Summary of the Sussex-Huawei Locomotion-Transportation Recognition Challenge 2019

Lin Wang
lin.wang@qmul.ac.uk
Centre for Intelligent Sensing
Queen Mary University of London, UK

Hristijan Gjoreski
hristijang@feit.ukim.edu.mk
Faculty of Electrical Engineering and Information Technologies
Ss. Cyril and Methodius University, MK

Mathias Ciliberto
m.ciliberto@sussex.ac.uk
Wearable Technologies Lab
University of Sussex, UK

Paula Lago
paula@mns.kyutech.ac.jp
Kyushu Institute of Technology, Japan

Kazuya Murao
murao@cs.ritsumei.ac.jp
College of Info. Sci. and Eng.
Ritsumeikan University, Japan

Tsuyoshi Okita
tsuyoshi.okita@gmail.com
Kyushu Institute of Technology, Japan

Daniel Roggen
daniel.roggen@ieee.org
Wearable Technologies Lab
University of Sussex, UK

ABSTRACT

In this paper we summarize the contributions of participants to the Sussex-Huawei Transportation-Locomotion (SHL) Recognition Challenge organized at the HASCA Workshop of UbiComp 2019. The goal of this machine learning/data science challenge is to recognize eight locomotion and transportation activities (Still, Walk, Run, Bike, Bus, Car, Train, Subway) from the inertial sensor data of a smartphone in a placement independent manner. The training data is collected with smartphones placed at three body positions (Torso, Bag and Hips), while the testing data is collected with a smartphone placed at another body position (Hand). We introduce the dataset used in the challenge and the protocol for the competition. We present a meta-analysis of the contributions from 14 submissions, their approaches, the software tools used, computational cost and the achieved results. Overall, three submissions achieved F1 scores between 70% and 80%, five with F1 scores between 60% and 70%, five between between 50% and 60%, and one below 50%, with a latency of a maximum of 5 seconds.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**; **Supervised learning by classification**.

KEYWORDS

Activity recognition; Deep learning; Machine learning; Mobile sensing; Transportation mode recognition

1 INTRODUCTION

The user's transportation mode is an important contextual information which enables adaptive services such as route or parking recommendation, proactive suggestions about transportation timetable, or more accurate measurements of energy expenditure. Several prior work looked at recognizing modes of transportation from smartphone sensors, including motion, GPS, sound, and image [15, 16, 21, 22]. To date, most research groups assess the performance of their algorithms using their own datasets on their own recognition tasks. These tasks often differ in the sensor modalities used or in the allowed recognition latency. This makes it difficult to compare methodologies and to systematically advance research in the field.

Following on our successful 2018 challenge [19], which saw 22 submissions, we organized the second Sussex-Huawei Locomotion-Transportation (SHL) recognition challenge in the year 2019[1]. In the previous SHL 2018, we focused on the development of position-specific recognition models using the data of a phone located in a front trousers pocket. In SHL 2019, we use previously unreleased data. The goal of this challenge is to recognize 8 modes of locomotion and transportation (activities) from the inertial sensor data of a smartphone in a mobile-phone placement independent manner. This paper introduces the dataset used for the challenge and the protocol

HASCA '19, September 10, 2019, London, UK
2019. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] http://www.shl-dataset.org/activity-recognition-challenge-2019/

for the competition, and summarizes and analyzes the achievements of the participants contributing to the challenge.

2 DATASET AND TASK

Dataset

The challenge uses a subset of the Sussex-Huawei Locomotion-Transportation (SHL) dataset [17, 18]. The SHL dataset was recorded over a period of 7 months in 2017 by 3 participants engaging in 8 different modes of transportation in real-life setting in the United Kingdom, i.e. Still, Walk, Run, Bike, Car, Bus, Train, and Subway. Each participant carried four smartphones at four body positions simultaneously: in the hand, at the torso, in the hip pocket, in a backpack or handbag (see Fig. 1). The smartphone logged data from 16 sensor modalities. The complete dataset contains up to 2812 hours of labeled data, corresponding to 16,732 km travel distance, and is considered as one of the biggest dataset in the research community.

The SHL Challenge 2019 uses the data recorded by the 4 phones of one user at the positions indicated in Fig. 1. It includes 82 days of recording (5-8 hours per day) during a 4-month period. The data is divided into three parts: train, validate and test The data comprises of 59 days of training data collected at three positions (hip, torso and bag), 20 days of test data collected at the hand position, and 3 days of validation data collected at all the four positions[2][3]. Fig. 2 depicts the duration of each transportation activity in the training, validation and testing datasets. In total, we have 90.7×3 hours of training data, 77 hours of testing data and 4.25×4 hours of validation data, respectively.

The challenge dataset contains the raw data from 7 sensors, including accelerometer, gyroscope, magnetometer, linear acceleration, gravity, orientation, and ambient pressure. The sampling rate of all these sensors is 100 Hz. The activity labels (class label) of the training and validation data is provided. The class label for the testing data is invisible to the participants for evaluation.

