Learning Disentangled Behaviour Patterns for Wearable-based Human Activity Recognition

JIE SU, Newcastle University, United Kingdom
ZHENYU WEN, Zhejiang University of Technology, China
TAO LIN, EPFL, Switzerland
YU GUAN, Newcastle University, United Kingdom

In wearable-based human activity recognition (HAR) research, one of the major challenges is the large intra-class variability problem. The collected activity signal is often, if not always, coupled with noises or bias caused by personal, environmental, or other factors, making it difficult to learn effective features for HAR tasks, especially when with inadequate data. To address this issue, in this work, we proposed a Behaviour Pattern Disentanglement (BPD) framework, which can disentangle the behavior patterns from the irrelevant noises such as personal styles or environmental noises, etc. Based on a disentanglement network, we designed several loss functions and used an adversarial training strategy for optimization, which can disentangle activity signals from the irrelevant noises with the least dependency (between them) in the feature space. Our BPD framework is flexible, and it can be used on top of existing deep learning (DL) approaches for feature refinement. Extensive experiments were conducted on four public HAR datasets, and the promising results of our proposed BPD scheme suggest its flexibility and effectiveness. This is an open-source project, and the code can be found at http://github.com/Jie-su/BPD

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**.

Additional Key Words and Phrases: Wearable Sensing; Human Activity Recognition; Deep learning; Machine Learning

ACM Reference Format:
Jie Su, Zhenyu Wen, Tao Lin, and Yu Guan. 2022. Learning Disentangled Behaviour Patterns for Wearable-based Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 28 (March 2022), 19 pages. https://doi.org/10.1145/3517252

1 INTRODUCTION

Wearable-based HAR is one of the most popular themes in ubiquitous and wearable computing, and it plays a major role in a wide range of applications such as health assessment [8, 42], sleeps monitoring [57], sports coaching [23], etc. The main tasks of wearable-based HAR involve partitioning the multi-variate data stream from one or more sensors into segments and assigning a corresponding activity label to each segment [44].

Previous studies in this field leveraged the hand-crafted features in statistical (e.g., mean, variance) and frequency (e.g., power spectral density) domain to represent segments of raw sensory streams and projected the feature vector to the corresponding activity labels based on traditional machine learning methods such as SVM [23], KNN [42] and Random Forest [39]. However, designing effective features tends to be a trial-and-error process, and discriminant features may vary from task to task, making system-developing expensive and less

Authors' addresses: Jie Su, Newcastle University, Newcastle Upon Tyne, United Kingdom, jieamsu@gmail.com; Zhenyu Wen, Zhejiang University of Technology, Hangzhou, China, wenluke427@gmail.com; Tao Lin, EPFL, Lausanne, Switzerland, tao.lin@epfl.ch; Yu Guan, Newcastle University, Newcastle Upon Tyne (Corresponding Author), United Kingdom, yu.guan@newcastle.ac.uk.

2474-9567/2022/3-ART28 $15.00
https://doi.org/10.1145/3517252

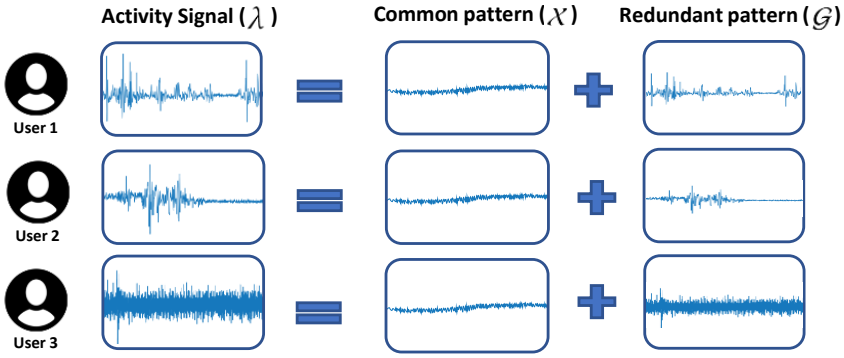


Fig. 1. Disentanglement factors in the feature representation for activity signal data: performing activity recognition over the disentangled features ($\mathcal{X}$) is much less challenging than that of raw sensor data (λ). The first column presents the raw sensor data (λ) for the standing activity across different user groups. The second and third columns indicate the disentangled common pattern ($\mathcal{X}$) and redundant/irrelevant patterns ($\mathcal{G}$) (e.g., gender, physical strength, etc.), respectively.

sustainable. To solve this problem, recent studies [13, 15, 35, 37, 56] leveraged the exceptional data representation ability of deep learning methods to expedite feature extraction. Such studies mainly utilized the deep neural networks (e.g., Convolutions neural networks(CNN) [28], Long-Short Term Memory(LSTM) [18]) to extract the features from the original input sensors in an end-to-end manner.

Although the deep learning approaches can extract decent representation from input sensor data, they may face challenges when dealing with multi-model sensor streams from diverse subjects/users. One of the crucial challenges is the intra-class variability problem. The discrepancy between subjects in performing activities was neglected in current studies - they usually map all subjects indiscriminately to the high-level feature representations with Deep Learning methods. However, the sensor record for the same activity may vary among different people due to their personal characteristics, such as gender, habits, physical strength, etc. Figure 1 shows the sensor reading of the standing activity $\lambda = \mathcal{X} + \mathcal{G}$. $\mathcal{X}$ represents standing activity which is the common pattern (distribution) across various groups of users. $\mathcal{G}$ are the *redundant* patterns which vary between different group of users, even each independent user. Thus, such redundant patterns bring challenges for developing a robust activity recognition system, serving million or billion of users. To overcome this, we can collect the new sensor data and retrain the model to increase the generalisation ability. However, this solution is time-consuming and it is very costly to label the new sensor data. Alternatively, removing or disentangling the redundant patterns from the sensor data can significantly improve robustness and generalisation of activity recognition system.

Recently, learning disentangled representations has attracted a lot of attention from the machine learning community. Such representations provide many advantages: improving the predictive performance on downstream tasks [31, 32], reducing the sample complexity [48, 53], offering interpretability [17], improving fairness [30] and have been identified as a way to overcome *short cut* learning in deep learning [9]. From the literature, disentanglement learning approaches proved to be effective in the computer vision field. However, applying disentanglement to sensor data (e.g., human activity recognition) is more challenging since the disentanglement should consider both context level (e.g., time dimension) information and feature level information.

To address this issue, in this work we proposed a Behaviour Pattern Disentanglement (BPD) scheme, which utilizes disentanglers to induce two groups of representations. Ideally, the activity features are captured as the common patterns on a certain class of activity, and the redundant representations are captured as the unpredictable personal patterns such as the lifestyle of a person. To effectively disentangle two groups of features, we develop an

adversarial disentangle mechanism. By using such mechanism, the generated activity feature representations are expected to be more invariant to other domains, compared to the original data. Moreover, our BPD framework is flexible, and is applicable on top of existing popular DL approaches, such as CNN [56], DeepConvLSTM [37], etc. for activity feature refinement. To evaluate our models, we leveraged the Leave-one-subject-out cross validation (LOSO-CV) protocol on the four public HAR datasets, which can demonstrate the performance at both overall and the subject-level. Our contributions can be summarised as follows:

- We proposed the BPD framework, which can separate the activity signal from the redundant feature in the feature space with the least dependency.
- Our BPD scheme is flexible, and it can be used on top of existing DL approaches for feature refinement, with improved HAR results. Our project is open source and the code can be found at http://github.com/Jie-su/BPD
- Extensive experiments were conducted, and we studied our BPD framework in details. The promising results suggested its effectiveness.

The rest of this paper is organized as follow. Section 2 introduces the related background knowledge. Section 3&4 present the problem definition and the details of the proposed BPD framework. Section 5 gives the experimental settings as well as evaluation results, and Section 6 concludes.

2 BACKGROUND

HAR has a long-standing history in the wider ubiquitous and wearable computing community. Recently, a multitude of methods have been proposed and facilitate a variety of applications. HAR has become one of the pillars of the third generation of computing [51]. In the following section, we will review the specific background for this paper, which spans three main subject areas: i) Deep learning for HAR in ubiquitous and wearable computing; ii) Adversarial Learning; iii) Representation Disentangle Learning.

