RESEARCH ARTICLE

Stacked deep analytic model for human activity recognition on a UCI HAR database [version 1; peer review: 2 approved with reservations]

Ying Han Pang1, Liew Yee Ping1, Goh Fan Ling2, Ooi Shih Yin1, Khoh Wee How1

1Faculty of Information Science and Technology, Multimedia University, Ayer Keroh, Melaka, 75450, Malaysia
2Millapp Sdn Bhd, Bangsar South, Kuala Lumpur, 59200, Malaysia

Abstract

Background
Owing to low cost and ubiquity, human activity recognition using smartphones is emerging as a trendy mobile application in diverse appliances such as assisted living, healthcare monitoring, etc. Analysing this one-dimensional time-series signal is rather challenging due to its spatial and temporal variances. Numerous deep neural networks (DNNs) are conducted to unveil deep features of complex real-world data. However, the drawback of DNNs is the uninterpretation of the network’s internal logic to achieve the output. Furthermore, a huge training sample size (i.e. millions of samples) is required to ensure great performance.

Methods
In this work, a simpler yet effective stacked deep network, known as Stacked Discriminant Feature Learning (SDFL), is proposed to analyse inertial motion data for activity recognition. Contrary to DNNs, this deep model extracts rich features without the prerequisite of a gigantic training sample set and tenuous hyper-parameter tuning. SDFL is a stacking deep network with multiple learning modules, appearing in a serialized layout for multi-level feature learning from shallow to deeper features. In each learning module, Rayleigh coefficient optimized learning is accomplished to extort discriminant features. A subject-independent protocol is implemented where the system model (trained by data from a group of users) is used to recognize data from another group of users.

Results
Empirical results demonstrate that SDFL surpasses state-of-the-art methods, including DNNs like Convolutional Neural Network, Deep Belief Network, etc., with ~97% accuracy from the UCI HAR database.

Open Peer Review

Approval Status

version 3
(revision)
01 Apr 2022

version 2
(revision)
18 Feb 2022

version 1
15 Oct 2021

1. Cheng-Yaw Low1, Yonsei University, Seoul, South Korea
   Institute for Basic Science, Daejeon, South Korea
2. Andrews Samraj1, Mahendra Engineering College, Namakkal, India

Any reports and responses or comments on the article can be found at the end of the article.
with thousands of training samples. Additionally, the model training time of SDFL is merely a few minutes, compared with DNNs, which require hours for model training.

**Conclusions**
The supremacy of SDFL is corroborated in analysing motion data for human activity recognition requiring no GPU but only a CPU with a fast-learning rate.

**Keywords**
smartphone, one-dimensional motion signal, activity recognition, stacking deep network, discriminant learning
Introduction
Human activity recognition (HAR) can be categorized into vision-based and sensor-based. In vision-based HAR, an image sequence, in the form of video, recording the human activity is captured by a camera. This sequence will be analysed to recognize the nature of an action. This system is applied for surveillance, human-computer interaction and healthcare monitoring. For sensor-based HAR, human activities are captured by inertial sensors, such as accelerometers, gyroscopes or magnetometers. Among these approaches, sensors are more favourable due to their lightweight nature, portability and low energy usage. With the advancement of mobile technology, smartphones are equipped with high-end components. Accelerometer and gyroscope sensors embedded in the smartphone make it feasible as an acquisition device for HAR. Smartphone-based HAR has been an area of contemporary research in recent years. In this work, we categorize the smartphone-based HAR as part of the sensor-based HAR. Activity inertial signals are collected through smartphone sensors.

Related work
Hand-crafted approaches using manually computed statistical features have been proposed. These authors applied various machine learning techniques such as decision tree, logistic regression, multilayer perceptron, naïve Bayes, Support Vector Machine etc. to classify the detected activities. The performance of the handcrafted approaches might be affected when dealing with complex scenarios due to their feature representation incapability. The algorithms could easily plummet into the local minimum despite the global optimal.

Hence, various deep neural networks (DNNs) were explored in HAR owing to the capability of extracting informative features. DNN is a machine learner that can automatically unearth the data characteristics hierarchically from lower to higher levels. The work of Ronao and Cho (2016), Lee et al. (2017) and Ignatov (2018) explored the deep convolutional neural networks by exploiting the activity characteristics in the one-dimensional time-series signals captured by the smartphone inertial sensors. The empirical results substantiated that the extracted deep features were crucial for data representation with promising recognition performance.