Data Format

The training, validation and testing data was generated by segmenting the whole data with a non-overlap sliding window of 5 seconds. The rationale for this is to force challenge participants to design algorithms which operate with a latency of a maximum of 5 seconds, which can be relevant in real-time interactive applications. The frames for the train data are consecutive in time. The frames in the validation and the testing data are randomly permuted.

[2]The exact dates for splitting the dataset will be released at the challenge website http://www.shl-dataset.org/activity-recognition-challenge-2019/.

[3]Note that the validation data is same as the one (the same user) released in a previewed version of the SHL dataset. http://www.shl-dataset.org/download/#shldataset-preview.

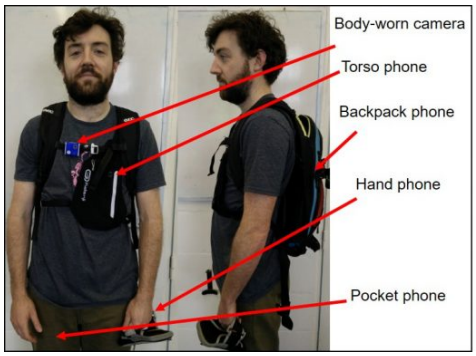

Figure 1: Smartphone positioning during data collection.

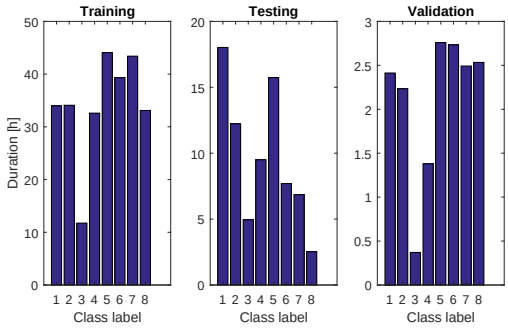

Figure 2: The duration of each class activity in the training and the testing dataset. The 8 classes are: 1 - Still; 2 - Walk; 3 - Run; 4 - Bike; 5 - Car; 6 - Bus; 7 - Train; 8 - Subway.

Table 1: Data files provided by the SHL recognition challenge. Position: B - Bag; T - Torso; Hi - Hips; Ha - Hand.

Modality	File	Train (B/T/Hi)	Validation (B/T/Hi/Ha)	Test (Ha)
Accelerometer	Acc_x.txt Acc_y.txt Acc_z.txt	✓	✓	✓
Gyroscope	Gyr_x.txt Gyr_y.txt Gyr_z.txt	✓	✓	✓
Magnetometer	Mag_x.txt Mag_y.txt Mag_z.txt	✓	✓	✓
Linear accelerometer	LAcc_x.txt LAcc_y.txt LAcc_z.txt	✓	✓	✓
Gravity	Gra_x.txt Gra_y.txt Gra_z.txt	✓	✓	✓
Orientation	Ori_w.txt Ori_x.txt Ori_y.txt Ori_z.txt	✓	✓	✓
Pressure	Pressure.txt	✓	✓	✓
Label	Label.txt	✓	✓	×

As shown in Table 1, the training data contains the data from three positions: Bag, Torso and Hips; the testing data contains one position: Hand; the validation data contains all the four positions: Bag, Torso, Hips and Hand. Each position in the training/validation dataset contains 21 plain text files, including 20 sensor files and 1 label file. Each position in the testing dataset only contains the 20 sensor files and excludes the label file.

Each sensor data file in each position of the training set contains a matrix of size 196,072 lines × 500 columns, corresponding to 196,072 frames each containing 500 samples (5 seconds at the sampling rate 100 Hz). The data in the label file is of the same size (196,072×500), indicating sample-wise transportation activity. Similarly, each sensor data file in each position of the validation set contains a matrix of size 12,177× 500. The label file is of same size as the sensor data. Each sensor data file in each position of the testing set contains a matrix of size 55,811×500. The label file of the testing set will remain confidential until after the challenge. It is used for performance evaluation by the challenge organizer. The total size of the data in ASCII format are 57.6, 5.4 and 4.8 GB for the training, testing and validation set, respectively.

The 8 numbers in the label file indicate the 8 activities: 1 - Still; 2 - Walk; 3 - Run; 4 - Bike; 5 - Car; 6 - Bus; 7 - Train; 8 - Subway. [4]

Task and Evaluation

The task is to train a recognition pipeline using the training/validation dataset and then use this system to recognize the transportation mode from the sensor data in the testing set. The recognition performance is evaluated with the F1 score averaged over all the activities.

