Summary of SHL Challenge 2023: Recognizing Locomotion and Transportation Mode from GPS and Motion Sensors

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

Hristijan Gjoreski
hristijang@feit.ukim.edu.mk
Ss. Cyril and Methodius University in Skopje, MK

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

Paula Lago
paula.lago@concordia.ca
Concordia University, Canada

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 fifth Sussex-Huawei Locomotion-Transportation (SHL) Recognition Challenge organized at the HASCA Workshop of UbiComp/ISWC 2023. 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 motion (accelerometer, gyroscope, magnetometer) and GPS (GPS location, GPS reception) sensor data of a smartphone in a user-independent manner. The training data of a train user is available from smartphones placed at four body positions (Hand, Torso, Bag and Hips). The testing data originates from test users with a smartphone placed at one, but unknown, body position. We introduce the dataset used in the challenge and the protocol of the competition. We present a meta-analysis of the contributions from 15 submissions, their approaches, the software tools used, computational cost and the achieved results. The challenge evaluates the recognition performance by comparing predicted to ground-truth labels at every 10 milliseconds, but puts no constraints on the maximum decision window length. Overall, five submissions achieved F1 scores above 90%, three between 80% and 90%, two between 70% and 80%, three between 50% and 70%, and two below 50%. While the task this year is facing the technical challenges of sensor unavailability, irregular sampling, and sensor diversity, the overall performance based on GPS and motion sensors is better than the previous four years (e.g. the best performance reported in SHL 2020, 2021 and 2023 are 88.5%, 75.4% and 96.0%, respectively). The is possibly due to the complementary between the GPS and motion sensors and also the removal of constraints on the decision window length. Finally, we present a baseline implementation to help understand the contribution of each sensor modality to the recognition task.

CCS CONCEPTS

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

KEYWORDS

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

ACM Reference Format:
Lin Wang, Hristijan Gjoreski, Mathias Ciliberto, Paula Lago, Kazuya Murao, Tsuyoshi Okita, and Daniel Roggen. 2023. Summary of SHL Challenge 2023: Recognizing Locomotion and Transportation Mode from GPS and Motion Sensors . In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2023 ACM International Symposium on Wearable Computers (UbiComp/ISWC '23 Adjunct), October 8–12, 2023, Cancun, Mexico.* ACM, New York, NY, USA, 12 pages. https://doi.org/xxxxxxx

1 INTRODUCTION

The mode of transportation or locomotion is an important contextual that enables applications such as route or parking recommendation, activity and health monitoring, individual environmental impact monitoring, and intelligent service adaptation [16–24]. Several prior works looked at recognizing modes of transportation from smartphone sensors, such as motion [25, 26], GPS [27–32], sound [33], image [34], GSM and WiFi [36, 39, 40], and the fusion of multiple sensors [35]. In particular, the combination of motion and GPS sensors

UbiComp/ISWC '23 Adjunct, October 8–12, 2023, Cancun, Mexico

ACM ISBN xxxxxxx. . . $15.00
https://doi.org/xxxxxxx

has been widely exploited for transportation mode recognition [37, 38]. 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 2018-2021 challenges [41–44], which saw 22, 14, 15 and 15 submissions, respectively, we organized the fifth Sussex-Huawei Locomotion-Transportation (SHL) recognition challenge in the year 2023[1]. The first three years (2018-2020) focused on recognition from motion sensors [45, 46], while the fourth year focused on recognition from GPS and radio signals. Being different from the previous years, the goal of this challenge is to recognize eight modes of locomotion and transportation (the activities include: being still, walking, running, cycling, driving a car, being in a bus, train or subway) in a user-independent manner based on GPS (GPS location and GPS reception) and motion (accelerometer, gyroscope, and magnetometer) sensors. This paper introduces the dataset used for the challenge and the protocol for the competition, and summarizes and analyzes the achievements of the participants contributing to the challenge.

2 DATASET AND TASK

2.1 Dataset

The challenge uses a subset of the complete Sussex-Huawei Locomotion-Transportation (SHL) dataset [47, 48]. The SHL dataset was recorded over a period of seven months in 2017 by three participants (called User1, User2 and User3) engaging in eight different modes of transportation and locomotion 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 (see Table 1). 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.

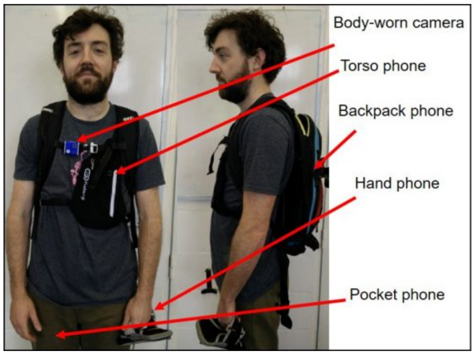


Figure 1: Smartphone positioning during data collection.

Table 1: Sensor modality of the complete SHL dataset.

Modality	SHL2018	SHL2019	SHL2020	SHL2021	SHL2023
Accelerometer	✓	✓	✓		✓
Gyroscope	✓	✓	✓		✓
Magnetometer	✓	✓	✓		✓
Linear accelerometer	✓	✓	✓		
Orientation	✓	✓	✓		
Gravity	✓	✓	✓		
Ambient pressure	✓	✓	✓		
GPS location				✓	✓
WiFi reception				✓	
GSM reception				✓	
GPS reception				✓	✓
Battery					
Ambient light					
Audio					
Video					
Google API					

The SHL Challenge 2023 provides a training, testing and validation dataset[2]. The training dataset comprises 59 days of data collected by a single "Train" user (User1), with all four smartphone positions available (Hips, Torso, Bags, Hand). The testing contains 28 days of data collected by a "Test" user with a smartphone placed at one position[3], which is unknown to the participants during the competition. The "Test" user is in reality a combination of data of User2 and User3, as none of these users could engage in all the eight activities, and this combination allows to obtain a balanced test dataset. The validation dataset contains six days of data from the four locations and from User2 and User3[4]. Fig. 2 depicts the duration of each transportation activity at one phone position in the training, validation and testing datasets. In total, we have 272×4 hours of training data, 40×4 hours of validation data and 129 hours of testing data, respectively. Fig. 3 visualizes the GPS location and trajectory of the user from the validation set.