2.1 Human Activity Recognition

Traditional machine learning[43] approaches such as K-Nearest Neighbor (KNN), Hidden Markov Model (HMM), Support Vector Machine (SVM), Random Forest (RF), and Naive Bayes (NB) have been successfully applied on HAR [2, 14, 27]. The main drawback of these models is that they are mainly relying on hand-crafted features or heuristic information.

Deep learning methods can automatically extract features from raw signals, reducing the efforts on feature engineering procedures. One of the most popular deep learning model is convolutional neural network (CNN), which can extract the HAR representation by stacking multiple convolutional layers [56]. DeepConvLSTM [37] extended CNN by adding LSTM layers for temporal information modelling. In [16], Hammerla *et al.* comprehensively studied the performance of DNNs, CNNs and RNNs for HAR tasks. Guan and Ploetz [13] explored sample-wise activity recognition by ensembles of deep LSTM learners using an epoch-wise bagging scheme. Murahari and Ploetz [35] added the attention layers to the DeepConvLSTM model to learn local temporal context from raw sensor data. Recent work [45] proposed a DDNN model to learn statistical, temporal and spatial correlation features from signals, before a final fusion for performing the activity recognition.

2.2 Adversarial Learning

As deep neural networks have found their way from labs to the real world, the security and integrity of the applications pose a great concern. Adversaries can craftily manipulate legitimate inputs, which may be imperceptible to the human eye, but can force a trained model to produce incorrect outputs [6]. Szegedy *et al.* [52] first discovered that well-trained deep neural networks were susceptible to adversarial attacks. Attacks on autonomous vehicles have been demonstrated by Kurakin *et al.* where the adversary manipulated traffic signs successfully confuse the learning model.

The Generative Adversarial Networks (GANs) were proposed by Goodfellow *et al.* [11] which brought the concept of adversarial to the network level. More precisely, the key idea of GANs is to create a competition between the generative model and an adversary: a discriminator model that learns to determine whether a sample is from the model distribution or the data distribution [11]. Imagining the generative model as a team of counterfeiters trying to produce fake currency that is non-detectable, while the discriminator as the police trying to detect the counterfeit currency. Such competition drives both teams to improve their intelligence until the counterfeits are indistinguishable from the genuine currency. Benefiting from GANs, the concept of adversarial learning has become a popular research topic in the deep learning community and is applied to many applications such as adversarial sample generation [55], style transfer [22], and autopilot [58].

2.3 Representation Disentangle Learning

Prior to the deep learning era, most computer vision systems made use of features that were hand-engineered and task-oriented. One of the desired goals and challenges for these features was to be invariant to certain nuisance/redundant factors in the data such as affine transforms, blur, etc. Early studies such as Gopalan *et al.* [12] and Lowe [33] have achieved it, but the drawback of these methods is that they are mainly relying on hand-crafted features. Recent advanced deep learning techniques are primarily data-driven where features are learned by adding suitable constraints on the learning paradigm. Being dependent on data enables those methods to learn covariate factors in the data (e.g., angle, shape, the noise in the data generating process). It should be noted that the 'noise' can be any undesired and unknown factors of original data which we parameterise with a mathematical model. Figure 1 illustrates the concept of covariate factors that might exist in the feature representation of multi-dimensional signals (i.e., time-series signal). Specifically, it illustrates common activity factor and personal/environmental factors for time-series data.

Disentangling those factors can help us to further explore the highly entangled high dimensional data, but the factors might become the bias/noise to the recognition system. There are a few common kinds of nuisance/noise factors that creep into the sensor signal datasets which can be used for training recognition systems: gender, physical strength variation, age, etc. Here, the nuisance/noise factors can be defined as 'task-based undesired factors of variations' since certain factors of variation are desirable for some tasks while not undesirable for others. For example, gender could be a noise factor when doing activity classification but could be a key factor for conditional signal sample generation (i.e., generate signal samples with gender condition). Thus, exploring the disentanglement and disentangling desired representation/factors is crucial for many downstream machine learning applications.

Benefiting from the advance of DL techniques, recent works [19–21, 29, 34, 36, 54] in computer vision started to learn the interpretable representations from images or videos by utilising generative adversarial networks. (GANs) [11] or Variational autoencoders(VAEs) [25]. InfoGAN [7] was proposed to learn the disentangled representation in an unsupervised manner while it may suffer from training instabilities. Beta-VAE [17] improved the poor disentanglement/reconstruction trade-off of the original VAEs. Later, Liu et al. [29] introduced a unified feature disentanglement framework to learn domain invariant features across different domains. Recently, disentangled representation learning has also been applied to some popular applications (e.g., Gait recognition [20], speaker recognition [50]). Hu et al. [20] proposed a disentanglement framework that can separate gait identity from the camera view for view-invariant gait recognition. DEAAN [50] was proposed to disentangle speaker-related features from speech signals to achieve robust speaker adaptation and recognition. Recently, GILE [46] was proposed by Qian et al. to disentangle ID information from raw sensor data. They utilised the additional subject ID label and the Independence Excitation mechanism to disentangle the id information and activity information. However, their design requires specific network design (complex network structure) and the additional meta information, which might be less practical.

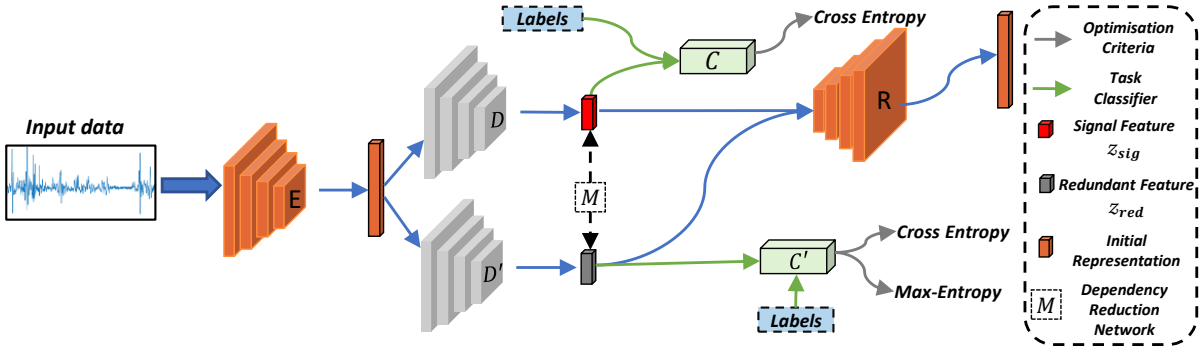


Fig. 2. Structure of our proposed BPD framework, where (E), (D, D'), (C, C'), (R), (M) represent Encoder, Disentanglers, Classifiers, Reconstructor, Dependency Reduction Network, respectively.

3 PROBLEM DEFINITION

Recognising human activity with multi-modal data involves multiple devices attached to different parts of the human body. Each device carries multiple sensors (e.g., 3-axis accelerometer, gyroscope, magnetometer). Following the standard HAR procedure [5], we divide the multi-variate sensory streams into segments with a fixed-size sliding window (detailed size information will be listed on experiment setting). Finally, given segment training data $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$ where N is the training sample number; $\mathbf{x}_i$ and y_i are the ith training example and label with $y_i \in [1, K]$ and K is the class number, the purpose of HAR is to learn a function $\mathbb{F}(x, \beta)$ to infer the correct activity label for the given segment data, where β represents all the parameters to be learned during the training process.

4 METHODOLOGY

For HAR, the collected sensor data often, if not always, includes other redundant features which can be subjects' personal style, gender, age, weight, etc. Such redundant features or factors may cause large intra-activity variability, making it challenging to learn discriminant behaviour patterns, especially when with inadequate data. In order to remove or separate the redundant features from the activity signal, we introduce the Behaviour Pattern Disentanglement (BPD) scheme, based on which the activity signal can be separated from the redundancy features in the latent feature space.