Let M_{ij} be the (i, j)-th element of the confusion matrix. It represents the number of samples originally belonging to class i which are recognized as class j. Let $C = 8$ be the number of classes. The F1 score is defined as below.

$$\text{recall}_i = \frac{M_{ii}}{\sum_{j=1}^{C} M_{ij}}, \quad \text{precision}_j = \frac{M_{jj}}{\sum_{i=1}^{C} M_{ij}}, \quad (1)$$

$$F1 = \frac{1}{C} \sum_{i=1}^{C} \frac{2 \cdot \text{recall}_i \cdot \text{precision}_i}{\text{recall}_i + \text{precision}_i}. \quad (2)$$

3 RESULTS

Twenty-five teams expressed interests in the initial registration stage. The teams had 1.5 months (15 May - 30 June 2019) to develop the methods and work on the challenge task. Eventually, 14 teams contributed 14 submissions in the final submission stage by the deadline of 30 June. Table 2 summarizes the 14 submissions and Table 3 shows the detailed confusion matrices computed on the testing dataset.

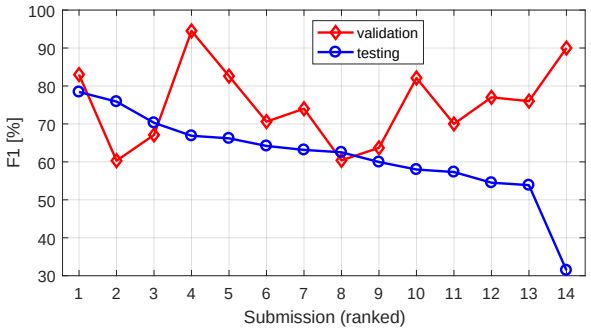

Figure 3: The submissions are ranked based on their F1 scores on the testing set (more details are given in Table 2).

Fig. 3 depicts the F1 scores of each submission for the testing set, as well as the F1 scores evaluated on the validation set (hand phone). The submissions are ranked based on their performance on the testing set (Table 2). The performance of the submissions ranges from 31.5% to 78.4%. There are 3 submissions achieving F1 scores above 70% on the testing set, 4 between 60% and 70%, 5 between 50% and 60%, and 1 below 50%.

In Fig. 3, the validation result on hand phone shows that the submissions 1, 3, 6, 7, 8, 9 generalize well between the training/validation and the testing data. The submission 2 shows certain under-fitting. The other submissions suffer form over-fitting. We briefly introduce the approaches used by the top three.

JSI-First achieves the highest F1 score of 78.%. The approach employs a cross-location transfer learning approach which trains two models: one using hand data in the validation set and one using non-hand data in both training and validation set [1]. A two-step classification method is employed, where the first model is used to classify all instances and the second model is used to reclassify all instances that were previously classified as still or vehicle. *Yonsei-MCML* achieves the second highest F1 score of 75.9% with a deep multimodal fusion model. The sensor data are independently pre-processed via a convolutional neural network (CNN), and the results are combined with the EmbraceNet fusion algorithm [2]. *We-can-fly* takes the third place with its F1 score 70.3%. It employs a 1D DenseNet model working on the multi-channel sensor data simultaneously [3].

[4] Note that we removed all the 'null' class from the raw data.

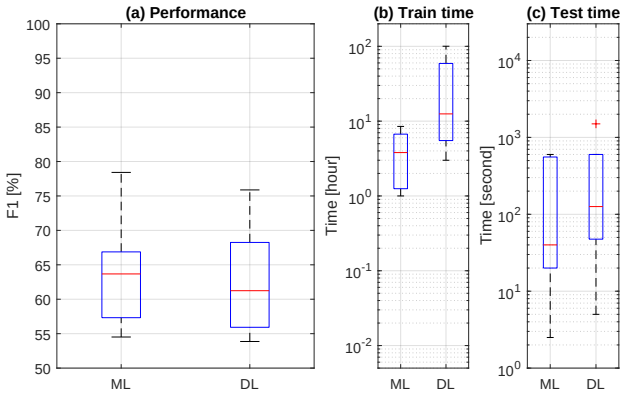

Figure 4: Comparison between machine learning and deep learning approaches. (a) F1 score for the testing data. (b) Training time. (c) Testing time.

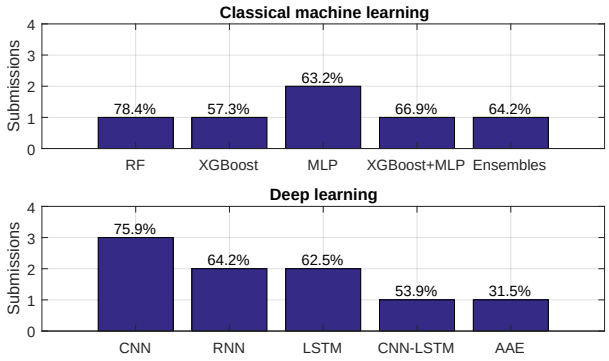

Figure 5: Classical machine learning and deep learning classifiers used by the submissions. The text on top of the bar indicates the highest F1 score achieved by each group of classifiers.