The challenge dataset contains the raw data from three motion sensors (accelerometer, gyroscope, and magnetometer) and two GPS sensors (GPS reception and GPS location). The three motion sensors are synchronously sampled at 100 Hz. The GPS sensors are asynchronously sampled, with a sampling rate of roughly 1 Hz but is time-varying for each sensor. Depending on the condition of GPS satellite, it may happen that the GPS sensor receives no signal and thus no data is recorded. The activity label (class label) of the training and validation data is provided and synchronized with the motion sensor, with a sampling rate 100 Hz. The class label for the testing data is invisible to the participants for evaluation.

[1] http://www.shl-dataset.org/activity-recognition-challenge-2023/

[2] The exact dates for splitting the dataset will be released at the challenge website http://www.shl-dataset.org/activity-recognition-challenge-2023/.

[3] The testing position is "Hand".

[4] Note that the validation data is the same as the preview version of the SHL dataset. http://www.shl-dataset.org/download/#shldataset-preview.

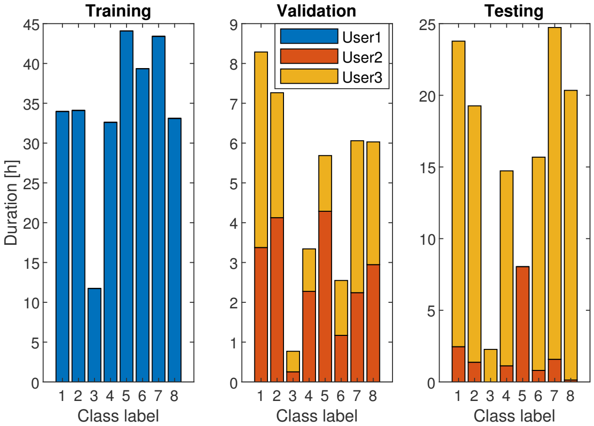


Figure 2: The duration of each class activity at one phone position in the training, validation, and testing datasets. The 8 classes are: 1 - Still; 2 - Walk; 3 - Run; 4 - Bike; 5 - Car; 6 - Bus; 7 - Train; 8 - Subway.

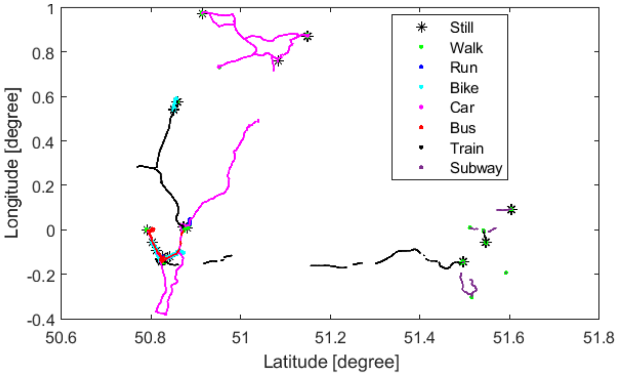


Figure 3: Visualization of the GPS location and trajectory in the complete validation dataset from User2 and User3 in 6 days.

2.2 Data Format

Tables 2, 3 and 4 list the data files provided in the training, validation and testing datasets, respectively[5]. In each sensor data file, the first column of the data contains the epoch time in ms. In addition,

- the GPS location file contains the latitude, longitude, altitude and the accuracy of the location.
- the GPS reception file contains the number of available satellites, and the SNR, the azimuth and elevation of reception;

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

Table 2: Data files provided in the training dataset (one phone position, four positions in total).

Filename	Size	Format	
		Column	Content
Acc.txt	98052438 × 4	1	Epoch time [ms]
		2	Accelerometer X [m^2/s]
		3	Accelerometer Y [m^2/s]
		4	Accelerometer Z [m^2/s]
Gyr.txt	98052438 × 4	1	Epoch time [ms]
		2	Gyroscope X [m^2/s]
		3	Gyroscope Y [m^2/s]
		4	Gyroscope Z [m^2/s]
Mag.txt	98052438 × 4	1	Epoch time [ms]
		2	Magnetometer X [m^2/s]
		3	Magnetometer Y [m^2/s]
		4	Magnetometer Z [m^2/s]
Location.txt	1088500 × 7	1	Epoch time [ms]
		2	Ignore
		3	Ignore
		4	Accuracy of this location [m]
		5	Latitude [degrees]
		6	Longitude [degrees]
		7	Altitude [m]
GPS.txt	1421690 × n_var	1	Epoch time [ms]
		2	Ignore
		3	Ignore
		4+	Variable number of entries for GPS data. If no satellite is visible the 4th column is 0. Otherwise, for each satellite visible 4 columns are added to the data file and an additional last column indicates the number of satellites. Each of the 4 columns contain in order: ID, SNR, Azimuth [degrees], Elevation [degrees]
Label.txt	98052438 × 2	1	Epoch time [ms]
		2	Label: Still=1, Walking=2, Run=3, Bike=4, Car=5, Bus=6, Train=7, Subway=8

Due to irregular sampling, the data file from each sensor has a different size. The Label data is sampled exactly at a frequency of 100 Hz. The label files at each phone position contain $98,052,438$ samples, $14,395,941$ samples, and $46,385,816$ samples, respectively for training, validation and testing. The total size of the sensor data in ASCII format is 54.8 GB, 8.11 GB and 6.42 GB for the training, validation and testing sets, respectively. The 'Label_idx.txt' file in the testing dataset provides the timestamps for which to predict the transportation mode. The participants would need to submit a plain text prediction file that contains the timestamps, as indicated in the 'Label_idx.txt' file, and the predicted labels.

SHL 2023 evaluates the recognition performance for each labeled sample, i.e. we compare ground truth to the predicted label at every sample (100 Hz) indicated in the 'Label_idx.txt' file. The challenge does not put constraints on the maximum decision window length, i.e. the sensor data samples are consecutive in time, with no segmentation and permutation applied. This is different from SHL 2018-2020, where the recognition recognition performance was evaluated with a frequency of 100 Hz, but to constrain the maximum decision window length of 1 minute (SHL 2018) or 5 seconds (SHL 2019 and 2020). This is also different from SHL 2021, where he recognition recognition performance was evaluated per

[5] More details of the data organization and file formate is available at the challenge website: http://www.shl-dataset.org/activity-recognition-challenge-2023/.

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

Table 3: Data files provided in the validation dataset (one phone position, four positions in total).

Filename	Size	Format
Acc.txt	14395941 × 4	Identical to the training data.
Gyr.txt	14395941 × 4	
Mag.txt	14395941 × 4	
Location.txt	138471 × 7	
GPS.txt	180377 × n_var	
Label.txt	14395941 × 2	

Table 4: Data files provided in the testing dataset.