Zeng et al. (2014) proposed a modified convolutional neural network to extract scale-invariant characteristics and local dependency of the acceleration time-series signal. The weight sharing mechanism in the convolutional layer was modified. Unlike in the vanilla model where the local filter weights were shared by all positions within the input space, the authors incorporated a more relax weight sharing strategy (partial weight sharing) to enhance the performance.

Recurrent Neural Network (RNN) was proposed to process sequential data by analysing previously inputted data and processing it linearly. Due to the vanishing gradient problem, RNN was enhanced and Long Short-Term Memory – LSTM was introduced. Chen et al. (2016) explored the feasibility of LSTM in predicting human activities. Empirical results demonstrated an encouraging performance of LSTM in HAR. Further, an enhanced version of LSTM, known as bidirectional LSTM, was proposed. Unlike LSTM, bidirectional LSTM tackles both past and future information during the feature analysis. With this, a richer description of features could be extracted for classification.

A cascade ensemble learning (CELearning) model was proposed for smartphone-based HAR. There are multiple layers in this aggregation network and the model goes deeper layer by layer. Each layer contains Extremely Gradient Boosting Trees, Random Forest, Extremely Randomized Trees and Softmax Regression. The CELearning model gains higher performance, and the training process is rather simple and efficient. Besides, Hierarchical Multi-View Aggregation Network (HMVAN) is also one of the aggregation models. This model integrates features from various feature spaces in a hierarchical context. In this network, three aggregation modules from the aspect of feature, position and modality levels are designed.

Motivation and contributions
In DNNs, there are learning modular components in multiple processing layers for multiple-level feature abstraction. These layers are trained based on a versatile learning principle, which does not require any manual design by experts. These DNNs accomplish excellent performances in pattern recognition. However, these networks are not well trained if they have limited training samples, leading to performance degradation. Furthermore, there is a lack of theoretical ground on how to fine-tune the gigantic hyper-parameter series. The outstanding accomplishment of DNNs can only be achieved if and only if sufficient training data is accessible for fine-tuning the large parameter set. A high specification of GPU is needed to train the network from gargantuan datasets.

Thus, a stacking-based deep learning model for smartphone-based HAR is proposed. Inspired by the hierarchical learning in the DNNs, the proposed stacked learning network is aggregated with multiple learning modules, one after another, in a hierarchical framework. Specifically, a discriminant learning function is implemented in each module for discriminant
mapping to generate discriminative features, level by level. The lower (generic) to higher level (deeper) features are input to a classifier for activity identification. This proposed approach is termed Stacked Discriminant Feature Learning, coined as SDLF.

The contributions of this work are summarized in three-fold:

1. A deep analytic model is proposed for smartphone based HAR to extract deep features without the need of a gigantic training set and tenuous hyper-parameter tuning.

2. An adaptable modular model is developed with a discriminant learning function in each module to extract discriminant features from lower to higher levels demanding no graphics processing unit (GPU) but only a central processing unit (CPU) with a fast-learning rate.

3. An experimental analysis using various performance evaluation metrics (i.e. recall, precision, the area under the curve, computational time, etc.) with subject-independent protocol implementation in which there is no overlap in subjects between training and testing sets.

**Methods**

Smartphone inertial sensors were used to capture 3-axial linear (total) acceleration and 3-axial angular velocity signals. These signals were pre-processed into time- and frequency-domain features, as listed in Table 1. Next, the pre-processed data was inputted into the Stacked Discriminant Feature Learning (SDFL) for feature learning. The extracted feature template was fed into the nearest-neighbour (NN) classifier for classification. The overview of the system is illustrated in Figure 1.

| Function        | Feature                        |
|-----------------|--------------------------------|
| Mean            | Average value                  |
| Std dev         | Standard deviation             |
| Median          | Median absolute value          |
| Max             | Largest value in array         |
| Min             | Smallest value in array        |
| Sma             | Signal magnitude area          |
| Energy          | Average sum of squares         |
| Iqr             | Interquartile range            |
| Entropy         | Signal entropy                 |
| ArCoeff         | Auto-regression coefficients   |
| Correlation     | Correlation coefficient        |
| MaxFreqInd      | Largest frequency component    |
| MeanFreq        | Frequency signal weighted average |
| Skewness        | Frequency signal skewness      |
| Kurtosis        | Frequency signal kurtosis      |
| EnergyBand      | Energy of a frequency interval |
| Angle           | Angle between two vectors      |