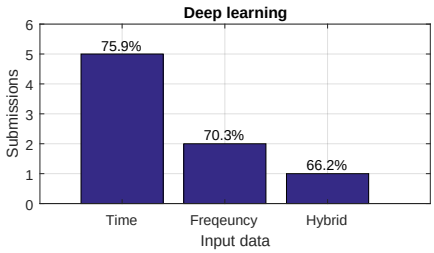

Figure 6: Type of input data to the deep-learning classifier. The text on top of the bar indicates the highest F1 score achieved by each type of input.

4 SUMMARY OF APPROACHES

We categorize the 14 submissions into two families: classical machine learning pipeline (ML) and deep learning pipeline (DL). There are 6 ML submissions and 8 DL submissions.

Fig. 4(a) box-plots the F1 scores obtained these two families. Interestingly, while ML has less submissions, it tends to outperform DL with higher upper bound and lower bound of the box. This is possibly because of the mismatch between the the training data (torso, bag, hips) and the testing data (hand). The features learned by DL for the source locations does not work well for the new target location. In contrast, hand crafted features, which incorporate 'human optimization', are shown to be robust dealing with this issues. The best performance achieved by the ML approach (JSI-First [1], 78.4%) is 2.5% higher than the best DL approach (Yonsei-MCML [2], 75.9%). Fig. 4(b)-(c) show in box-plot the training and testing time by ML and DL approaches, respectively. DL usually takes much more time for training than ML, and takes slightly more time for testing.

Fig. 5 depicts the specific classifiers employed by ML and DL pipelines. ML involves five classifiers: extreme gradient boost (XGBoost), random forest (RF), multi-layer perceptron neural network with less than 2 hidden layers (MLP), XGBoost+MLP, and ensembles of classifiers. DL involves five classifiers: convolutional neural network (CNN), recursive neural network (RNN), long-short term memory neural network (LSTM), CNN+LSTM, and adversial autoencoder (AAE).

For classical machine learning, MLP (2S - 2 submissions) is the most popular classifier, with the other four classifiers each with one submission. Note that RF and XGBoost use ensembles of decision trees. The approach in [7] uses ensembles of MLPs. Among these classifiers, RF achieves the highest F1 score of 78.4%, followed by XGBoost+MLP (66.9%). For deep learning, CNN (3S - submissions) is the most popular classifier, followed by RNN (2S) and LSTM (2S). CNN achieves the highest F1 score of 75.9%, followed by RNN (64.2%) and LSTM (62.5%).

All the 6 ML approaches uses hand-crafted features as input to the classifier. DL may use different types of raw data as input to the classifier (Fig. 6), either in the time domain (5S), in the frequency domain (2S), or hybrid (1S). The hybrid input achieves the highest F1 score of 75.9%, followed by time-domain raw data (70.3%) and frequency- domain raw data (66.2%).

Post-processing

Three submissions combine ensemble method with a post-processing scheme, which leads to quite good results. The approach in [1] employs hidden Markov model (HMM) to temporally smooth the results from RF, achieving the highest F1 score (78.4%). Note that while the frames in the testing set are randomly shuffled, the submission [1]

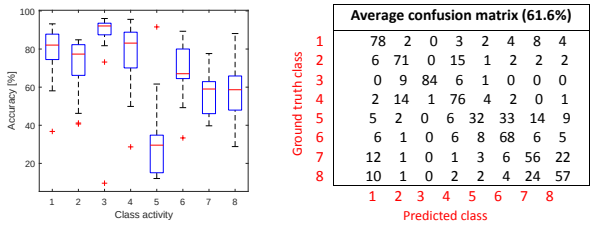

Figure 7: Recognition accuracy for each class activity by the top 13 submissions and the average confusion matrix. The 8 class activities are: 1 - Still; 2 - Walk; 3 - Run; 4 - Bike; 5 - Car; 6 - Bus; 7 - Train; 8 - Subway.

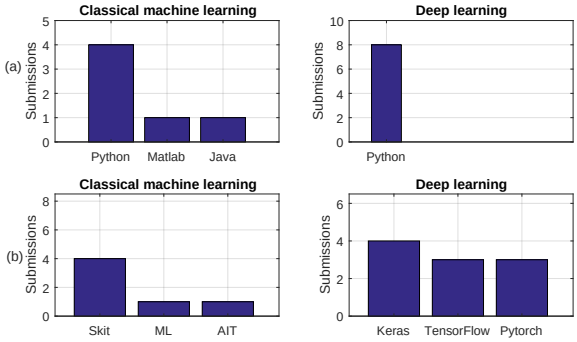

Figure 8: Programming languages and machine libraries used by the submissions for classical machine learning and deep learning. (a) Programming. (b) Library.

proposed a method, which, by looking at the correlation of the sensor data, can roughly recover the temporal order of frames before applying HMM. The approach in [2] employs the EmbraceNet to fuse the decisions from multiple independent single-modality classifiers, achieving the second highest F1 score (75.9%). Another approach [6] employs an RNN network to combines the decisions from an ensemble of classifiers, achieving the 6th highest F1 score (64.2%).

5 PERFORMANCE ANALYSIS

In Fig. 3, 13 out of 14 submissions achieve F1 scores between 50% and 80%. We analyze the results from the top 13 submissions.

