 2016. Recognition of human activities using continuous autoencoders with wearable sensors. Sensors, 16(2), p.189.

[17] Blundell, C., Cornebise, J., Kavukcuoglu, K. and Wierstra, D., 2015, June. Weight uncertainty in neural network. In International conference on machine learning (pp. 1613-1622). PMLR.

[18] Vyas, A., Jammalamadaka, N., Zhu, X., Das, D., Kaul, B. and Willke, T.L., 2018. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 550-564).
[19] Gal, Y. and Ghahramani, Z., 2016, June. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In international conference on machine learning (pp. 1050-1059). PMLR.
[20] Kingma, D.P. and Welling, M., 2014, April. Stochastic gradient VB and the variational auto-encoder. In Second International Conference on Learning Representations, ICLR (Vol. 19, p. 121).
[21] Lakshminarayanan, B., Pritzel, A. and Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems, 30.
[22] Liang, S., Li, Y. and Srikant, R., 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690.
[23] Malinin, A., Mlodozeniec, B. and Gales, M., 2019. Ensemble distribution distillation. arXiv preprint arXiv:1905.00076.
[24] Hendrycks, D. and Gimpel, K., 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136.
[25] Lee, K., Lee, K., Lee, H. and Shin, J., 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems, 31.
[26] Henry Friday Nweke, Ying Wah Teh, Ghulam Mujtaba, Uzoma Rita Alo, and Mohammed Ali Al-garadi. 2019. Multi-sensor fusion based on multiple classi- fier systems for human activity identification. Human-centric Computing and Information Sciences 9, 1 (2019), 34
[27] Hu, N., Englebienne, G., Lou, Z. and Kröse, B., 2016. Learning to recognize human activities using soft labels. IEEE transactions on pattern analysis and machine intelligence, 39(10), pp.1973-1984.
[28] Akbari, A. and Jafari, R., 2019, May. A deep learning assisted method for measuring uncertainty in activity recognition with wearable sensors. In 2019 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) (pp. 1-5). IEEE.
[29] Akbari, A. and Jafari, R., 2019, April. Transferring activity recognition models for new wearable sensors with deep generative domain adaptation. In Proceedings of the 18th International Conference on Information Processing in Sensor Networks (pp. 85-96).
[30] Kingma, D.P. and Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.
[31] Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A. and Vandergheynst, P., 2017. Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine, 34(4), pp.18-42.
[32] Dhaker, H., Ngom, P. and Mbodj, M., 2017. Overlap coefficients based on Kullback-Leibler divergence: exponential populations case. arXiv preprint arXiv:1704.02671.
[33] Weiss, G.M., Yoneda, K. and Hayajneh, T., 2019. Smartphone and smartwatch-based biometrics using activities of daily living. IEEE Access, 7, pp.133190-133202.
[34] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013. Bruges, Belgium 24-26 April 2013.
[35] Malekzadeh, M., Clegg, R.G., Cavallaro, A. and Haddadi, H., 2019, April. Mobile sensor data anonymization. In Proceedings of the international conference on internet of things design and implementation (pp. 49-58).
[36] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G. and Petersen, S., 2015. Human-level control through deep reinforcement learning. nature, 518(7540), pp.529-533.
[37] Kendall, A. and Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision?. Advances in neural information processing systems, 30.
[38] Bakar, U.A.B.U.A., Ghayvat, H., Hasanm, S.F. and Mukhopadhyay, S.C., 2016. Activity and anomaly detection in smart home: A survey. Next Generation Sensors and Systems, pp.191-220.
[39] Jakkula, V. and Cook, D.J., 2008. Anomaly detection using temporal data mining in a smart home environment. Methods of information in medicine, 47(01), pp.70-75.
[40] Kanev, Anton, et al. "Anomaly detection in wireless sensor network of the "smart home" system." 2017 20th Conference of Open Innovations Association (FRUCT). IEEE, 2017.
[41] Yamauchi, Masaaki, et al. "Anomaly detection for smart home based on user behavior." 2019 IEEE International Conference on Consumer Electronics (ICCE). IEEE, 2019.
[42] Yamauchi, M., Ohsita, Y., Murata, M., Ueda, K. and Kato, Y., 2020. Anomaly detection in smart home operation from user behaviors and home conditions. IEEE Transactions on Consumer Electronics, 66(2), pp.183-192.