Filename	Size	Format
Acc.txt	46385816 × 4	Identical to the training data.
Gyr.txt	46385816 × 4	
Mag.txt	46385816 × 4	
Location.txt	450932 × 7	
GPS.txt	600576 × n_var	
Label_idx.txt	46385816 × 1	Epoch time [ms]. The timestamps for which to predict the transportation mode.

second and without a constraint on the decision window length.

2.3 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 macro F1 score is defined as below.

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

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

3 RESULTS

Thirty-four teams expressed interest in the initial registration stage. The teams had 2.2 months (20 April - 25 June 2023) to develop the methods and work on the challenge task. Eventually, 15 teams contributed 15 submissions in the final submission stage by the deadline of 05 June[7]. Table 8 summarizes the technical details of the 15 submissions and

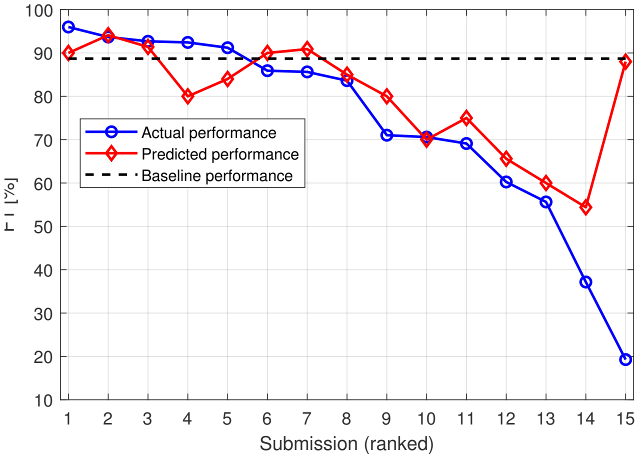


Figure 4: The submissions are ranked based on their F1 scores on the testing set (more details are given in Table 8).

Table 9 shows the detailed confusion matrices computed on the testing dataset.

3.1 Ranking

Fig. 4 depicts the F1 score of each submission for the testing set. The submissions are ranked based on their performance on the testing set (Table 8). Overall, five submissions achieved F1 scores above 90%, three between 80% and 90%, two between 70% and 80%, three between 50% and 70%, and two below 50%. In this paper, we also provide a baseline implementation that achieves an F1 score of 88.7% (see Sec. 6).

Since each team employs a distinct strategy for cross-validation, we requested each team to predict their performance for the testing dataset based on the available dataset (i.e. training and validation). We report this predicted performance, as well as the actual performance on the test set in Fig. 4. The predicted result shows that most submissions (except 4, 14 and 15) generalize well between the training/validation and the testing data, and with a difference between the actual and predicted F1 score lower than 10 percentage points. Submission 4 shows under-fitting. Submissions 14 and 15 suffer from significant over-fitting. We briefly introduce the approaches used by the top four contributions, which achieve an F1 score above 92%.

HELP takes the first place with F1 score 96.0% [1]. The submission computes 106 features in total from the five sensors (accelerometer, gyroscope, magnetometer, GPS location and GPS receiver) per 5-second data frame, and trains multiple DL (fully connected DNN) and ML (RF) models at various decision window length ranging from 5 to 240 seconds,

[7] Submissions [13] and [15] withdrew their technical paper. But we still include their results in the analysis.

and also train specific ML models to distinguish ambiguous activities, such as Car and Bus, Train and Subway. The decision at one time point is made by combining an ensemble of the abovementioned classifiers via majority voting and further improved by temporal smoothing.

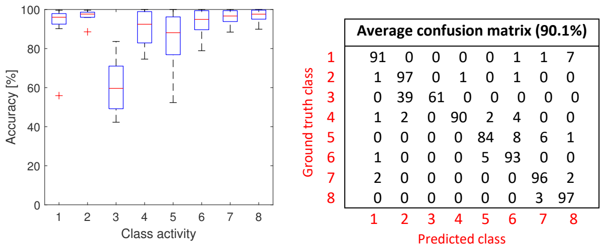
HYU-CSE achieves the second-highest F1 score of 93.7%, which is 2.3 percentage points lower than the champion [2]. The submission computes 8 GPS location features per 60-second data frame, and feeds the location feature as well as the raw data of motion sensors (accelerometer, gyroscope and magnetometer) to a dense convolutional neural network (DenseNet) to predict the transportation mode. The DenseNet receives separate inputs from the motion sensor data and the GPS feature data, and fuses them in the middle of the network, and outputs a final decision at the end of the network. The decision window slides the sensor data with a small skip size of 1 second, resulting in multiple decisions at each time point. A majority voting is thus employed to get a joint decision. The submission divides the trajectory into trips with the same transportation activity in each trip segment, and further improves the prediction accuracy by exploiting this property. The precision of trip segmentation however is not reported.

Juliet takes third place with an F1 score of 92.7% [3]. The submission divides the sensor data into 5-second windows and extracts features from the five sensors (accelerometer, gyroscope, magnetometer and GPS location and reception) per window. The submission additionally incorporates OpenStreetMap data (such as road, railway, transport, traffic, and land use) to augment the geographical information. An ensemble of ML and DL model are employed for a joint prediction. The ML approach includes XGBoost, LightGBM, and CatBoost models with hand-crafted features as input. The DL approach employs a CNN-RNN-Transformer framework with input consisting of both the raw data and hand-crafted features. The submission further improves the prediction accuracy by applying temporal smoothing within a large segment of 300 seconds.

Fighting_zsn takes fourth place with an F1 score of 92.4% [4], which is very close the third place. The submission extracts hand-crafted features from the five sensors (accelerometer, gyroscope, magnetometer and GPS location and reception) and feeds them to an ensemble of ML classifiers, including XGBoost, LightGMB and RF for a joint decision. Similarly to [3], the submission incorporates external knowledge from OpenStreetMap to enrich geographical context information. The submission employs data augmentation to address label distribution imbalance, ensuring a balanced label distribution for improved training. The submission also improves the prediction performance by applying temporal smoothing within a large window of 150 seconds.

3.2 Average Performance

As shown in Fig. 4, eight out of 15 submissions achieve F1 scores above 80%. We analyze the results from the top eight submissions (with F1 scores larger than 80%).



Average confusion matrix (90.1%)

Ground truth class

	1	2	3	4	5	6	7	8
1	91	0	0	0	0	1	1	7
2	1	97	0	1	0	1	0	0
3	0	39	61	0	0	0	0	0
4	1	2	0	90	2	4	0	0
5	0	0	0	0	84	8	6	1
6	1	0	0	0	5	93	0	0
7	2	0	0	0	0	0	96	2
8	0	0	0	0	0	0	3	97

Predicted class

Figure 5: Recognition accuracy for each class activity by the top 8 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.