**Figure 1.** Overview of the proposed Stacked Discriminant Feature Learning (SDFL) system.
SDFL is a pile of multiple discriminant learning layers interleaved with a nonlinear activation unit, as illustrated in Figure 2. By cascading multiple discriminant learning modules, each layer of SDFL learns based on the input data and the learned nonlinear features of the preceding module. The depth of the stacking layer is determined using the database subset. If the performance is not improving but showing degradation, the depth of the stacking layer is determined. In this case, the depth of three showed the optimal performance, so we adopted this architecture with three layers. To be detailed, the first discriminant learning module learns based on the input data and the second learning module learns based on an input vector (concatenating the input data and the learned features of the first learning module). This is similar to the third learning process where the third learning module learns based on an input vector (comprising the input data and the learned features of the second learning module).

Let \( x_i, y_i \) be a set of \( N \) transformed data, \( y_i \) is the class label of \( x_i \), \( C \) is the number of training classes, each of \( C \) classes has a mean \( \mu_j \) and total mean vector \( \mu = \frac{1}{N} \sum_{i=1}^{N} x_i \) with \( m_j \) denotes the number of training samples for \( j \)th class. In the first learning layer, the input vector is the transformed data \( x_i \). The computation of the intrapersonal scatter matrix \( \Sigma_{\text{intra}} \) and interpersonal scatter matrix \( \Sigma_{\text{inter}} \) are defined as:

\[
\Sigma_{\text{intra}} = \sum_{j=1}^{C} \sum_{i,j \in C_j} (x_i - \mu_j) (x_j - \mu_j)^T
\]

\[
\Sigma_{\text{inter}} = \sum_{j=1}^{C} \mu_j (\mu_j - \mu) (\mu_j - \mu)^T
\]

where T denotes a transpose operation. Next, a linear transformation \( \Phi \) is computed by maximizing the Rayleigh coefficient. With this optimization, the data from the same person could be projected close to each other, while data from different people is projected as far apart as possible. This optimization function is termed as Fisher’s criterion,

\[
J(\Phi) = \frac{\Phi^T \Sigma_{\text{intra}} \Phi}{\Phi^T \Sigma_{\text{inter}} \Phi}
\]

The mapping \( \Phi \) is constructed through solving the generalized eigenvalue problem,

\[
\Sigma_{\text{inter}} \Phi = \lambda \Sigma_{\text{intra}} \Phi
\]

The learned features are produced through the projection of the input data \( x_i \) onto the mapping subspace,

\[
\hat{x}_i = \Phi^T x_i
\]
\( \hat{x} \) is transformed to \( C - 1 \) dimensions. We denote \( l \) for the index of modular layer in SDFL. The learned feature vector of the first modular unit is notated as \( \hat{x}^{(1)} = \hat{x}^{(1)} \). A nonlinear input-output mapping is applied to \( \hat{x}^{(1)} \) via a nonlinear activation function. In this study, we adopt a sigmoid function, \( \hat{x}^{(1)} = S(\hat{x}) = \frac{1}{1+e^{-\hat{x}}} \) for the nonlinear projection. To be specific, \( \hat{x}^{(1)} = \frac{1}{1+e^{-\hat{x}^{(1)}}} \) is the nonlinear learned features of the first modular unit.

For deeper modules, the input vector of the respective module is a stacking vector containing the input data and the learned features, i.e. \( \mathbf{z}^{(l)} = [ \mathbf{x}^{(l)}, \hat{x}^{(l-1)} ] \) where \( l = 2 \) and \( 3 \). The intrapersonal scatter matrix \( \Sigma_{\text{intra}}^{(l)} \) and interpersonal scatter matrix \( \Sigma_{\text{inter}}^{(l)} \) are formulated,

\[
\Sigma_{\text{intra}}^{(l)} = \sum_{j=1}^{C} \sum_{\mathbf{z}^{(l)} \in C_j} (\mathbf{z}^{(l)} - \mu^{(l)}) (\mathbf{z}^{(l)} - \mu^{(l)})^T
\]

\[
\Sigma_{\text{inter}}^{(l)} = \sum_{j=1}^{C} \mu^{(l)} (\mu^{(l)} - \mu^{(l)}) (\mu^{(l)} - \mu^{(l)})^T
\]