Our proposed BPD scheme includes two key components as illustrated in Figure 2: (i) *A Signal & Redundant feature disentanglement network*, which learns to disentangle the input features into activity signal and redundant features; (ii) *A Dependency Reduction Networks*, which aims to reduce the correlations between the activity signal and redundant features. These two blocks together with a feature reconstruction module maximises the effects of feature disentanglement while ensuring the minimal information loss. The whole framework can be trained in an end-to-end manner, and the disentanglement networks and the activity signal classifier will be used during inference (i.e., activity classification).

In the following subsections, we will elaborate our design choices on Signal & Redundant Feature Disentanglement (Section 4.1), the Signal & Redundant Feature Dependency Reduction (Section 4.2) and the adversarial training and optimisation process (Section 4.3).

4.1 Signal & Redundant Feature Disentanglement

The activity pattern of different users is often associated with users' own personal information, like gender, age and other factors. However, these personalised attributes hinder the representation learning for the later classification. To this end, we propose to utilise a more general and robust feature representation with less irrelevant user information to alleviate the classification difficulty across different users, so as to further improve the generalisation ability of the HAR models. In the following section, we refer to the aforementioned robust feature representation and irrelevant user information as "activity features" $\mathbf{z}_{sig}$ and "redundant features" $\mathbf{z}_{red}$ respectively.

We leverage the idea from the generative adversarial network and introduce the concept of *adversarial disentanglement* to remove/disentangle the redundant features from the activity features. We feed the learnt features to disentanglers D and D' to decompose the features representation retrieved from the encoder E. The representation extracted from disentangler D will be supervised by the corresponding activity labels in the classifier C to ensure the activity classification ability, while a two-player game happens in the classifier C' and its' adversary disentangler D' so as to generate irrelevant representations. Note that the classifier C' aims to map the representations to the correct activity labels while the D' on the contrary generates the irrelevant representations to fool the classifier C'. To guard the representation integrity of the disentangled features from D and D', we add a feature reconstructor R to recover the initial feature representation.

The proposed BPD framework trains the aforementioned encoder (E), disentanglers (D and D') and classifiers (C and C') in an alternative way. More precisely, two disentanglers D and D' along with two K-way classifiers (i.e., C and C') will be first trained by minimising the cross-entropy loss in Eq. (1):

$$\underset{\theta_{E,D,D',C,C'}}{\mathcal{L}_{ce}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} 1[y_i = k]\log(C(\mathbf{z}_{sig}^i)) - \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} 1[y_i = k]\log(C'(\mathbf{z}_{red}^i))\,. \tag{1}$$

where $\mathbf{z}_{sig}^i = D(E(\mathbf{x}_i))$ and $\mathbf{z}_{red}^i = D'(E(\mathbf{x}_i))$.

Then, to produce features with less discriminative information, or equivalently, increasing the uncertainty of classification, we minimise the negative entropy on the feature extracted by disentangler D'. The negative entropy function can be written as:

$$\underset{\theta_{E,D'}}{\mathcal{L}_{ne}} = -\frac{1}{N}\sum_{i=1}^{N}\log C'(\mathbf{z}_{red}^i)\,, \tag{2}$$

and the parameter of classifiers is fixed when optimising Eq. (2).

To ensure the integrity of the feature representation, that is, the disentangled activity and redundant features can be reconstructed back to the initial representation, we forward the features $\mathbf{z}_{sig}$ and $\mathbf{z}_{red}$ that are extracted from disentanglers to the reconstructor R and optimise it simultaneously with the negative entropy minimisation (Eq. (2)). To achieve the effects of reconstruction, we utilise the L2 loss function to constrain the equivalence between reconstructed features and initial features. Such constraint can be written as:

$$\underset{\theta_{D,D',R}}{\mathcal{L}_{recon}} = \frac{1}{N}\sum_{i=1}^{N}\left\| E(\mathbf{x}_i) - R(\mathbf{z}_{sig}^i, \mathbf{z}_{red}^i)\right\|\,, \tag{3}$$

where $E(x_i)$ is the initial representation. Note, the activity feature $\mathbf{z}_{sig}$ and redundant feature $\mathbf{z}_{red}$ will be fed into the reconstructor with a concatenation operation, which means only one concatenated feature will be forwarded into reconstructor.

4.2 Signal & Redundant Feature Dependency Reduction

The previous section presents the disentanglement scheme of the activity and redundant features. To ensure less correlated/dependent disentangled features for a good disentanglement learning, we leverage the mutual information—a measure of non-linear dependencies between variables (e.g. learned feature representation) [26, 38, 49]—to reduce the mutual information between the activity signal $\mathbf{z}_{sig}$ and redundant features $\mathbf{z}_{red}$, so as to reduce the dependence of between them.

Here, the mutual information between the activity features $\mathbf{z}_{sig}$ and redundant features $\mathbf{z}_{red}$ can be defined as:

$$I(\mathbf{z}_{sig}, \mathbf{z}_{red}) = \int_{\mathbf{z}_{sig}} \int_{\mathbf{z}_{red}} \log \frac{P(\mathbf{z}_{sig}, \mathbf{z}_{red})}{P(\mathbf{z}_{sig})P(\mathbf{z}_{red})} d\mathbf{z}_{sig} d\mathbf{z}_{red}\,, \tag{4}$$

where $P(\mathbf{z}_{sig}, \mathbf{z}_{red})$ is the joint probability density distribution (pdf); $P(\mathbf{z}_{sig})$ and $P(\mathbf{z}_{red})$ are marginal pdfs. $I(\mathbf{z}_{sig}, \mathbf{z}_{red})$ measures the dependency between the two features, and minimising $I(\mathbf{z}_{sig}, \mathbf{z}_{red})$ may further push these two features apart in the feature space.

Despite being a pivotal measure across different domains, the mutual information is only tractable for discrete variables, or for a limited family of problems where the probability distributions are unknown [4]. Following [4], we adopt Mutual Information Neural Estimator (MINE) as an unbiased estimation of mutual information on *i.i.d* samples through a neural network M_θ (as shown in Figure 2). Specifically, we leverage the lower-bound calculation from [4] to formulate the loss function as follows:

$$\underset{\theta_{D,D',M}}{\mathcal{L}_{MINE}} = I(\mathbf{z}_{sig}, \mathbf{z}_{red}) = \frac{1}{n}\sum_{i=1}^{n} M(\mathbf{z}^i_{sig}, \mathbf{z}^i_{red}, \theta) - log(\frac{1}{n}\sum_{i=1}^{n} e^{M(\mathbf{z}^i_{sig}, \hat{\mathbf{z}}^i_{red}, \theta)})\,, \tag{5}$$

where $\{\mathbf{z}^i_{sig}, \mathbf{z}^i_{red}\}^n_{i=1}$ are n pairs sampled from the joint distribution $P(\mathbf{z}_{sig}, \mathbf{z}_{red})$; $\hat{\mathbf{z}}^i_{red}$ is sampled from the marginal distribution $P(\mathbf{z}_{red})$, and $M(\mathbf{z}^i_{sig}, \mathbf{z}^i_{red}, \theta)$ is a neural network parameterised by θ to estimate the mutual information between two distributions. Model parameters of the disentanglers (D and D') as well as the dependency reduction network M will be updated by minimising Eq.(5).

4.3 Algorithm & Implementation

For our BPD structure, we implement the components as follows: 1) Encoder (E): CNN[56] or DeepConvLSTM[37] with the same network structure as original works; 2) Disentanglers (D and D'): Single fully-connected layer with a batch normalisation layer; 3) Reconstructor (R): single fully-connected layer. 4) Dependency Reduction Network (M): two fully-connected layers. 5) Classifiers (C and C'): two fully-connected layers and a dropout function. To ensure the learned consistent representation space can lead to accurate activity classification, we jointly minimise the activity classification error (Classifier C) and conduct adversarial training.