Fig. 7 box-plots the recognition accuracy for each class activity (i.e. the diagonal elements of the confusion matrix), among the top 13 submissions, and also presents the average confusion matrix of their results. It can be observed from the box-plot that the class Car is the most difficult activity to recognize, followed by Train, Subway and Bus. It can be observed from the average confusion matrix that, Car tends to be misclassified as Bus, and Train and Subway tend to be misclassified as each other.

Overall, the first four activities (Still, Walk, Run, and Bike) are better recognized compared to the last four (Car, Bus, Train, and Subway). The motion of the smartphones during walk, run and bike is more distinctive than when the person is sitting or standing in the car, bus, train or subway, thus making the first four activities more distinctive than the last four. There is mutual confusion between the motor vehicles (Car vs Bus), and between the rail vehicles (Train vs Subway). The reason for this is the similar motion patterns during these activities. Some confusion between Still and the four vehicle activities (Car, Bus, Train and Subway) is also observed. This is similar to previous results reported in SHL 2018 [19] and in our baseline evaluation [18].

6 IMPLEMENTATION

Fig. 8(a) summarizes the programming languages used by the submissions. For ML, Python (4S) is the most popular languages among 6 submissions, followed by Matlab (1S) and Java (1S). For DL, Python is the dominant language used by all 8 submissions. Fig. 8(b) summarizes the machine learning libraries used by the submissions. For ML, Scikit-Learn (Python) is the mostly used library (4S), followed by Matlab Machine Learning Toolbox (1S) and AIT (1S). For DL, Keras (4S) is the most popular library, followed by Tensorflow (3S) and Pytorch (3S). Keras is a high-level library building on low-level libraries including Tensorflow, Theano and CNTK, where all the four submissions use the Tensorflow backend.

7 DISCUSSION

The F1 scores reported in SHL 2019 (the highest 78.4%) are much lower than the ones reported in SHL 2018 (the highest 93.4%) [19]. There are mainly two reasons for the performance drop. First, SHL 2018 focused on temporal-invariant evaluation while SHL 2019 considers both temporal-invariant and position-independent evaluation. All the participant teams reported that it is very challenging to train a model using the smartphone data collected at a specific body position and test the model using the data at a new position. The mismatch between the training and testing data degrades the performance significantly. Second, SHL 2018 split the data into frames of 1 minute long, while SHL 2019 uses frames of 5 seconds long. The decision has to be made per each 5-second frames, making it difficult to apply sequence modeling or temporal smoothing scheme. It has been reported in SHL 2018 that a temporal smoothing scheme within a 1-minute frame can improve the F1 performance by more than 10 percentage points over the 5-second frame.

The participant teams have employed various techniques to tackle the position-independent challenge. We summarize them as four schemes below.

Robust representation. The first scheme is to use orientation/position independent representation of the sensor data. For instance, the magnitude of sensor data, which is a combination of the data at three coordinates, has been widely used across the teams for feature computation or classifier training. The submission [6] proposed a robust representation at the three coordinates.

Exploiting target-domain data. The challenge provides a small amount of validation data collected at hand phones. Several submissions proposed to use the validation data to assist the training procedure [1, 2, 7, 10]. The submissions [1] and [2] obtain the top 2 performance among all the candidates. One challenge is that the frames in the validation set is randomly shuffled. It has been reported that a random train-test splits neglecting the temporal dependencies between the frames may lead to an upward scoring bias [19]. To cope with this issue, the submission [1] applies a order-recovering approach which aims to roughly recover the temporal orders of the frames. Since the validation set is generated using the preview version of the SHL dataset, which is made available online. The submission [7] generated their own validation set from this preview dataset (with the correct temporal information) instead of using the one given by the challenge.

Transfer learning. The third scheme is to employ transfer learning or cross-location training techniques, which train the model at source locations and generalize it to new target locations, exploiting the small amount of validation data [1, 10].

Random rotation. The last scheme is to randomly changes the orientation of the sensor data in the training set, aiming to increase the robustness of the trained model to new phone positioning. This scheme is employed in the submission [2], which takes the second place in the challenge.

8 CONCLUSION

We reported the achievements obtained during the SHL recognition challenge 2019, where 3 submissions achieved F1 scores between 70% and 80%, 5 submissions between 60% and 70%, 5 between 50% and 60%. We summarized the approaches used by these submissions and analyzed their performance. Because the approaches are implemented by different research groups with varying expertise, the conclusions drawn will be confined to the submissions of the challenge.

The submissions can be divided into ML and DL pipelines. While the advantage of deep learning has been well recognized, in this challenge DL does not show significant advantage over ML as expected. In contrast, the overall perform of ML is slightly better than DL. The highest performance is achieved by an ML approach (78.4%), which is 2.5% higher than the best DL approach (75.9%). This is possibly because hand crafted features, which incorporate human expert knowledge, play more important roles in position-independent evaluation, where the training data (torso, bag, hips) is very different from the testing data (hand).