In this case, \( \mu^{(l)} \) is the \( j \)th class mean computed from the input vectors of \( j \)th class, \( \mathbf{z}^{(l)} \in \mathcal{C}_j \) and the total mean vector \( \mu^{(l)} = \frac{1}{N} \sum_{j=1}^{N} \mu^{(l)} \) at \( l \)th modular unit. The final feature vector is the nonlinear learned features of each modular layer,

\[
\hat{x}^{\text{final}} = [\hat{x}^{(1)}, \hat{x}^{(2)}, \hat{x}^{(3)}]
\]

**Results**

We scrutinized how well SDFL could analyse the inertial data and correctly classify those activities. The experimental hardware platform was constructed on a desktop with an Intel® Core™ i7-7700 processor with 4.20 GHz and 48.0 GB main memory; whereas the experimental software platform was a 64-bit operating system of Windows 10 with Matlab R2018a (MATLAB, RRID:SCR_001622) software (An open-access alternative that provides an equivalent function is GNU Octave (GNU Octave, RRID:SCR_014398)).

We used the UCI HAR dataset\(^2\): There were 30 subjects with 7352 training samples and 2947 testing samples. Each subject was required to carry a smartphone (Samsung Galaxy SII) on the waist and perform six different activities. The activities were “walking”, “walking_upstairs”, “walking_downstairs”, “sitting”, “standing” and “laying”.

The generalization level of SDFL was evaluated in a user-independent scenario. SDFL was trained using training samples from a group of users. Then, the model was applied to new users without the necessity of collecting additional samples of these new users to retrain the model. In this experiment, the UCI HAR dataset was partitioned into two sets: 70% of the volunteers were selected to generate the training data and the remaining 30% of the volunteers’ data was used as the testing data. There was no subject overlapping between the training and test data sets. Table 2 records the performance of SDFL and Table 3 records the performance comparison with other approaches.

Table 4 tabulates the computational time. The computational time of SDFL is benchmarked with that of the ordinary methodology, which is directly performing classification on the pre-processed data. Instead of using a multiclass support

| Table 2. Performance of Stacked Discriminant Feature Learning (SDFL). |
|--------------------------|-----------|
| **Metric**               | **Performance** |
| True Positive (TP) rate  | 0.963     |
| False Positive (FP) rate | 0.008     |
| Precision                | 0.964     |
| Recall                   | 0.963     |
| F-score                  | 0.963     |
| Area Under the Curve     | 0.977     |
| Accuracy (%)             | 96.2674   |
vector machine as in, we adopt Nearest Neighbour (NN) classifier for classification because the focus of this work is the feature extraction capability and the classification is standardized with the simplest classifier, i.e. NN.

Discussion
From the empirical results, we observed that the proposed SDFL was able to demonstrate superior classification performance compared to most of the existing techniques, even though a simple classifier was adopted in the system. The exceptional performance of SDFL explains the capability of SDFL in capturing the essence of the inertial data without heavily depending on the classifier. Furthermore, SDFL also exhibited its superiority to most of the existing approaches, including deep learning models. To be specific, SDFL obtained an accuracy of 96.3%, whilst Deep Belief Network’s accuracy was 95.8%,4 CNN achieved 95.75% accuracy14 and ANN’s accuracy was 91.08%.

Last but not least, it was discerned that the performance of SDFL is on a par with the Cascade Ensemble Learning model (CELearning).20 Both approaches are ensemble learning methods with multiple layers for data learning. The key difference between these approaches is the analysis algorithms in each layer. CELearning is comprised of four different classifiers, i.e. Random Forest, Extremely Gradient Boosting Trees, Softmax Regression and Extremely Randomized Trees and the final classification result is obtained through the last layer via the score-level fusion of the four complex classifiers. On the other hand, in SDFL, merely Rayleigh coefficient optimization is implemented to extract the low-to-higher level of discriminant features. Further, a simple classifier, i.e. NN classifier, is adopted in SDFL. This deduces that the discrimination capability of SDFL primarily depends on the SDFL modular model to extract discriminant features demanding no complex classifier.