Fig. 5 box-plots the recognition accuracy for each class activity (i.e. the diagonal elements of the confusion matrix in Table 9), among the top eight submissions, and also presents the average confusion matrix of their results. From the box-plot, most classes can be identified accurately except the class Run and the class Car. From the average confusion matrix, a major part of Run activity is misclassified as Walk, which is possible because of different users' behaving habits, i.e. the running speed of the testing user may be close to the walking speed of the training user. An ambiguity between Car and Bus is also observed from the confusion matrix, possibly due to their similar vibration pattern and traveling speed.

4 SUMMARY OF APPROACHES

4.1 Classical machine learning vs Deep learning

We categorize the 15 submissions into two families: classical machine learning pipeline (ML) and deep learning pipeline (DL). There are 6 ML submissions and 9 DL submissions.

Fig. 6(a) box-plots the F1 scores obtained by these two families. DL has a higher upper bound than DL, while achieving a much lower lower bound. ML has a smaller dynamic range than DL. DL achieves the top three F1 scores (96.0% [1], 93.7% [2] and 92.7% [3]), while the best ML approach (92.4% [4]) was ranked the 4th place. Fig. 6(b)-(c) show in box-plot the training and testing time by ML and DL approaches, respectively. DL takes much more time for training than ML, and also takes more time for testing.

Fig. 7 depicts the specific classifiers employed by ML and DL approaches. ML involves three classifiers: extreme gradient boost (XGBoost), random forest (RF), and ensembles of classifiers (XGBoost + RF + LGBM), where LGBM refers to light gradient-boosting machine. DL involves four types of classifiers: convolutional neural network (CNN), Long short-term memory (LSTM), Transformer, and ensembles of deep leaning and machine learning.

For classical machine learning, XGBoost has three submissions, followed by Ensembles with two submissions, and RF with one submission. Among these classifiers, Ensemble achieves the highest F1 score (92.4%), followed by XGBoost (91.2%), while RF achieves the lowest F1 score (60.2%). For

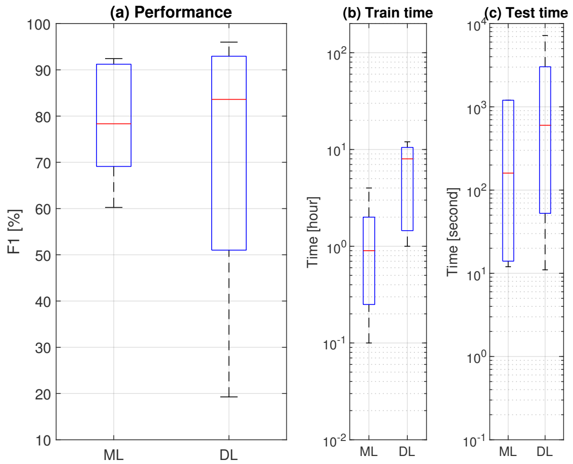


Figure 6: Comparison between machine learning and deep learning approaches. (a) F1 score for the testing data. (b) Training time. (c) Testing time.

deep learning, Ensembles (ML + DL) is the most popular classifier with four submissions, followed by CNN with three submissions, while LSTM and Transformer each have one submission. Ensembles (ML + DL) achieve the highest F1 score (96.0%), followed by CNN (93.7%) and Transformer (70.6%), while LSTM achieves the lowest F1 score (19.3%).

All six ML approaches use hand-crafted features as input to the classifier. Among the nine DL approaches, three use hand-crafted features as input to the classifier, three use raw data, and three combine raw data and hand-crafted features as input. Feature-only approaches achieve the highest F1 score (96.0%), followed by Feature-raw approaches (93.7%), while raw-only approaches achieve the lowest F1 score (83.6%).

4.2 Sensor modalities

Most submissions (13 out of 15) employ all five sensors: accelerometer (A), gyroscope (G), magnetometer (M), GPS location (L) and GPS reception (R). One submission uses only the motion data (AGM) and is ranked 8th. One submission uses only the accelerometer data (A), and is ranked 15th. It seems that the employment of all five sensors is a sensible choice.

4.3 Software Implementation

For both ML and DL, Python is the only programming language used by the submissions. For ML, Scikit-Learn (Python) is the most used library. For DL, Pytorch is the most popular library. Keras (based on Tensorflow) and Tensorflow each only have one submission.

5 DISCUSSION

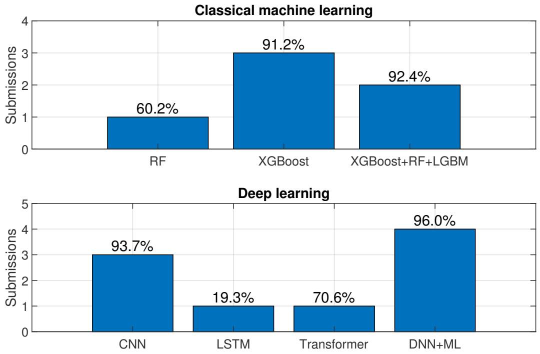


Figure 7: Classical machine learning and deep learning classifiers used by the submissions. The text on top of the bar indicates the highest F1 score achieved by each group of classifiers.

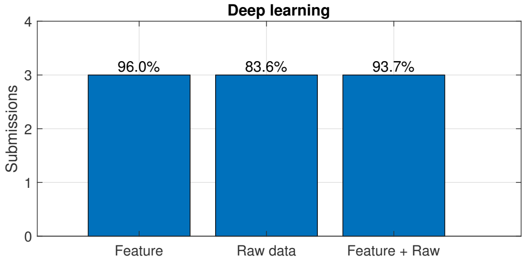


Figure 8: Type of input data to the deep-learning classifier. The text on top of the bar indicates the highest F1 score achieved by each type of input.

The overall performance based on GPS and motion sensors (SHL 2023) is much higher than the one achieved by motion sensors (SHL 2018-2020) and the one achieved by GPS and radio sensors (SHL 2021). For instance, for user-independent testing, SHL 2023 has 8 (out of 15) submissions with F1 scores above 80%, while SHL 2021 and 2020 only have 1 (out of 15) and 0 (out of 15) submissions, respectively, with F1 scores above 80%. There are mainly two reasons for the high performance of SHL 2023. First, the GPS and motion sensors provide complementary information that is beneficial to transportation mode recognition [48]. Second, SHL 2023 did not have a constraint on the maximum decision length (e.g. 5 seconds in SHL 2020), thus allowing a large decision window that can improve the transportation mode recognition performance effectively [35].