The training strategy of our BPD framework is detailed in Algorithm 1. More precisely, on the training stage, given data from training users, with selected Encoder E (e.g., CNN, ConvLSTM), the BPD framework is optimised in an end-to-end manner using Algorithm 1. At the inference stage, given trained components E (CNN or ConvLSTM), C (Classifier), D (Disentangler) from Algorithm 1, for any query data $\mathbf{x}$ from any unseen user, classification can be performed by inputting the disentangled activity features to the classifier C. Specifically, the label is assigned to $\hat{y}$ such that:

$$\hat{y} = \arg\max_{[1,K]} C(\mathbf{z}_{sig}), \quad \text{where} \quad \mathbf{z}_{sig} = D(E(\mathbf{x})). \tag{6}$$

Algorithm 1: Training Strategy of the BPD framework

Input: Training data $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$, Encoder E, Disentanglers (D and D'); Classifiers (C and C'); Dependency Reduction Network M, and Reconstructor R;
Result: Trained Encoder E, trained disentangler D and trained classifier C
Initialisation;
for *j=1 : maxEpoch* **do**
  if *not converged* **then**
    Sampling training mini-batch data from $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$;
    Signal&**Noise Disentanglement:**
    Updating E, D, D', C, C' by minimising Eq. (1);
    Updating E, D' by minimising Eq. (2);
    Signal&**Noise Dependency Reduction:**
    Updating D, D', M by minimising Eq. (5) ;
    Reconstruction:
    Updating D, D', R by minimising Eq. (3) ;
    $j = j + 1$;
  else
    break;
  end
end
return Trained E; trained C; trained D

5 EXPERIMENT

5.1 Datasets

To evaluate the effectiveness of our BPD framework, we perform it on four public datasets: PAMAP2 [47], MHEALTH [1], DSADS [3] and GOTOV [40].

Table 1. Description of the four public HAR datasets used in our study; #Dim represents the dimension of the input data.

Dataset	#Subject	#Activity	Frequency	#Sample	#Dim	Wearing Position
PAMAP2	8	12	100Hz	2.84M	52	Wrist,Chest,Ankle
MHEALTH	10	12	50Hz	0.34M	23	Chest,Ankle,Arm
DSADS	8	19	25Hz	1.14M	45	Tarso, Right/Left Arm, Right/Left Leg
GOTOV	35	16	83HZ	5.9M	3	Wrist

Physical Activity Monitoring (PAMAP2) [47] dataset includes data recorded from 9 subjects performing 18 different activities, such as vacuum cleaning, ironing, rope jumping, etc. The data were collected with three IMUs placed on the subject's chest, dominant wrist, and dominant ankle, respectively. In our study, 12 activities were selected (as shown in Fig. 3), and all the IMU data channels (i.e., 52 dimensions) from 8 subjects were used.

Mobile Health (MHEALTH) [1] dataset contains body motion and vital signs recording for 10 subjects of diverse profiles while performing 12 activities in an out-of-lab environment with no constraints. The total dimension of the input data is 23, which include the data recorded by inertial measurement units (IMUs) that are placed on the subject's chest, right wrist and left ankle. The IMUs collect a 3-axis acceleration, a 3-axis gyroscope and a 3-axis

magnetic field of motion, respectively. Also, the IMUs positioned on the chest provides 2-lead ECG measurement, which can be used for basic heart monitoring.

Daily and Sports Activities Data Set (DSADS) [3] dataset contains motion sensor data of 19 daily and sports activities performed by 8 subjects. Each activity was performed for 5 minutes in their style without constraints. 5 IMUs were positioned on the torso, right arm, left arm, right leg and left leg with 9 sensors on each unit (3-axis accelerometers, 3-axis gyroscopes, and 3-axis magnetometers) which produces 45-dimensional sensor data.

Growing Old Together Validation (GOTOV) [40] dataset contains 16 daily activities sensor data of 35 elder participants (21 male and 14 female). The inertial measurement units were placed on the subject's ankle, chest, and wrist to collect 3-axis acceleration data.

For DSADS dataset, we used all subjects' data. For PAMAP2, we removed 6 activities (i.e., Watching TV, Computer work, Car driving, Folding laundry, House Cleaning, and Playing soccer), as they were only performed by one subject, which was also removed in our study. For MHEALTH dataset, we used all subjects' data. For the GOTOV dataset, since there are some missing channels in the sensors attached to ankle/chest, only wrist-worn accelerometer data (i.e., with 3 dimensions) were used in our study. Details of these used datasets can be found in Table 1.

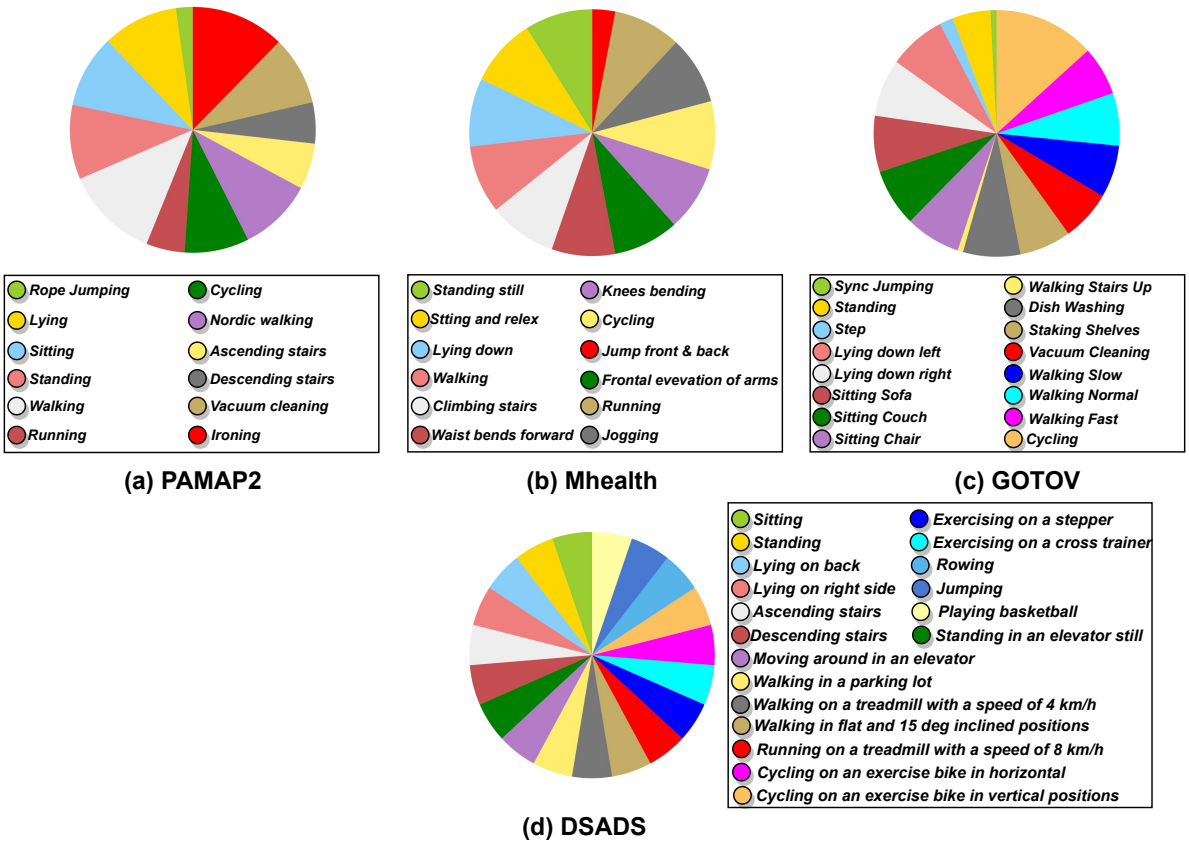


Fig. 3. Activity distribution of the four datasets in our study (best viewed in colour)

Figure 3 presents the activity distribution for the four datasets. In terms of activity classes, DSADS dataset is more balanced than the other three. Since the unbalanced class distribution might affect the performance of the algorithms, we further report the class-wise performance.