Ensemble-based approaches in combination with a post-processing or fusion scheme tends to produce good result. For instance, the submission [1] combining RF and HMM achieves the highest F1 score 78.4%, while the submission [2] fusing multi-modal classifiers achieves the second highest F1 score 75.9%.

Various schemes have been employed by the participant teams to tackle the challenge of position-independent training, including orientation/position robust data representation, sensor data random rotation, exploiting the validation hand phone data, and transfer learning. The provides a good insight for developing novel algorithms for position-independent activity recognition.

Finally, for reference, we present the baseline performance obtained with the baseline pipeline (CNN-freq) that was employed in SHL 2018 [20]. We simply applied the same pipeline to the challenge data without fine tuning. When using the training set only for model training, we obtain an F1 score of 60.3% for the testing set. When using both the training and validation set for model training, we obtain an F1 score of 66.6%. This demonstrates that the performance can be improved effectively by incorporating the validation data for model training. The confusion matrix for the highest F1 score (66.6%) is given in Table 3.

ACKNOWLEDGEMENT

This work was supported by HUAWEI Technologies within the project "Activity Sensing Technologies for Mobile Users".

REFERENCES

[1] V. janko, M. Gjoreski, C. M. De Masi, et al. Cross-location transfer learning for the Sussex-Huawei locomotion recognition challenge. Proc. HASCA 2019.
[2] J. Choi and J. Lee. EmbraceNet for activity: A deep multimodal fusion architecture for activity recognition. Proc. HASCA 2019.
[3] Y. Zhu, F. Zhao, R. Chen. Applying 1D sensor denseNet to Sussex-Huawei locomotion-transportation recognition challenge. Proc. HASCA 2019.
[4] H. Lu, M. Pinaroc, M. Lv, S. Sun, H. Han, R. C. Shah. Locomotion recognition using XGBoost and neural network ensemble, Proc. HASCA 2019.
[5] L. Zheng, S. Li, C. Zhu, Y. Gao. Application of IndRNN for human acivity recognition - the Sussex-Huawei locomotion. Proc. HASCA 2019.

Table 2: Summary of the SHL recognition challenge 2019 result.

Approach	Rank	Team	Classifier	Input	Post-process	Sensor modality	Performance		Computational resource		Time		Implementation		Model size (MB)	Ref
							Train	Test	CPU	GPU	Train [h]	Test [s]	Lang.	Library		
ML	1	JSI-First	Random Forest	Features	HMM	LAGMOPR	83.0%	78.4%	4-core@3.6GHz RAM-16G	/	8.5	20	Python	ScikitLearn	43	[1]
	4	Jellyfish	XGBoost+ MLP	Features		AGMP	94.5%	66.9%	28-core@2.5GHz RAM-64G	5xGTX 2080	1.25	50	Python	ScikitLearn TensorFlow	40.54	[4]
	6	Gradient Descent	Classifier ensembles	Features	RNN	LAGMOPR	70.6%	64.2%	4-core@2.5GHz RAM-8G	?	6.7	556	Python	ScikitLearn TensorFlow	383.1	[6]
	7	S304	MLP ensembles	Features		AGM	74.0%	63.2%	4-core@2.8GHz RAM-8G	/	1	30	Java	AIT	0.2	[7]
	11	QMUL-IOT	XGBoost	Features		LAGMOPR	70.0%	57.3%	4-core@3.4GHz RAM-8G	/	5	600	Python	ScikitLearn	30	[11]
	12	Orion	MLP	Features		LAGMOR	77.0%	54.5%	4-core@3.4GHz RAM-16G	?	2.6	2.5	Matlab	ML Toolbox	1.44	[12]
DL	2	Yonsei-MCML	CNN	Time + Frequency	Embrace Net	LAGMOPR	60.3%	75.9%	4-core@4.2GHz RAM-32G	GTX 1080	17	1500	Python	TensorFlow	210.9	[2]
	3	We-can-fly	CNN	Time		LAGMPR	67.1%	70.3%	14-core@2.6GHz RAM-64G	TESLA V100	6	120	Python	Pytorch	11	[3]
	5	UESTC_IndRNN	RNN	Frequency		LAGMR	82.6%	66.2%	10-core@2.4GHz RAM-256G	Titan XP	3	600	Python	Pytorch	34.6	[5]
	8	Orange Lab	LSTM	Time		LA	60.4%	62.5%	8-core@2.5GHz RAM-16G	GTX 1080	5	5	Python	Pytorch	2.1	[8]
	9	GanbareAMT	LSTM	Time		AGMP	63.7%	60.0%	12-core@2.2GHz RAM-120G	TESLA P100	100.3	21	Python	Keras (Tensorflow)	0.7	[9]
	10	TDU-DSML	CNN	Frequency		AG	82.1%	58.0%	24-core@2.9GHz RAM-128G	2xGTX 1080	6.7	600	Python	Keras (Tensorflow)	10.2	[10]
	13	ICT-BUPT	CNN-LSTM	Time		LAGMOPR	76.0%	53.9%	28-core@2.4GHz RAM-64G	2x Tesla K80	17.8	132	Python	Keras (Tensorflow)	13	[13]
	14	DB	AAE	Time		AG	90.0%	31.5%	4-core@2.5GHz RAM-15G	TESLA V100	8	74	Python	Keras (Tensorflow)	431	[14]

Sensor modalitiy: L - Linear accelerometer; A - Accelerometer; G - Gyroscope; M - Magnetometer; O - Orientation; P - Pressure; R - Gravity.