On the other hand, sensor unavailability, irregular sampling, and sensor diversity impose technical challenges to data processing. The participant teams have employed various techniques to tackle these challenges at various stages: preprocessing, feature extraction and sensor fusion, and postprocessing.

5.1 Pre-processing

The sensor data are captured by independent sensors, leading to unsynchronized timestamps in each data file, i.e. the epochtime columns are not identical in the data files. The data from the three motion sensors (accelerometer, gyroscope, magnetometer) are synchronously sampled with a sample rate 100 Hz. The data from the GPS location and reception sensors are captured with a time-varying sampling rate around 1 Hz. Depending on the environment and device, the GPS reception might not be available (e.g. underground subway) and thus the GPS sensor does not provide any data.

In SHL 2021 we were facing similar challenges of sensor unavailability and irregular sampling and the participant teams have proposed a variety of solutions, such as label matching and data interpolation to synchronize the data and fill missing sample [44]. Due to these previous achievements, most submissions in SHL 2023 are able to deal with these two challenges successfully with similar strategies. For instance, the submissions employ label alignment [1, 4, 6, 8, 9, 11, 14] to synchronize the data and the submissions [2, 3, 7] employ data interpolation to fill missing samples. One submission [5] exploits the availability of GPS signals to distinguish indoor and outdoor scenarios.

5.2 Feature engineering

The data diversity makes it important to extract suitable features for the classification task.

The motion (accelerometer, gyroscope and magnetometer) sensor can be exploited in the form of raw data or hand-crafted features. The raw motion data is typically used as input to DL classifiers, after being converted to the Euclidean magnitude [2, 3, 6, 8, 10]. Alternatively, hand-crafted features can be extracted form the raw motion sensor data and used as input to either DL or ML classifiers [1, 3–9, 11]. Due to the large variety of hand-crafted features extracted from the motion sensors, we did not summarize the details here.

The GPS sensor data is usually exploited in the form of hand-crafted features. Table 5 summarizes the GPS features used by the submissions. For GPS reception data, the basic features include the number of satellites and the SNR of each satellite.

For GPS location data, three sets of basic features are considered. One is the raw location data in terms of longitude and altitude [1–3, 6, 8, 12]. The second set includes distance, heading, velocity and acceleration derived from longitude and altitude [1–12]. From the change of distance, velocity and acceleration can be computed by differentiating the distance with respect to the time. The third set is GIS information, e.g. the distance to nearby railways and bus stops in the city [1, 3, 4, 7, 11, 12]. These submissions utilize the map information from OpenStreetmap [3, 4] or OverPass [6], an external geographic database that contains road map data on earth. Based on the phone position indicated with latitude and longitude, the closest distance to various types of roads (e.g. railway, busway, motorway, residential, pedestrian, living streets) and landmarks (e.g. railway and bus station) can be calculated.

Most submissions compute the statistical information of the basic GPS features within a time window, such as the max, min, mean and median value in a decision window period.

Table 5: GPS features used by the submissions to the SHL recognition challenge 2023.

Modality	Basic features	Submissions
GPS location	Velocity, acceleration distance, heading	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
	Longitude, altitude	[1, 2, 3, 6, 8, 12]
	GIS (distance to road type and landmarks)	[3, 4, 6, 7, 9, 11, 12]
GPS reception	Number of satellites and SNR	[1, 3, 4, 7, 11, 12]

Table 6: Multimodal fusion schemes used by the submissions to the SHL recognition challenge 2023.

Fusion scheme	Submissions
Early fusion	[1, 3, 4, 5, 6, 7, 9, 11, 12, 14]
Middle fusion	[2, 6, 10]
Late fusion	[1, 2, 3, 4, 6, 8, 9]

5.3 Sensor fusion

With different sampling rates, the motion and GPS data can be fused in different ways for the classification task. Several multimodal fusion approaches have been used, which can be categorized as early fusion, middle fusion and later fusion.

The early fusion scheme concatenates the data of all modalities as a single input vector for classification, and thus only needs a single classification model. The synchronization of multiple modalities and the handling of different data sizes and sampling rates remain challenging problems. To address this issue, the submissions employing the early fusion scheme extract hand-crafted features from the motion and GPS sensors, respectively, and concatenate them as a single input vector to the classifier [1, 3–7, 9, 11, 12, 14]. By doing this, the classifier does not need to consider inconsistent sampling rates and data size from multiple sensors.

The middle fusion scheme is typically applied to deep neural network works. The classifier receives separate input from multi-modal sensors, merges the multi-modal information in the middle of the architecture, and outputs a decision at the final layer. A few submissions employ the middle fusion scheme, where a deep neural network receives input from motion sensors and the GPS sensors, separately, and merges them in the middle of the network [2, 6, 10]. In this way, the inconsistent sampling rate of the motion and GPS sensors can be handled naturally.

The late fusion scheme trains a separate classifier for each modality independently, and draws a final decision by combining the outputs of the classifiers. In the late fusion scheme, each separate classifier is optimized, which brings additional benefits of flexibility and scalability. Rather than fusing the modalities, several submissions fuse multiple classifiers, either an ensemble of ML classifiers [4, 9], an ensemble of DL classifiers [2, 8], or the combination of ML and DL [1, 3, 6]. Since each individual classifier is already optimized with an early fusion or middle fusion scheme, the late fusion scheme tends to maximize the classification performance. For instance, the submissions [1, 3] achieve the first and third place by combining early fusion and late fusion, the submission [2] achieves the second place by combining middle fusion and late fusion.

5.4 Post-processing

Similar to SHL 2021, the challenge this year does not put constraints on the size of the decision window and the participants have the freedom to choose any window length as appropriate. Most submissions employ a two-step processing, i.e. perform prediction with a short decision window (ranging from 1 second, 5 seconds to 60 seconds) followed by a large temporal smoothing window up to 480 seconds (see details in Table 8). Since the transportation activity usually remains for a long period, the temporal smoothing can improve the prediction accuracy significantly. One submission [2] divides the sensor data into trips, each having only on transportation activity, and further improve the prediction performance by applying major voting within this trip segment. While the accuracy of trip segmentation is not reported, this submission achieves second place among all the participants.