APPENDIX

A. *Elderly Fall Detection*

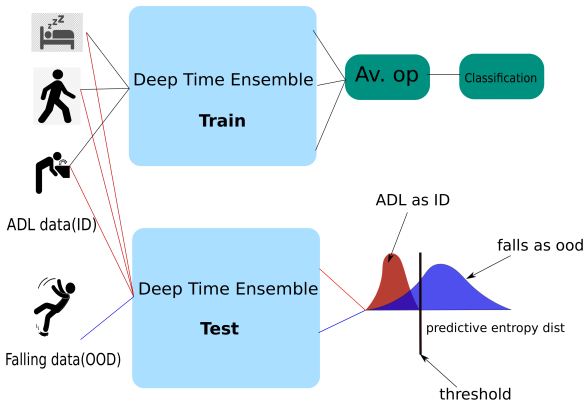

Fig. 7: Fall Detection Usecase

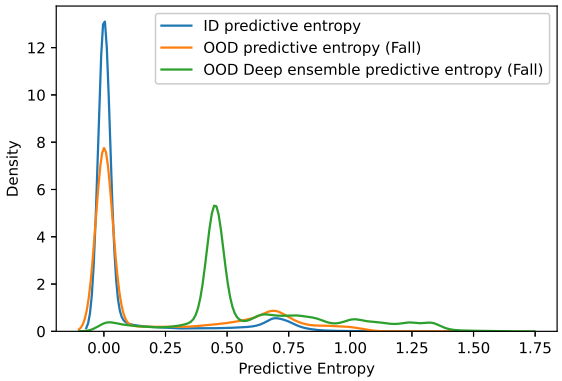

Fig. 8: Predictive Entropy Fall Detection Use-case

Elderly fall detection being a vital application in geriatric care is a use-case where the application of *Deep time-ensemble* showed promising outcomes with lesser data.

The traditional *fall detection* task involves training DL models on ADL activities along and simulated/real-falls. The problem with this approach is that to detect falls reliably; the model needs to learn the fall patterns from the training dataset. This calls for an extensive and well-curated fall dataset with examples of all the different types of falls. The highly stochastic nature of the *falling down* activity and the corresponding wearable signal pattern it can produce makes this problematic inspiring an exploration towards alternative directions.

TABLE III: Implementational details for results

Datasets	UCI	WISDM
Ensemble size	6	6
Timesteps	[128,100,80,70,60,50]	[200,180,160,140,120,100]
Train/test split	78/22	78/22
Convolution Filter	196	196
Filter Size	12	12
Dense Layer Size	1024	1024
Batch Size	256	256
Learning Rate	1e-4	1e-4
Optimizer	ADAM	ADAM
Dropout	0.15	0.15

Based on the primary proposition of this paper, a new fall detection system is proposed where the falls are detected as OOD. The process is explained in Figure 7. The proposed system is trained only with non-fall activities from a standard fall dataset using the *Deep time-ensemble* method. Thus when probed with *Falling* samples, it produces a higher predictive entropy at the output, detecting it as a *fall*. Furthermore, a threshold can be drawn based on the confidence of the predictions from the ensemble, that rejects samples below the threshold as probable falls. Not only this system allows to raise alarms and have a human-in-the-loop scenario for emergency geriatric care services but also it drastically reduces the amount of training data required to train a fall detection system. An experiment is performed using data from the *Sisfall* dataset that shows, using only 10 out of 30 subjects, and 6 out of 17 ADL types, a system that detects falls as OOD could be formulated.

The *Weitzman measure* metric decreases by a factor **3.75** using *Deep time-ensemble* compared to the *baseline* model. This behaviour can also be inferred by looking at Figure 8, where the clear separation of OOD and ID predictive entropy values are visible using the proposed training method. Also the *classification accuracy* is up by **0.6** for *Deep time-ensembles*.

The purpose of this experiment was to show that even with less data, a good result is achieved in the fall detection task. Better architecture selection, model composition and training focused towards *fall detection* would further improve the system in terms of OOD detection and classification, which is primarily is major work for the future.

B. Convolution Architecture and Implementation Details

In this section we present the convolution architecture in Figure 9 and the model hyper-parameters in Table III.

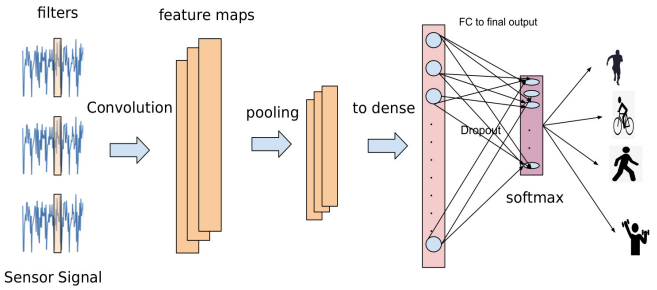

Fig. 9: Convolution Architecture used for modelling baseline and *DTE*