[6] M. Ahmed, A. D. Antar, T. Hossain. POIDEN: position and orientation independent deep ensemble network for the classification of locomotion and transportation modes. Proc. HASCA 2019.

[7] P. Widhalm, M. Leodolter, N. Brandle. Ensemble-based domain adaptation for transport mode recognition with mobile sensors. Proc. HASCA 2019.

[8] A. Alwan, V. Frey, G. L. La. Orange labs contribution to the Sussex-Huawei locomotion-transportation recognition challenge. Proc. HASCA 2019.

[9] B. Friedich, B. Cauchi, A. Hein, S. Fudickar. Transportation mode classification from smartphone sensors via a long-short-term memory network. Proc. HASCA 2019.

[10] C. Ito, M. Shuzo, E. Maeda. CNN for human activity recognition on small datasets of acceleration and gyro sensors using transfer learning. Proc. HASCA 2019.

[11] B. Wu, J. Men, Z. Ma, W. Wu. Applying XGBoost for Sussex Huawei locomotion challenge. Proc. HASCA 2019.

[12] S. S. Saha, S. Rahman, Z. R. R. Haque, et al. Position independent activity recognition using shallow neural architecture and empirical modeling. Proc. HASCA 2019.

[13] Y. Qin, C. Wang, H. Luo. Transportation recognition with the Sussex-Huawei locomotion challenge. Proc. HASCA 2019.

[14] D. Balabka, Semi-supervised learning for human activity recognition using adversarial autoencoders. Proc. HASCA 2019.

[15] H. Xia, Y. Xiao, J. Jian, Y. Chang. Using smart phone sensors to detect transportation modes. Sensors, 14(11): 20843-20865, 2014.

[16] M. C. Yu, T. Yu, S. C. Wang, et al. Big data small footprint: the design of a low-power classifier for detecting transportation modes. Proc. Very Large Data Base Endowment, 2014, 1429-1440.

[17] H. Gjoreski, M. Ciliberto, L. Wang, F.J.O. Morales, S. Mekki, S. Valentin, D. Roggen. The university of sussex-huawei locomotion and transportation dataset for multimodal analytics with mobile devices. IEEE Access, 2018, 42592-42604.

[18] L. Wang, H. Gjoreski, M. Ciliberto, S. Mekki, S. Valentin, D. Roggen. Enabling reproducible research in sensor-based transportation mode recognition with the Sussex-Huawei dataset. IEEE Access, 2019, 10870-10891.

[19] L. Wang, H. Gjoreski, K. Murao, T. Okita, D. Roggen. Summary of the sussex-huawei locomotion-transportation recognition challenge. Proc. 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, 2018, 1521-1530.

[20] L. Wang, H. Gjoreski, M. Ciliberto, S. Mekki, S. Valentin, D. Roggen. Benchmarking the SHL recognition challenge with classical and deep-learning pipelines. Proc. 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, 2018, 1626-1635.

[21] S. Richoz, M. Ciliberto, L. Wang, P. Birch, H. Gjoreski, A. Perez-Uribe, D. Roggen. Human and machine recognition of transportation modes from body-worn camera images. Proc. Joint 8th Int. Conf. Informatics, Electronics & Vision and 3rd Int. Conf. Imaging, Vision & Pattern Recognition, 2019, 1-6.

[22] L. Wang, D. Roggen. Sound-based transportation mode recognition with smartphones. Proc. ICASSP 2019.

Table 3: Confusion matrix (F1 score) of each submission for the testing dataset. The 8 class activities are: 1 - Still; 2 - Walk; 3 - Run; 4 - Bike; 5 - Car; 6 - Bus; 7 - Train; 8 - Subway.