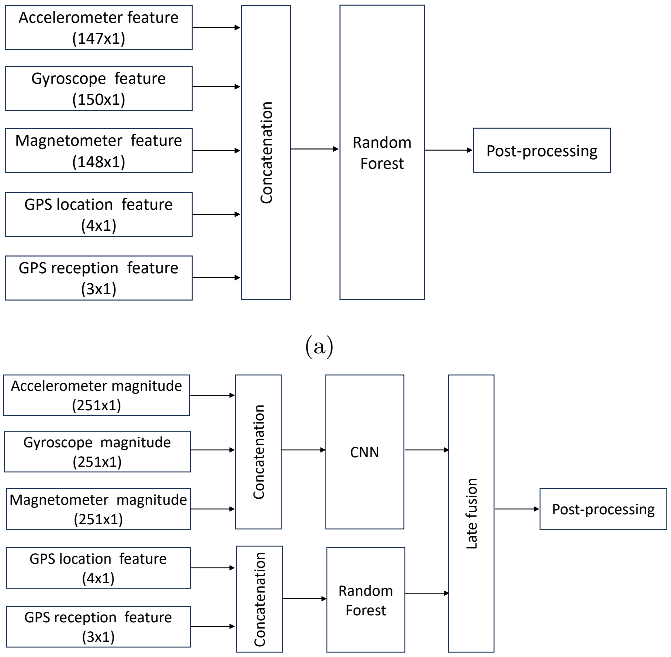
5.5 Phone location estimation

In this challenge, the position of the testing phone is unknown to the participant. It has been proved in SHL 2019 [42] that the phone position will affect the motion sensor significantly. Several submissions [5–8] designed a classifier to estimate the position of the phone so that the validation can be performed more precisely. They all estimate the phone position correctly to be “Hand”, which helps them to choose the right validation dataset. On the other hand, the GPS sensor is robust to phone placement, and can mitigate the influence on the motion sensors. The top four submissions [1–4] still achieve the best performance without knowing the phone position.

6 BASELINE IMPLEMENTATION

To investigate the contribution of each sensor to the recognition task, we present a baseline solution for SHL 2023. We use all five sensor modalities for the prediction.

In the *pre-processing* we handle the synchronization and missing data problem with the label-matching method proposed in the baseline of SHL 2021 [44]. After synchronization, the GPS data is aligned with the label with a sampling rate of 1 Hz. In the *classification* stage, we proposed two pipelines: ML and DL, as shown in Fig. 9(a) and (b), respectively. For both pipelines, we perform recognition per 5-second decision window, followed by post-processing. In the *post-processing* stage, we perform median filtering within a window of length 125 seconds, assuming a transportation activity will continue for at least 2 minutes.


Accelerometer feature (147x1)
Gyroscope feature (150x1)
Magnetometer feature (148x1)
GPS location feature (4x1)
GPS reception feature (3x1)
Concatenation
Random Forest
Post-processing

(a)

Accelerometer magnitude (251x1)
Gyroscope magnitude (251x1)
Magnetometer magnitude (251x1)
GPS location feature (4x1)
GPS reception feature (3x1)
Concatenation
Concatenation
CNN
Random Forest
Late fusion
Post-processing

(b)

Figure 9: Baseline recognition pipeline. The recognition is performed per 5-second decision window, followed by post-processing. (a). ML pipeline. (b) DL pipeline.

6.1 ML pipeline

For the ML pipeline as shown in Fig. 9(a), we compute the hand-crafted features per each sensor modality, and concatenate them into a vector before feeding into a random forest classifier. The hand-crafted features of the motion sensors are defined in [46], where we compute 147, 150 and 148 features for the accelerometer, gyroscope and magnetometer, respectively. The hand-crafted features of the GPS sensors is defined in [44], where we compute 4 and 3 features for the GPS location and reception sensors, respectively. In total, we compute 452 hand-crafted features for the recognition task in each 5-second decision window. This is an early fusion scheme. We implement the random forest classifier with the Matlab Machine Learning Toolbox, using 20 trees and setting ‘minleafsize’ to 1000. We use data from the training and validation set to train the classifier. When using an Intel i7-1165G7 4-core@2.8GHz CPU with 32G RAM, the training and testing time are 340 seconds and 3 seconds, respectively.

Table 7(a) gives the confusion matrices and F1 scores obtained by the ML pipeline combining various sensor modalities for the testing data. It can be observed that, when using a single modality, the accelerometer contributes most to the recognition task with F1 score 46.9%, followed by GPS location (42.5%), gyroscope (41.6%), magnetometer (38.6%), and GPS reception (26.5%). The combination of GPS location and reception can only increase the F1 score from 42.5% (location only) to 44.5% (both). The combination of accelerometer and gyroscope can increase the F1 score to 57.1%, with additional magnetometer increasing F1 score to 63.8%. The combination of the five sensors further increases the F1 score to 68.3%, which can be improved to 76.0%

Table 7: Confusion matrix (F1 score) of the baseline pipeline for the testing dataset, with different combinations of sensor modalities. (a) ML pipeline. (b) DL baseline. Key: A - accelerometer; G: gyroscope; M - magnetometer; L - GPS location; R - GPS reception; The 8 class activities are: 1 - Still; 2 - Walk; 3 - Run; 4 - Bike; 5 - Car; 6 - Bus; 7 - Train; 8 - Subway.

Ground truth class

A (46.9%)	1	2	3	4	5	6	7	8
1	71	1	0	1	8	5	8	8
2	1	78	0	9	5	4	1	1
3	0	19	61	19	0	0	0	0
4	0	0	0	24	64	11	0	0
5	7	3	0	2	12	14	56	7
6	9	1	0	4	24	46	14	2
7	14	0	0	1	12	12	55	5
8	28	0	0	1	19	7	34	11

G (41.6%)	1	2	3	4	5	6	7	8
1	57	5	0	2	7	2	23	4
2	1	74	0	22	2	0	0	0
3	0	45	33	22	0	0	0	0
4	0	45	0	42	11	1	0	0
5	3	4	0	5	77	6	4	0
6	14	3	0	2	39	22	19	2
7	17	2	0	1	22	7	48	4
8	28	2	0	1	19	3	43	4

M (38.6%)	1	2	3	4	5	6	7	8
1	54	8	0	14	7	12	3	1
2	3	42	1	28	2	3	11	12
3	3	8	17	64	0	0	8	0
4	12	18	0	45	5	6	8	6
5	4	7	1	5	38	22	19	5
6	13	7	1	12	31	25	10	2
7	2	10	0	4	13	7	40	24
8	1	7	0	3	5	3	33	48