From Table 4, we can notice that the overall training and testing time of SDFL are much lesser than those of the benchmark method. On average, SDFL just needs $\sim 4.3 \times 10^{-4}$ seconds per sample (sps) for the training phase and $\sim 4.2 \times 10^{-4}$ sps for the testing phase. The fast feature learning of SDFL and the dimensionality reduction in SDFL to project the data onto a lower-dimensional subspace are the main reasons for having such an efficient computation.
Conclusions
A cascading learning network for human activity recognition using smartphones is proposed. In this network, a chain of independent discriminant learning modules is aggregated, layer by layer in a stackable framework. Each layer is constituted by a discriminant analysis function and a nonlinear activation function to effectively extract the rich features from the inertial data. This proposed SDFL network possesses characteristics of good performance even on small-scale training sample sets, as well as less hyper-parameter fine-tuning, and fast computation compared with the other deep learning networks. Despite showing computational efficiency, the proposed network also demonstrated its classification superiority to most of the state-of-the-art approaches with an accuracy score of ~97% in differentiating human activity classes.

Data availability
All data underlying the results are available as part of the article and no additional source data are required.

References

1. Poppe R: A survey on vision-based human action recognition. Image Vis. Comput. 2010; 28(6): 976–990.
2. Ahmed N, Rafiq JI, Islam MR: Enhanced Human Activity Recognition Based on Smartphone Sensor Data Using Hybrid Feature Selection Model. Sensors. Jan. 2020; 20(1): 317. PubMed Abstract | Publisher Full Text | Free Full Text
3. Cao L, Trocan M: Deep learning of smartphone sensor data for personal health assistance. Microelectronics J. Jun. 2019; 88: 164–172. Publisher Full Text
4. Hernández F, Suárez LF, Villamizar J, et al.: Human Activity Recognition on Smartphones Using a Bidirectional LSTM Network. 2019 22nd Symp. Image, Signal Process. Artif. Vision, STSIVA 2019 - Conf. Proc. 2019: 1–5. Publisher Full Text
5. Li H, Trocan M: Deep learning of smartphone sensor data for personal health assistance. Microelectronics J. Jun. 2019; 88: 164–172. Publisher Full Text
6. Yang J, MN N, PP S, et al.: Human Activity Recognition Using Multichannel Time Series for Human Activity Recognition. ICAI 2015.
7. Sun J, Fu Y, Li S, et al.: Sequential Human Activity Recognition Based on Deep Convolutional Network and Extreme Learning Machine Using Wearable Sensors. J. Sensors. 2018; 8580959. Publisher Full Text
8. Nweke HF, Teh YW, Al-garadi MA, et al.: Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. Expert Systems with Applications. Elsevier Ltd; Sep. 01, 2018; vol. 105. pp. 233–261. Publisher Full Text
9. Kwapisz JR, Weiss GM, Moore SA: Activity recognition using cell phone accelerometers. ACM SIGKDD Explor. Newsl. 2011; 12(2): 74–82. Publisher Full Text
10. Wu W, Dasqupta S, Ramirez EE, et al.: Classification accuracies of physical activities using smartphone motion sensors. J. Med. Internet Res. Sep. 2012; 14(5): e130. PubMed Abstract | Publisher Full Text | Free Full Text
11. Anguita D, Ghio A, Oneto L, et al.: A public domain dataset for human activity recognition using smartphones. 2013. Publisher Full Text
12. Temitope Yekeen S, Balogun AL, Wan Yusof KB: A novel deep learning instance segmentation model for automated marine oil spill detection. ISPRS J. Photogramm. Remote Sens. Sep. 2020; 167: 190–200. Publisher Full Text
13. Ronao CA, Cho SB: Human activity recognition with smartphone sensors using deep learning neural networks. Expert Syst. Appl. 2016; 59: 235–244. Publisher Full Text
14. Lee SM, S M, Cho H, et al.: Human Activity Recognition From Accelerometer Data Using Convolutional Neural Network. IEEE Int. Conf. Big Data Smart Comput. (BigComp), 2017; vol. 62, pp. 131–134. Publisher Full Text
15. Ignatov A: Real-time human activity recognition from accelerometer data using Convolutional Neural Networks. Appl. Soft Comput. 2018; 62: 915–922. Publisher Full Text
16. Chen Y, Zhong K, Zhang J, et al.: Dynamic Fusion Networks for Accurate Human Activity Recognition Based on Smartphone Sensor Data. Proc. - 2019 IEEE Intl Conf. Big Data Smart Comput. (BigComp), 2019: 1–10. Publisher Full Text
17. Zeng M, et al.: Convolutional Neural Networks for human activity recognition using mobile sensors Article. 2014: 381–388. Publisher Full Text
18. Chen Y, Zhong K, Zhang J, et al.: LSTM Networks for Mobile Human Activity Recognition. no. Icaita. 2016; pp. 50–53. Publisher Full Text
19. Xu S, Qin L: Human activity recognition with smartphone inertial sensors using bidir-LSTM networks. Proc. - 2018 3rd Int. Conf. Mech. Control Comput. Eng. ICMCCCE 2018: 219–224. Publisher Full Text
20. Xu S, Tang Q, Jin L, et al.: A cascade ensemble learning model for human activity recognition with smartphones. Sensors (Switzerland), May 2018; 18(10). PubMed Abstract | Publisher Full Text | Free Full Text
21. Zhang X, Wang Y, Kankanhalli MS, et al.: Hierarchical multi-view aggregation network for sensor-based human activity recognition. PLoS One. 2019; 14(9): e0221390. PubMed Abstract | Publisher Full Text | Free Full Text
22. Lecun Y, Bengio Y, Hinton G: Deep learning. Nature. 2015; 521: 436–444.
23. Fukunaga K: Introduction to Statistical Pattern Recognition. Elsevier; 1990.
24. Seto S, Zhang W, Zhou Y: Multivariate time series classification using dynamic time warping template selection for human activity recognition. Proc. - 2015 IEEE Symposium Series on Computational Intelligence, SSCI 2015: 1399–1406. Publisher Full Text
25. Ronao CA, Cho SB: Recognizing human activities from smartphone sensors using hierarchical continuous hidden Markov models. Int. J. Distrib. Sens. Networks. 2017; 13(1). Publisher Full Text
Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 30 November 2021