5.2 Experimental Settings

Data Pre-processing. In our study, we divided the raw sensory data streams into small data segments with a fixed-sized sliding window (168 samples) with an overlap of 50%. Since the sampling rates for the four datasets are different, it yields window lengths of 1.68, 3.36, 6.72, 2.02 seconds for the PAMAP2, MHEALTH, DSADS, and GOTOV datasets, respectively. These windows/segments can be fed into the network directly without any hand-crafted feature engineering or transformation.

Baseline Models. We compared our proposed BPD framework with the closely related baselines. CNN[56] and DeepConvLSTM[37] were the state-of-the-art feature learning approaches for human activity recognition; beta-VAE[17] was a conventional disentangle learning framework in the computer vision field and GILE [46] was a recent method aiming at disentangling the identity information from raw data streams, yet subject identity label is required by GILE as an extra information. For all baseline methods, we used the released code if available, and reproduced the unavailable methods using Pytorch[41].

Training Setting. Our network parameters were initialised by Xavier Normal [10] and optimised by Adam Optimiser [24] with a learning rate of 0.0001 for four datasets. Due to the computational limitation, we set the training batch size to 64 with 300 maximum training epoch. For the latent representation dimension (i.e., feature dimension of activity code $\mathbf{z}_{sig}$/redundant code $\mathbf{z}_{red}$), we empirically set 592 to encoder CNN and 32 to DeepConvLSTM respectively. All algorithms were implemented by Pytorch and running on NVIDIA RTX 3090 GPU.

Evaluation Protocol. Initially, we used the Leave-One-Subject-Out Cross Validation (LOSO-CV) strategy for each dataset to evaluate the models' performance. For each dataset, both the overall performance and the subject-level performance were reported. Moreover, we conducted the ablation study on the large GOTOV dataset (35 subjects in total) using the hold-out validation, where the training set included 28 subjects while the test set included the other 7 subjects.

Evaluation Metric. To measure the performance of our proposed BPD framework, we used the mean F1 score as the evaluation metric, which is widely used in the human activity recognition literature [13, 37]. Moreover, we also reported the class-wise F1-score for the four datasets to investigate the effect of redundant information on the class level (i.e., to see which class benefits more when removing the irrelevant information) for a better understanding of our BPD framework.

5.3 Result on Four public HAR Datasets

We evaluated our BPD framework on the four public HAR datasets. In Table 2, 3, 4, the mean F1-scores of the baseline models as well as our BPD framework (based on two encoders, i.e., CNN and DeepConvLSTM) were reported in both subject-level and overall average for datasets PAMAP2, MHEALTH, and DSADS. We can see the superior performance improvements in all settings brought by the proposed BPD framework, irrespective of the encoders. We also compared two disentanglement learning baselines, namely, beta-VAE[17] and GILE [46]. From these tables, we can see both methods yield lower performance even than the baselines without disentanglement. Since beta-VAE was an approach borrowed from the computer vision field, it may not generalise well to HAR tasks. On the other hand, the low performance of the GILE might be due to the enforcement of mapping features into subject-identifiable-level which may be unnecessary and hard to train. We grid-searched the hyper-parameters

Table 2. Mean F1-score for each subject on the PAMAP2 dataset (in leave-one-subject-out CV setting)

Dataset	Subject	CNN	DeepConvLSTM	beta-VAE	GILE	BPD (CNN)	BPD (DeepConvLSTM)
PAMAP2	1	0.6539	0.6340	0.5970	0.6032	0.6826	**0.6915**
	2	0.7563	0.7363	0.7084	0.7181	**0.8719**	0.8381
	3	0.8099	0.7154	0.5713	0.6950	**0.8262**	0.8117
	4	0.8044	0.8183	0.7115	0.7542	**0.8307**	0.8244
	5	0.8886	0.8588	0.7391	0.8181	**0.8900**	0.8675
	6	0.8791	0.7924	0.7175	0.7723	**0.8827**	0.8281
	7	0.9243	0.9100	0.8376	0.8932	**0.9311**	0.9101
	8	0.3952	0.4495	0.3814	0.3521	0.4011	**0.4921**
	Avg.	0.7640	0.7393	0.6580	0.7008	**0.7895**	0.7829

Table 3. Mean F1-score for each subject on the Mhealth dataset (in leave-one-subject-out CV setting)

Dataset	Subject	CNN	DeepConvLSTM	beta-VAE	GILE	BPD (CNN)	BPD (DeepConvLSTM)
MHEALTH	1	0.9514	0.9074	0.7866	0.8452	**0.9575**	0.9554
	2	0.8530	0.8760	0.8195	0.8135	**0.9348**	0.9085
	3	0.8441	0.8688	0.7091	0.8322	0.8659	**0.8711**
	4	0.9351	0.9112	0.8101	0.8810	0.9510	**0.9573**
	5	0.8781	0.8583	0.7752	0.8150	**0.9904**	0.9804
	6	0.9849	0.9241	0.8753	0.8832	**0.9934**	0.9766
	7	0.9793	0.9735	0.6985	0.8923	**0.9965**	0.9760
	8	0.9685	0.9566	0.8706	0.8842	**0.9848**	0.9775
	9	0.9894	0.9812	0.9113	0.9237	**0.9913**	0.9869
	10	0.9829	0.9383	0.8737	0.9224	**0.9943**	0.9865
	Avg.	0.9367	0.9195	0.8130	0.8693	**0.9660**	0.9576

of GILE for the best results, yet the results were less promising when compared with our approach. Moreover, GILE requires human identity labels, which can be less flexible than ours.

In Table 2, we also noticed the low results from subject 8. Although the BPD scheme can refine the activity feature and improve the results substantially (about $2-4\%$), they are still far from satisfactory. One major reason can be the limited number of subjects for training. In the LOSO-CV setting, only 7 subjects were used for training, and the trained model may not generalise well to unseen subjects that are very different from the (small) population.

On the MHEALTH and DSADS datasets, although with different activity types, from Table 3, and Table 4, we can see more significant results: 1) BPD can boost the performance much further, irrespective of the encoder. 2) when compared with other disentanglement learning baselines (beta-VAE, GILE), our BPD yields much higher results.

Table 4. Mean F1-score for each subject on the DSADS dataset (in leave-one-subject-out CV setting)

Dataset	Subject	CNN	DeepConvLSTM	beta-VAE	GILE	BPD (CNN)	BPD (DeepConvLSTM)
DSADS	1	0.7354	0.7316	0.6135	0.7467	**0.7599**	0.7267
	2	0.7661	0.7804	0.6545	0.7746	**0.8919**	0.8850
	3	0.7468	0.8327	0.5858	0.7305	0.8755	**0.8947**
	4	0.6734	0.6551	0.5012	0.6579	0.6626	**0.6756**
	5	0.6472	0.6504	0.4915	0.7228	**0.7894**	0.7661
	6	0.8023	0.9027	0.5225	0.8185	**0.9551**	0.9302
	7	0.7242	0.7341	0.5724	0.6363	**0.8615**	0.8537
	8	0.6139	0.6255	0.5584	0.5875	**0.7498**	0.7247
	Avg.	0.7136	0.7390	0.5625	0.7094	**0.8182**	0.8070

Compared with PAMAP2, MHEALTH, and DSADS datasets, GOTOV is a much larger dataset with 35 subjects, based on which we conducted LOSO-CV and reported the corresponding results in Table 5 on appendix A. We can observe that the results can benefit from our BPD scheme, with more significant improvement on the CNN encoder than the DeepConvLSTM encoder. Specifically, as shown in Table 5, BPD(CNN) and BPD(DeepConvLSTM) can yield about 3.08%, and 2% performance gain (in terms of overall average), respectively.