A+G (57.1%)	1	2	3	4	5	6	7	8
1	79	3	0	1	5	2	5	6
2	3	82	0	13	2	0	0	0
3	0	31	62	7	0	0	0	0
4	1	1	0	67	30	1	0	0
5	8	5	0	2	66	2	15	2
6	8	1	0	3	28	41	16	3
7	16	1	0	1	12	12	53	5
8	24	1	0	1	18	6	36	15

A+G+M (63.8%)	1	2	3	4	5	6	7	8
1	82	2	0	2	3	4	4	2
2	3	79	0	15	1	0	0	0
3	0	36	60	4	0	0	0	0
4	2	1	0	73	22	2	0	1
5	4	4	0	0	55	20	11	4
6	8	1	0	3	30	50	7	2
7	9	1	0	1	10	6	55	18
8	5	1	0	1	8	1	37	48

L (42.5%)	1	2	3	4	5	6	7	8
1	54	6	0	0	0	2	7	29
2	7	85	2	1	0	2	1	1
3	0	84	12	3	0	1	0	0
4	2	38	40	17	0	3	0	0
5	3	3	0	5	63	17	9	0
6	7	4	1	12	18	56	1	0
7	3	2	1	4	40	15	32	4
8	1	1	0	2	23	16	10	47

R (26.5%)	1	2	3	4	5	6	7	8
1	24	30	0	7	10	2	13	14
2	1	79	0	7	9	2	2	1
3	0	70	0	10	14	6	0	0
4	0	86	0	5	9	1	0	0
5	0	44	0	13	38	5	1	0
6	0	46	0	9	32	9	3	0
7	6	10	0	5	36	15	24	4
8	2	3	0	5	21	9	12	49

L+R (44.5%)	1	2	3	4	5	6	7	8
1	71	5	0	0	0	1	8	14
2	7	87	2	1	0	1	1	1
3	0	86	11	2	0	0	0	0
4	2	44	36	16	0	1	0	0
5	2	3	0	7	70	14	2	0
6	7	6	1	20	20	46	1	0
7	5	2	0	4	37	15	33	4
8	1	1	0	2	22	16	9	49

A+G+M+L+R (68.3%)	1	2	3	4	5	6	7	8
1	83	1	0	2	3	5	6	1
2	3	82	0	12	1	0	1	0
3	0	38	59	3	0	0	0	0
4	1	0	0	74	19	5	0	0
5	4	3	0	1	67	11	15	0
6	8	1	0	3	26	55	7	0
7	8	1	0	1	13	4	69	3
8	5	1	0	0	5	1	37	50

Post (76.0%)	1	2	3	4	5	6	7	8
1	90	2	0	1	3	2	2	0
2	1	91	0	8	0	0	0	0
3	0	43	57	0	0	0	0	0
4	0	0	0	80	19	1	0	0
5	0	1	0	0	86	9	4	0
6	2	1	0	1	24	72	0	0
7	4	0	0	0	8	6	82	0
8	1	0	0	0	1	0	48	49

Predicted class

(a)

Ground truth class

DL: A+G+M (73.1%)	1	2	3	4	5	6	7	8
1	82	3	0	1	2	2	8	3
2	4	85	0	8	1	1	1	1
3	0	35	60	5	0	0	0	0
4	6	3	0	73	7	10	1	1
5	5	3	0	1	67	15	6	4
6	4	1	0	1	13	74	5	2
7	6	1	0	0	4	3	70	16
8	4	0	0	0	3	1	29	62

ML: L+R (44.5%)	1	2	3	4	5	6	7	8
1	71	5	0	0	0	1	8	14
2	7	87	2	1	0	1	1	1
3	0	86	11	2	0	0	0	0
4	2	44	36	16	0	1	0	0
5	2	3	0	7	70	14	2	0
6	7	6	1	20	20	46	1	0
7	5	2	0	4	37	15	33	4
8	1	1	0	2	22	16	9	49

Fusion: A+G+M+L+R (77.0%)	1	2	3	4	5	6	7	8
1	88	2	0	0	1	1	5	2
2	4	90	0	5	0	1	0	0
3	0	42	55	3	0	0	0	0
4	2	7	0	77	5	8	0	0
5	3	3	0	1	75	11	5	2
6	3	1	0	1	13	77	4	1
7	4	1	0	0	4	3	77	11
8	3	0	0	0	2	1	26	68

Post (88.7%)	1	2	3	4	5	6	7	8
1	94	2	0	0	0	0	2	0
2	1	98	0	1	0	0	0	0
3	0	41	59	0	0	0	0	0
4	0	3	0	90	6	1	0	0
5	0	1	0	0	96	2	0	0
6	1	0	0	0	6	92	0	0
7	2	1	0	0	1	1	94	1
8	1	0	0	0	0	0	23	77

Predicted class

(b)

through post-processing. The confusion matrix obtained after post-processing shows certain ambiguities between Walk vs Run, Car vs Bus and Train vs Subway.

6.2 DL pipeline

For the DL pipeline as shown in Fig. 9(b), we use two classifiers to fuse the multimodal sensor data efficiently. For the three motion sensors, we concatenate the raw data (Euclidean magnitude) of each sensor into a vector before fed into a CNN classifier. The Euclidean magnitude is defined in [46], where we convert the 500 raw data points in a 5-second decision into a magnitude vector of length 251. In total we have a vector of length 753 for the three motion sensors. For the two GPS sensors, we compute the hand-crafted features as defined in [44], where we have 4 and 3 features for the GPS location and reception sensors. The hand-crafted features were fed to a random forest classifier, with the same parameter as the one in the ML pipeline. Finally, we use a product-based ensemble scheme, as defined in [35] to fuse the decision form the CNN and RF classifiers. This is a late-fusion scheme. The CNN classifier is implemented with the Matlab Deep Learning Toolbox. We use data from the training and validation set to train the classifier. When using a GeForce GTX 1080 Ti GPU with 3584 CUDA cores@1.58 GHz and 11 GB memory, the training and testing time are 2.8 hours and 10 seconds, respectively.

Table 7(a) gives the confusion matrices and F1 scores obtained by the ML pipeline combining various sensor modalities for the testing data. It can be observed that the CNN combining three motion sensors achieve an F1 score of 72.3%, and the RF combining two GPS sensors achieves an F1 score of 44.5%. The fusion of the two classifiers can increase the F1 score to 77.0%, which is further improved to 88.7% by post-processing. The confusion matrix obtained after post-processing shows certain ambiguities between Walk vs Run and Train vs Subway. The baseline performance (88.7%) locates between the fifth (91.2%) and sixth place (85.9%) of submissions.