https://doi.org/10.5256/f1000research.76806.r99185

© 2021 Samraj A. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Andrews Samraj

Department of Computer Science and Engineering, Mahendra Engineering College, Namakkal, Tamil Nadu, India

1. The motivation and contributions may be rewritten to make it simpler to understand the purpose and achievements of the work (purpose of deep features have to be mentioned).

2. Any chances of an Intra personal scatter matrix with any other inter personal scattermatrix? How was the threshold or borderline decided? There needs to be explanations with sample values.

3. The used hardware and software details, and details about dataset should be shifted from result to any other section in methodology.

4. Slight corrections on technical writings need to be carried out. (eg: Table 4 tabulates: can be changed as either: Table4 presents or it is tabulated in table 4)

5. A sample subject wise (Personnel) data table would be nice if presented to know the inter personnel differences.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
I cannot comment. A qualified statistician is required.
Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** AI, Patterns, Bionics, ML and DEEP Learning, signals,sensors

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

---

Author Response 15 Feb 2022

**Ying Han Pang**, Multimedia University, Ayer Keroh, Malaysia

First of all, we would like to express our heartiest thanks to the Editor-in-Chief and Reviewers who have provided us with many insightful comments and a chance to improve our works.

1. The motivation and contributions may be rewritten to make it simpler to understand the purpose and achievements of the work (purpose of deep features have to be mentioned).

   Response: The motivation and contribution have been revised for better clarification.

2. Any chances of an Intra personal scatter matrix with any other inter personal scatter matrix? How was the threshold or borderline decided? There needs to be explanations with sample values.

   Response: The Method section has been revised for better clarification. The optimal projection in SDFL is computed by optimizing the Rayleigh coefficient for modelling the difference between the classes of data through maximizing the ratio of the inter-personal scatter matrix and intra-personal scatter matrix. The variability between data is contained in a subspace spanned by the eigenvectors corresponding to the number of class -1 largest eigenvalues. Hence, the threshold is $C^{-1}$ in this work.

3. The used hardware and software details, and details about dataset should be shifted from result to any other section in methodology.

   Response: The amendment has been done and the details have been shifted to the Method section.

4. Slight corrections on technical writings need to be carried out. (eg: Table 4 tabulates: can be changed as either: Table4 presents or it is tabulated in table 4)

   Response: The correction has been done.
5. A sample subject-wise (Personnel) data table would be nice if presented to know the inter-personnel differences.

Response: Figure 3 is added to illustrate the inter-personnel inertial signal differences of standing activity.