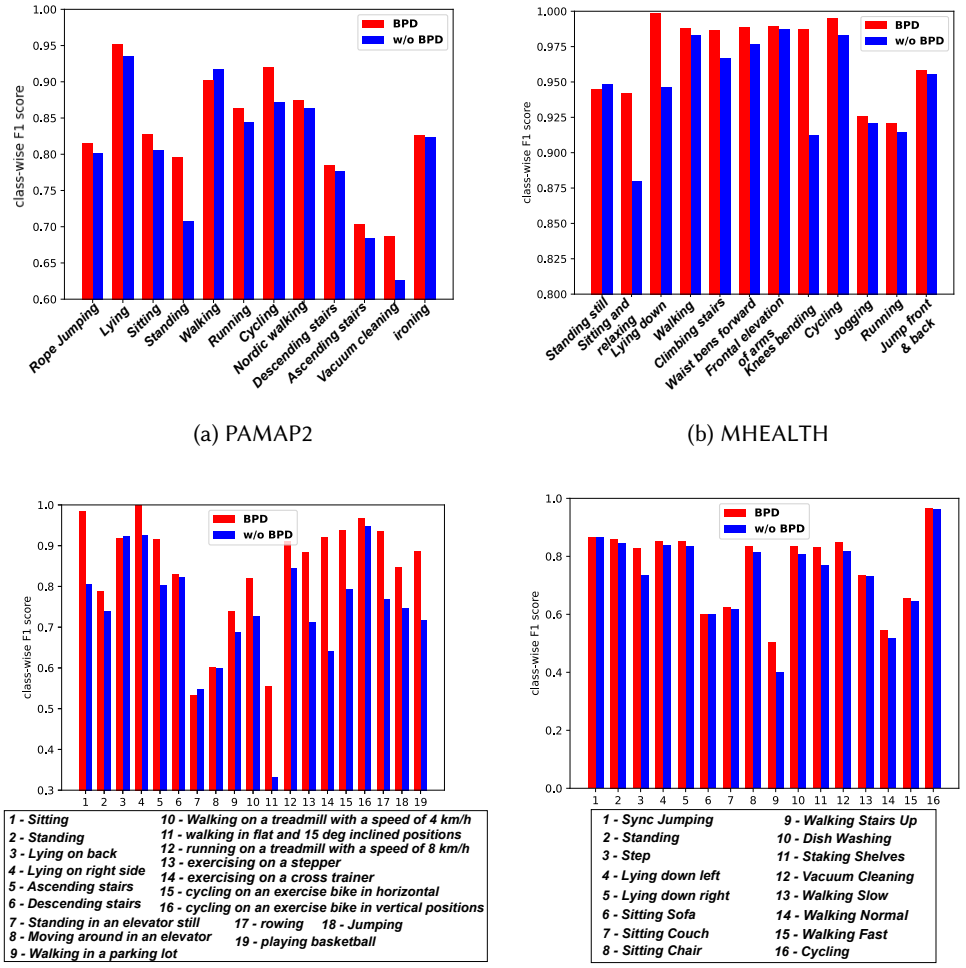
5.4 Class-wise Analysis

Previous experimental results present the performance in both overall and subject-level. To get more insight into our BPD scheme, it is also crucial to show the class-wise or activity-wise results to see how BPD can disentangle redundancy from different activities.

We conducted the experiments using CNN and BPD(CNN) on the four datasets, and reported the class-wise results in Figure 4. We can see the general performance gains (by using BPD), yet they vary at the activity level. For the PAMAP2 dataset, substantial improvements are from the following activities: “Vacuum cleaning”, “Standing”, “Cycling”, “Ascending stairs”. It is quite interesting to see that in the MHEALTH, DSADS and GOTOV datasets (Figure 4b,4c, 4d), the “Ascending stairs” activity (or analogously, “Climbing Stairs” on MHEALTH dataset and “Walking Stairs Up” on GOTOV dataset) were also the activities that benefit significantly from the BPD scheme, indicating these activities may be easily affected by personal/environmental factors. Similarly, some strenuous activities such as “Cycling”, ‘Vacuum cleaning” , “Knees bending”, “Exercising”, and “ Rowing” are more likely to be affected by physical factors such as vital capacity so that might cause large variance for different subjects. On the contrary, the activities with less energy consumption (e.g., “Lying” activities) will benefit less from the BPD framework since the patterns for those activities tend to be less personal. It is interesting to see that the “Standing” activity on PAMAP2, and “Sitting” activity on MHEALTH and DSADS dataset gain large improvement. A possible explanation for that is these activities may be affected by the personal habit (i.e., sitting/standing posture).

5.5 Ablation Study

To study the effectiveness of the major components in the BPD framework, we conducted ablation studies. Due to the computational limitation, we used the GOTOV dataset with a hold out setting. The training group contains 28 subjects (17 males and 11 females) while the testing group contains 7 subjects (4 males and 3 females).



(a) PAMAP2 (b) MHEALTH

(c) DSADS (d) GOTOV

Fig. 4. Class-wise F1-score of CNN with and w/o the proposed BPD scheme on the four datasets

With encoders (CNN[56] and DeepConvLSTM[37]), we studied the baseline (i.e., w/o BPD), the signal & redundancy disentanglement component (i.e., BPD w/o dependency reduction), and the proposed full BPD. In addition, we also studied the contribution of reconstructor (which can keep the integrity of the feature representation) by removing it from BPD (i.e., BPD w/o reconstructor).

Figure 5 reports the detailed results of the ablation studies, and we can see that each component in the BPD framework contributes positively to the final results, and the encoder CNN benefits more from the BPD framework than DeepConvLSTM. For encoder CNN, the application of the signal & redundancy disentanglement component (i.e., BPD w/o dependency reduction) can achieve about 3% performance improvement, in contrast to only 1% for DeepConvLSTM. For both encoders, performing the dependency reduction mechanism can further push

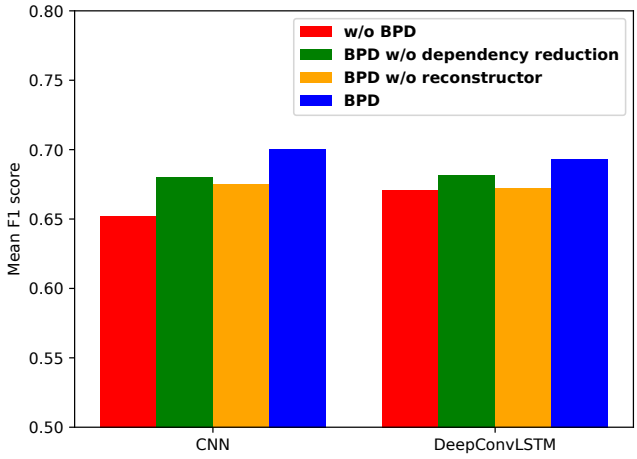


Fig. 5. Mean F1-score of the ablation study on GOTOV datasets

signal & redundancy features apart, with further performance gains (i.e., BPD in Fig. 5). We can also observe that for both encoders, the performance of the BPD drop substantially (about 2% in mean F1 score) without using reconstructor (i.e., BPD w/o reconstructor in Fig.5), indicating the importance of keeping the integrity of the feature representation.

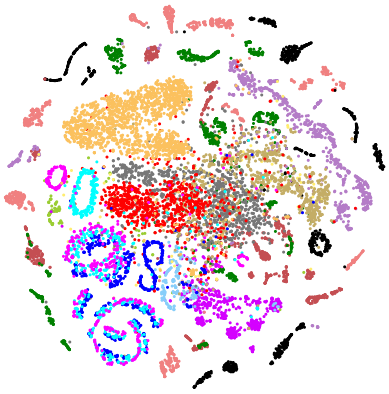
5.6 Disentanglement Analysis

To further verify the effectiveness of our BPD framework–(whether the intra-class variability is reduced), we applied t-SNE to generate visualisation on latent features for GOTOV dataset.