Ground truth class	JSI-First (78.42%)								Yonsei-MCML (75.88%)								We_can_fly (70.30%)								Jellyfish (66.88%)							
1	88	1	0	4	2	4	0	0	89	2	0	2	1	1	4	2	58	1	0	4	2	13	16	5	84	1	0	3	2	4	4	1
2	11	85	0	0	1	1	0	1	6	84	0	6	0	2	2	1	2	78	0	7	1	6	3	2	8	81	0	5	0	1	4	1
3	0	5	94	0	0	0	0	0	0	5	90	4	0	0	0	0	0	5	90	2	0	0	2	1	0	5	94	1	0	0	0	0
4	3	26	0	70	0	1	0	0	2	2	0	96	0	0	0	0	1	6	0	84	2	6	1	0	3	10	0	86	0	1	0	0
5	1	0	0	0	92	5	1	1	4	1	0	4	46	28	8	8	0	0	0	1	62	23	10	4	5	1	0	5	35	28	21	5
6	3	1	0	0	7	86	1	2	2	1	0	1	5	89	1	1	2	1	0	2	9	80	4	3	7	2	0	9	9	66	5	3
7	7	0	0	0	4	1	43	44	8	1	0	1	1	2	75	13	1	0	0	1	1	4	78	15	11	1	0	1	5	7	60	16
8	2	0	0	0	2	3	4	88	7	1	0	1	1	2	24	64	2	0	0	1	2	4	28	63	4	1	0	2	0	0	41	52
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8

Ground truth class	UESTC_IndRNN (66.20%)								Gradient_Descent (64.20%)								S304 (63.15%)								OrangeLabs (62.52%)							
1	93	1	0	0	1	2	1	2	85	1	0	4	2	3	2	2	88	1	0	1	0	8	1	1	77	1	0	3	2	1	6	9
2	9	82	0	1	1	2	0	3	5	66	0	22	1	2	2	2	8	80	0	5	0	3	2	1	3	74	1	15	1	2	3	2
3	1	7	92	0	0	0	0	0	0	3	93	4	0	0	0	0	0	6	92	1	0	0	0	0	0	2	96	2	0	0	0	0
4	4	44	0	50	1	2	0	0	2	7	2	89	0	0	0	0	2	23	1	71	0	2	0	0	1	12	3	80	0	1	1	1
5	6	1	0	0	32	44	3	14	4	4	0	5	31	39	6	12	7	0	0	2	19	58	10	3	5	2	0	18	30	32	10	4
6	9	2	0	0	8	77	2	2	5	2	0	8	11	66	2	5	6	2	0	2	5	82	3	1	1	0	0	11	8	73	5	3
7	11	0	0	0	2	3	61	23	5	1	0	1	5	5	46	37	14	1	0	0	4	16	40	26	17	0	0	2	1	4	63	12
8	7	0	0	0	1	1	19	72	4	1	0	2	1	3	12	78	4	1	0	0	3	18	8	66	18	2	0	5	9	7	23	35
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8

Ground truth class	GanbareAMT (59.96%)								TDU-DSML (57.99%)								QMUL-IoTLab (57.32%)								TeamOrion (54.51%)							
1	82	1	0	4	1	2	5	4	74	1	0	0	5	1	12	7	82	1	0	8	1	4	2	2	74	1	0	5	3	4	9	3
2	5	71	1	18	1	1	2	3	12	83	0	1	1	1	2	1	4	41	0	50	1	2	1	1	5	46	2	42	1	2	1	1
3	0	4	95	1	0	0	0	0	0	5	93	0	0	0	0	0	0	3	73	23	0	0	0	0	1	4	87	8	0	0	0	0
4	2	3	3	90	1	1	0	0	3	30	0	57	1	6	0	2	1	12	0	86	0	0	0	0	2	5	6	82	0	3	0	0
5	3	1	0	13	29	28	17	8	12	1	0	0	22	43	16	6	5	1	0	20	15	46	9	5	6	1	0	12	12	46	17	7
6	2	1	0	17	11	49	15	5	11	2	0	0	3	68	9	6	6	1	0	17	7	65	2	2	3	1	0	9	12	64	7	3
7	9	0	0	1	3	4	46	37	23	1	0	0	2	3	57	14	7	1	0	4	6	5	63	15	15	0	0	2	5	7	59	12
8	2	1	0	2	0	0	47	48	42	1	0	1	1	7	20	29	3	1	0	7	0	1	24	66	7	2	0	2	0	3	31	55
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8

Ground truth class	ICT-BUPT (53.87%)								DB (31.45%)																Baseline (66.6%) [20]							
1	78	0	0	2	0	1	7	13	37	9	0	0	2	7	37	9									83	3	0	4	1	3	4	3
2	4	41	0	40	1	5	2	8	6	76	0	0	2	5	7	2									6	79	0	9	1	2	2	2
3	0	7	82	11	0	0	0	0	0	61	10	20	7	2	0	0									0	4	71	25	0	0	0	0
4	2	1	0	92	1	3	0	2	1	22	2	29	40	4	0	2									2	10	2	84	0	1	0	0
5	6	0	0	3	14	32	20	25	5	10	0	0	14	6	44	21									4	2	0	6	45	24	10	8
6	3	0	0	4	12	57	12	12	17	4	0	0	7	33	20	17									5	3	0	8	9	71	3	1
7	17	0	0	1	1	2	59	20	20	4	0	0	1	18	40	17									7	1	0	1	1	4	63	22
8	17	0	0	2	1	3	28	50	21	6	0	0	2	11	27	33									7	2	0	2	3	3	28	56
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8									1	2	3	4	5	6	7	8

Predicted class