**Competing Interests:** No competing interests were disclosed.
a) For reproducibility, the parameter configurations should be disclosed.

b) It would be better if an ablation study is presented, e.g., exploration of the number of SDFL layers, etc. (refer to [14]). Furthermore, the baseline performance, the accuracy for the preprocessed data prior to SDFL learning, is not presented.

c) I suggest the authors include additional small-scale HAR datasets for more extensive experiments.

d) I think cross-validation should also be performed.

e) I suggest the authors double-check the results reported for precision, recall, f-score, and accuracy. It is rare that these varying matrices give the same value, approximately 96.3%. Alternatively, the confusion matrices for these evaluation matrices should be disclosed.

(f) In Table 4, I think the inference time should be computed for the SDFL and other conventional models, in place of the pre-processed data? Otherwise, this comparison may not be meaningful.

(g) In Table 3, the authors compare SDFL with [24], [25], [4], [3], etc., however, not all are reviewed in the related work section.

References
1. Ronao C, Cho S: Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications*. 2016; 59: 235-244 Publisher Full Text

*Is the work clearly and accurately presented and does it cite the current literature?*
Partly

*Is the study design appropriate and is the work technically sound?*
Partly

*Are sufficient details of methods and analysis provided to allow replication by others?*
Partly

*If applicable, is the statistical analysis and its interpretation appropriate?*
Partly

*Are all the source data underlying the results available to ensure full reproducibility?*
Partly

*Are the conclusions drawn adequately supported by the results?*
Partly

*Competing Interests:* No competing interests were disclosed.
Reviewer Expertise: Stacked neural networks, deep neural networks

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 15 Feb 2022

Ying Han Pang, Multimedia University, Ayer Keroh, Malaysia

First of all, we would like to express our heartiest thanks to the Editor-in-Chief and Reviewers who have provided us with many insightful comments and a chance to improve our works.

1. Objective 2: The authors claim that SDFL learns low-level to higher-level feature representation. This hypothesis should be demonstrated, if possible. Otherwise, I think this objective should be revised accordingly.

Response: Thanks for the feedback. Authors have revised Objective 2 for better clarification.

2. Objective 3: The authors should elaborate the subject-independent protocol from the motivation section as a problem statement to be resolved. In addition to that, the advantages of the subject-independent protocol should also be revealed.

Response: Authors have included the preference of subject-independent solution in real-time applications in the Motivation section for better relating the motivation and the Objective 3.

3. a) For clarity, the data/feature dimensionalities for all mathematical equations should be indicated accordingly. b) The SDFL hyper-parameters are unknown?

Response: Authors have included dimensionalities of data/feature, number of training classes \( C \) and the generated final feature vector in the Methods section for better clarification. Parameters of SDFL are the number of stacking layers, nonlinear activation function, dimensions of intermediate feature vectors and dimensions of the final feature vector. The Method section has been revised to include the information of these parameters for better clarification.

4. a) For reproducibility, the parameter configurations should be disclosed. b) It would be better if an ablation study is presented, e.g., exploration of the number of SDFL layers, etc. (refer to [14]). Furthermore, the baseline performance, the accuracy for the preprocessed data prior to SDFL learning, is not presented. c) I suggest the authors include additional small-scale HAR datasets for more extensive experiments. d) I think cross-validation should also be performed. e) I suggest the authors double-check the results reported for precision, recall, f-score, and accuracy. It is rare that these varying matrices give the same value, approximately 96.3%. Alternatively, the confusion matrices for these evaluation matrices should be disclosed. 

(f) In Table 4, I think the inference time should be computed for the SDFL and other conventional ...
models, in place of the pre-processed data? Otherwise, this comparison may not be meaningful. (g) In Table 3, the authors compare SDFL with [24], [25], [4], [3], etc., however, not all are reviewed in the related work section.

Response: The information of these parameters has been included in the Methods section for better clarification. Table 3 has been revised to include the benchmark method, that is Multiclass Support Vector Machine, proposed by the original author of the UCI database. This baseline performance which is the accuracy of the preprocessed data with Support Vector Machine has been included in the table. In order to have a better performance comparison with the existing methods, the testing protocol of this study follows the train-test split protocol defined by the database provider, i.e. this database has a version that is already split into training and test sets that contain data from different participants. This paper is using this version of the database. The results have been double-checked and the confusion matrix of SDFL has been included. Table 4 has been revised by just presenting the computational time of SDFL. With this, readers can have a picture of the training and inference time of the proposed SDFL. The Related Work section has been revised to include the reference in Table 3.

**Competing Interests:** No competing interests were disclosed.