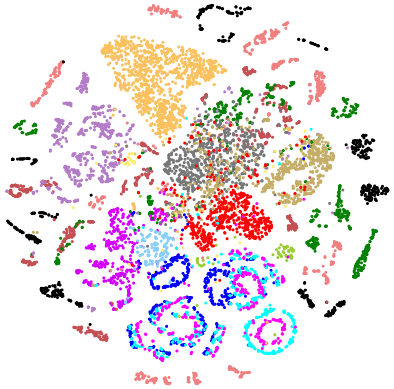
Figure 6 illustrates the t-SNE plot for the original CNN feature (Figure 6a), activity feature (Figure 6b) and redundant features (Figure 6c). We can witness that the clusters of activity embeddings of BPD (i.e., Figure 6b) are more distinct and organised than those of CNN and redundant features (i.e., Figure 6a and 6c), and samples with the same activity class tend to group into the same cluster, i.e., smaller intra-class variability. For example, "Dish washing", "Vacuum Cleaning", "Stacking Shelves" and "Step" seems more distinguishable in activity feature space than the other two. However, we also noticed that the redundant feature still contain some substantial activity patterns, which should be removed. One possible conjecture can be the limitation of this GOTOV dataset. Although it is a large dataset with 35 subjects, the population are older people (e.g., with age ranging from 61-73). The lack of diversity makes it challenging to remove the redundancy caused by personal factors completely.

5.7 Discussion and Limitation

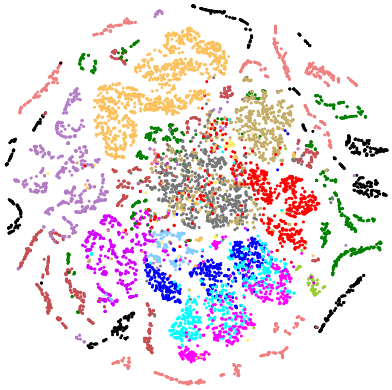
The proposed BPD framework aims to remove redundant features that do not contribute to the classification/recognition (in the training set), in order to reduce the intra-class variability for improved performance. For HAR scenarios, these redundant features correspond to the coupled effect of various covariate factors, e.g., the coupled effect of age, gender, weight, etc. and we expect performance gain by disentangling and removing these redundancies. However, current HAR datasets are normally limited due to population diversity, making it challenging to remove all the factors completely, and in Figure 6c, we can still observe the activity patterns in the



(a) w/o BPD(CNN)



(b) BPD(CNN) Activity Features



(c) BPD(CNN) Redundant Features

Fig. 6. Feature visualisation: t-SNE plot of CNN features, activity features and redundant features on GOTOV Dataset. We use different colours to denote different categories.(best viewed in colour)

redundant features, suggesting the activity signals and the redundancy were not completely separated in the

feature space. Nevertheless, our BPD framework can still reduce the effect of covariate factors, as suggested by the substantial performance gains on the 4 public datasets.

Although our framework can reduce the effect of covariate factors with improved performance,it remains unclear what these factors are. With additional metadata, a recent work GILE [46] attempted to disentangle human identity from activity signal, and this motivates us to explore further the attribute-oriented disentanglement frameworks. Although it may require additional meta-information for the disentanglement (between the attribute and the activity), it may provide a solution with higher interpretability.

Another major challenge is the cross-dataset evaluation. Different datasets may be affected by more challenging external factors such as unpredictable wearing locations or unknown hardware settings. For example, accelerometer devices from various manufacturers may have different xyz orientations. Although our BPD can be used to reduce the intra-class variability to some extent at the person level, it is hard to generalise to unknown hardware setups. Some wearing location protocol or device calibration should be applied for the cross-dataset evaluation, which will be explored in the future.

6 CONCLUSION

The main focus of this work is developing a feature disentanglement method that can effectively disentangle the redundant information from the initial signal to reduce the intra-class variability and improve the performance of HAR models. Such a method could become a prerequisite for wider adoption of HAR models/algorithms in real-world applications. In this case, we proposed the Behaviour Pattern Disentanglement(BPD) scheme for sensor-based human activity recognition. Specifically, we first design a novel signal&redundant feature disentanglement module that leverages the concept of adversary training to separate the activity feature and redundant feature from the initial representation. Then, we present a signal&redundant feature dependency reduction module to reduce the correlation between two disentangled features so to improve the disentanglement. Finally, we present an adversarial training algorithm to ensure the proposed BPD framework can be trained properly. To evaluate the HAR models more thoroughly, we conducted extensive experiments on four public datasets. Experimental results suggested it can further improve the performance of existing DL approaches (e.g., CNN or DeepConvLSTM), making it a flexible solution for the HAR research community.

REFERENCES

[1] Oresti Banos, Rafael Garcia, Juan A Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. 2014. mHealthDroid: a novel framework for agile development of mobile health applications. In *International workshop on ambient assisted living*. Springer, 91–98.

[2] Ling Bao and Stephen S Intille. 2004. Activity recognition from user-annotated acceleration data. In *International conference on pervasive computing*. Springer, 1–17.

[3] Billur Barshan and Murat Cihan Yüksek. 2014. Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units. *Comput. J.* 57, 11 (2014), 1649–1667.

[4] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *International Conference on Machine Learning*. PMLR, 531–540.

[5] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)* 46, 3 (2014), 1–33.

[6] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2018. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069* (2018).

[7] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint arXiv:1606.03657* (2016).

[8] Yan Gao, Yang Long, Yu Guan, Anna Basu, Jessica Baggaley, and Thomas Ploetz. 2019. Towards Reliable, Automated General Movement Assessment for Perinatal Stroke Screening in Infants Using Wearable Accelerometers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1, Article 12 (March 2019), 22 pages. https://doi.org/10.1145/3314399

[9] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673.

[10] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 249–256.
[11] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *arXiv preprint arXiv:1406.2661* (2014).
[12] Raghuraman Gopalan, Sima Taheri, Pavan Turaga, and Rama Chellappa. 2012. A blur-robust descriptor with applications to face recognition. *IEEE transactions on pattern analysis and machine intelligence* 34, 6 (2012), 1220–1226.
[13] Yu Guan and Thomas Plötz. 2017. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–28.
[14] Haodong Guo, Ling Chen, Liangying Peng, and Gencai Chen. 2016. Wearable sensor based multimodal human activity recognition exploiting the diversity of classifier ensemble. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 1112–1123.
[15] Nils Y Hammerla, Shane Halloran, and Thomas Plötz. 2016. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880* (2016).
[16] Nils Y Hammerla, Shane Halloran, and Thomas Plötz. 2016. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880* (2016).
[17] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework. (2016).
[18] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[19] BingZhang Hu, Yan Gao, Yu Guan, Yang Long, Nicholas Lane, and Thomas Ploetz. 2018. Robust cross-view gait identification with evidence: A discriminant gait gan (diggan) approach on 10000 people. *arXiv e-prints* (2018), arXiv–1811.
[20] BingZhang Hu, Yu Guan, Yan Gao, Yang Long, Nicholas Lane, and Thomas Ploetz. 2020. Robust Cross-View Gait Recognition with Evidence: A Discriminant Gait GAN (DiGGAN) Approach. arXiv:1811.10493 [cs.CV]
[21] BingZhang Hu, Feng Zheng, and Ling Shao. 2018. Dual-Reference Face Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[22] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. 2019. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics* 26, 11 (2019), 3365–3385.
[23] Aftab Khan, James Nicholson, and Thomas Plötz. 2017. Activity Recognition for Quality Assessment of Batting Shots in Cricket Using a Hierarchical Representation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 62 (Sept. 2017), 31 pages. https://doi.org/10.1145/3130927
[24] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[25] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
[26] Justin B Kinney and Gurinder S Atwal. 2014. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences* 111, 9 (2014), 3354–3359.
[27] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. 2011. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter* 12, 2 (2011), 74–82.
[28] Yann LeCun, Yoshua Bengio, et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* 3361, 10 (1995), 1995.
[29] Alexander H Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. 2018. A unified feature disentangler for multi-domain image translation and manipulation. *arXiv preprint arXiv:1809.01361* (2018).
[30] Francesco Locatello, Gabriele Abbati, Tom Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. 2019. On the fairness of disentangled representations. *arXiv preprint arXiv:1905.13662* (2019).
[31] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*. PMLR, 4114–4124.
[32] Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem. 2019. Disentangling factors of variation using few labels. *arXiv preprint arXiv:1905.01258* (2019).
[33] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.
[34] Michael Mathieu, Junbo Zhao, Pablo Sprechmann, Aditya Ramesh, and Yann LeCun. 2016. Disentangling factors of variation in deep representations using adversarial training. *arXiv preprint arXiv:1611.03383* (2016).
[35] Vishvak S Murahari and Thomas Plötz. 2018. On attention models for human activity recognition. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*. 100–103.
[36] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2017. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*. PMLR, 2642–2651.

[37] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16, 1 (2016), 115.
[38] Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron van den Oord, Sergey Levine, and Pierre Sermanet. 2019. Wasserstein dependency measure for representation learning. *arXiv preprint arXiv:1903.11780* (2019).
[39] Mahesh Pal. 2005. Random forest classifier for remote sensing classification. *International journal of remote sensing* 26, 1 (2005), 217–222.
[40] Stylianos Paraschiakos, Ricardo Cachucho, Matthijs Moed, Diana van Heemst, Simon Mooijaart, Eline P Slagboom, Arno Knobbe, and Marian Beekman. 2020. Activity recognition using wearable sensors for tracking the elderly. *User Modeling and User-Adapted Interaction* 30, 3 (2020), 567–605.
[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
[42] Thomas Plötz, Nils Y. Hammerla, Agata Rozga, Andrea Reavis, Nathan Call, and Gregory D. Abowd. 2012. Automatic Assessment of Problem Behavior in Individuals with Developmental Disabilities. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (Pittsburgh, Pennsylvania) *(UbiComp '12)*. Association for Computing Machinery, New York, NY, USA, 391–400. https://doi.org/10.1145/2370216.2370276
[43] Bin Qian, Jie Su, Zhenyu Wen, Devki Nandan Jha, Yinhao Li, Yu Guan, Deepak Puthal, Philip James, Renyu Yang, Albert Y Zomaya, et al. 2020. Orchestrating the development lifecycle of machine learning-based iot applications: A taxonomy and survey. *ACM Computing Surveys (CSUR)* 53, 4 (2020), 1–47.
[44] Hangwei Qian, Sinno Jialin Pan, Bingshui Da, and Chunyan Miao. 2019. A Novel Distribution-Embedded Neural Network for Sensor-Based Activity Recognition.. In *IJCAI*. 5614–5620.
[45] Hangwei Qian, Sinno Jialin Pan, Bingshui Da, and Chunyan Miao. 2019. A novel distribution-embedded neural network for sensor-based activity recognition. (2019).
[46] Hangwei Qian, Sinno Jialin Pan, and Chunyan Miao. 2021. Latent Independent Excitation for Generalizable Sensor-based Cross-Person Activity Recognition. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 11921–11929. https://ojs.aaai.org/index.php/AAAI/article/view/17416
[47] Attila Reiss and Didier Stricker. 2012. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th International Symposium on Wearable Computers*. IEEE, 108–109.
[48] Karl Ridgeway and Michael C Mozer. 2018. Learning deep disentangled embeddings with the f-statistic loss. *arXiv preprint arXiv:1802.05312* (2018).
[49] Eduardo Hugo Sanchez, Mathieu Serrurier, and Mathias Ortner. 2020. Learning disentangled representations via mutual information estimation. In *European Conference on Computer Vision*. Springer, 205–221.
[50] Mufan Sang, Wei Xia, and John HL Hansen. 2020. DEAAN: Disentangled Embedding and Adversarial Adaptation Network for Robust Speaker Representation Learning. *arXiv preprint arXiv:2012.06896* (2020).
[51] Albrecht Schmidt, Michael Beigl, and Hans-W Gellersen. 1999. There is more to context than location. *Computers & Graphics* 23, 6 (1999), 893–901.
[52] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).
[53] Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. 2019. Are disentangled representations helpful for abstract visual reasoning? *arXiv preprint arXiv:1905.12506* (2019).
[54] Junyan Wang, Bingzhang Hu, Yang Long, and Yu Guan. 2019. Order matters: Shuffling sequence generation for video prediction. *arXiv preprint arXiv:1907.08845* (2019).
[55] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. 2018. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610* (2018).
[56] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiaoli Li, and Shonali Krishnaswamy. 2015. Deep convolutional neural networks on multichannel time series for human activity recognition.. In *Ijcai*, Vol. 15. Buenos Aires, Argentina, 3995–4001.
[57] Bing Zhai, Ignacio Perez-Pozuelo, Emma A. D. Clifton, Joao Palotti, and Yu Guan. 2020. Making Sense of Sleep: Multimodal Sleep Stage Classification in a Large, Diverse Population Using Movement and Cardiac Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 67 (June 2020), 33 pages. https://doi.org/10.1145/3397325
[58] Mengshi Zhang, Yuqun Zhang, Lingming Zhang, Cong Liu, and Sarfraz Khurshid. 2018. DeepRoad: GAN-based metamorphic testing and input validation framework for autonomous driving systems. In *2018 33rd IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 132–142.

A GOTOV DATASET RESULT

Table 5. Mean F1-score for each subject on the GOTOV dataset (in leave-one-subject-out CV setting)

Dataset	Subject	CNN	DeepConvLSTM	beta-VAE	GILE	BPD (CNN)	BPD (DeepConvLSTM)
GOTOV	1	0.6852	0.7156	0.6200	0.6401	**0.7541**	0.7493
	2	0.7585	0.7241	0.7064	0.7120	**0.7730**	0.7287
	3	0.6770	0.6432	0.6221	0.6373	**0.7357**	0.6469
	4	0.7480	0.6993	0.5751	0.6931	**0.7497**	0.7263
	5	0.8295	0.8157	0.6757	0.7832	**0.8378**	0.8271
	6	0.6608	0.6569	0.5572	0.6212	**0.6742**	0.6670
	7	0.6581	0.7125	0.5845	0.6101	0.7194	**0.7391**
	8	0.7843	0.8176	0.6750	0.7532	0.8118	**0.8378**
	9	0.8945	0.8997	0.7365	0.8106	0.8968	**0.9018**
	10	0.6584	0.6527	0.5780	0.5701	**0.6621**	0.6537
	11	0.7950	0.7915	0.6649	0.7432	0.7984	**0.8164**
	12	0.6802	0.6816	0.5479	0.6330	**0.7458**	0.6894
	13	0.7515	0.6415	0.5382	0.6179	**0.7880**	0.6725
	14	0.5257	0.5047	0.3700	0.3841	**0.6243**	0.5231
	15	0.7852	0.7746	0.6751	0.7597	0.7946	**0.7961**
	16	0.5837	0.5936	0.5311	0.5332	**0.6537**	0.6210
	17	0.6124	0.6055	0.5328	0.5421	**0.6476**	0.6319
	18	0.5664	0.5102	0.4213	0.4979	**0.5709**	0.5305
	19	0.6900	0.6533	0.6192	0.6127	**0.7530**	0.6825
	20	0.8163	0.8190	0.7243	0.7904	**0.8725**	0.8219
	21	0.7902	0.7756	0.5745	0.6920	**0.8037**	0.7869
	22	0.4930	0.4564	0.3693	0.3988	**0.5122**	0.4826
	23	0.5523	0.5647	0.5030	0.4988	0.5540	**0.5841**
	24	0.6332	0.6250	0.6078	0.6320	**0.6643**	0.6374
	25	0.5303	0.4571	0.3908	0.4450	**0.5340**	0.4842
	26	0.4463	0.4204	0.3347	0.4102	**0.4813**	0.4455
	27	0.8107	0.7798	0.7157	0.7322	**0.8261**	0.7910
	28	0.6305	0.6046	0.5876	0.5886	**0.6529**	0.6377
	29	0.4893	0.4578	0.3921	0.4172	**0.4951**	0.4692
	30	0.6301	0.6405	0.5354	0.6078	0.6799	**0.6840**
	31	0.5527	0.5454	0.4941	0.5132	**0.5929**	0.5595
	32	0.7328	0.7065	0.6184	0.6932	**0.7620**	0.7388
	33	0.6779	0.6427	0.5948	0.6321	**0.6986**	0.6505
	34	0.6307	0.6169	0.5212	0.6039	**0.6490**	0.6453
	35	0.4982	0.5163	0.3670	0.4821	**0.5670**	0.5639
	Avg.	0.6645	0.6492	0.5589	0.6083	**0.6953**	0.6692