Table 7 gives the confusion matrices and F1 scores obtained by combining various sensor modalities for the testing data. It can be observed that, when using a single modality, GPS location contributes most to the recognition task with F1 score 45.4%, followed by GPS reception (27.8%), WiFi (17.7%) and Cell (15.8%). The combination of WiFi and Cell only increases the F1 score slightly to 20.0%. The combination of GPS reception with other sensors can increase the recognition performance effectively. For instance, the F1 score increases from 45.4% to 51.3% when combining GPS reception with GPS location; and increases from 20.0% to 33.6% when combining GPS reception with WiFi and Cell. The combination of the four sensors leads to the highest F1 score of 54.9%, which is slightly lower than the submission that ranks in the 7th position (56.6%). The performance can be further improved by incorporating statistical features and by exploiting the temporal dependencies in transportation activities.

7 CONCLUSION

We reported the achievements obtained during the SHL recognition challenge 2023, where five submissions achieved F1 scores above 90%, three between 80% and 90%, two between 70% and 80%, three between 50% and 70%, two below 50%. 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. DL has a higher upper bound than DL, while achieving a much lower lower bound. On the other hand, ML has a smaller dynamic range than DL. DL achieves the top three F1 scores (96.0%, 93.7% and 92.7%), while the best ML approach (92.4%) was ranked the 4th place.

While the task this year is facing the technical challenges of sensor unavailability, irregular sampling, and sensor diversity, the overall performance based on GPS and motion sensors is better than the previous four years. The is possibly due to the complementary between the GPS and motion sensors and also the removal of constraints on the decision window length. Various hand-crafted features were extracted from the GPS and motion sensors. An external GIS dataset has been widely used to augment geographical information. Three types of fusion schemes (early, middle and late fusion) were employed to integrate the information from multimodal sensors. The late fusion scheme, in the form of an ensemble of ML and DL classifiers, has shown the best performance.

ACKNOWLEDGEMENT

The dataset used in this work was supported by HUAWEI Technologies within the project "Activity Sensing Technologies for Mobile Users". This work received additional support from the EU H2020-ICT-2019-3 project "HumanE AI Net" (project number 952026).

REFERENCES

[1] S. Oh, et al. Multimodal Sensor Data Fusion and Ensemble Modeling for Human Locomotion Activity Recognition. Proc. UbiComp/ISWC 2023.

[2] S. Huang, et al. User-independent Motion and Location Analysis for Sussex-Huawei Locomotion Data. Proc. UbiComp/ISWC 2023.

[3] M. Li, et al. Enhanced SHL Recognition Using Machine Learning and Deep Learning Models with Multi-source Data. Proc. UbiComp/ISWC 2023.

[4] Y. Zhao, et al. Road Network Enhanced Transportation Mode Recognition with an Ensemble Machine Learning Model. Proc. UbiComp/ISWC 2023.

[5] L. Alecci, et al. Enhancing XGBoost with heuristic smoothing for transportation mode and activity recognition. Proc. UbiComp/ISWC 2023.

[6] R. Sekiguchi, et al. Ensemble learning using motion sensors and location for human activity recognition. Proc. UbiComp/ISWC 2023.

[7] J. Deng, et al. Enhancing Locomotion Recognition with Specialized Features and Map Information via XGBoost. Proc. UbiComp/ISWC 2023.

[8] T. Hyugagi, et al. Moving state estimation by CNN from long time data of smartphone sensors. Proc. UbiComp/ISWC 2023.

[9] Z. Zeng, et al. An Ensemble Framework Based on Fine Multi-Window Feature Engineering and Overfitting Prevention for Transportation Mode Recognition. Proc. UbiComp/ISWC 2023.

[10] R. Chen. Enhancing Transportation Mode Detection using Multi-scale Sensor Fusion and Spatial-topological Attention. Proc. UbiComp/ISWC 2023.

[11] G. Habault, et al. A Classical Machine Learning Method for Locomotion and Transportation Recognition using both Motion and Location Data. Proc. UbiComp/ISWC 2023.

[12] S. Huang, et al. A post-processing machine learning for activity recognition challenge with OpenStreetMap data. Proc. UbiComp/ISWC 2023.

[13] TeamX. Withdrawn.

[14] H. Yan, et al. AttenDenseNet for the Sussex-Huawei Locomotion-Transportation (SHL) recognition challenge. Proc. UbiComp/ISWC 2023.

[15] TeamY. Human activity recognition using accelerometer, gyroscope and magnetometer with LSTM. Withdrawn.

[16] J. Engelbrecht, M.J. Booysen, G. Rooyen, et al. Survey of smartphone-based sensing in vehicles for intelligent transportation system applications. IET Intelligent Transport Systems, 9(10): 924-935, 2015.

[17] Y. Vaizman, K. Ellis, G. Lanckriet. Recognizing detailed human context in the wild from smartphones and smartwatches. IEEE Pervasive Computing, 16(4): 62-74, 2017.

[18] E. Anagnostopoulou, J. Urbancic, E. Bothos, et al. From mobility patterns to behavioural change: leveraging travel behaviour and personality profiles to nudge for sustainable transportation. Journal of Intelligent Information Systems, 2018: 1-22, 2018.

[19] D.A. Johnson, M.M. Trivedi. Driving style recognition using a smartphone as a sensor platform. Proc. IEEE Conf. Intelligent Transportation Systems, 2011, 1609-1615.

[20] W. Brazil, B. Caulfield. Does green make a difference: The potential role of smartphone technology in transport behaviour. Transportation Research Part C: Emerging Technologies, 37: 93-101, 2013.

[21] J. Froehlich, T. Dillahunt, P. Klasnja, et al. Ubigreen: Investigating a mobile tool for tracking and supporting green transportation habits. Proc. SIGCHI Conf. Human Factors Computing Systems, 2009, 1043-1052.

[22] N.D. Lane, E. Miluzzo, H. Lu, et al. A survey of mobile phone sensing. IEEE Communications Magazine, 48(9): 140-150, 2010.

[23] S. C. Mukhopadhyay. Wearable sensors for human activity monitoring: A review. IEEE Sensors Journal, 15(3): 1321-1330, 2015.

[24] G. Castignani, T. Derrmann, R. Frank, T. Engel. Driver behavior profiling using smartphones: A low-cost platform for driver monitoring. IEEE Intelligent Transportation Systems Magazine. 7(1): 91-102, 2015.

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

[26] 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.

[27] Y. Zheng, L. Liu, L. Wang, X. Xie. Learning transportation mode from raw GPS data for geographic applications on the Web. Proc. Int. Conf. World Wide Web, 2008, 247-256.

[28] L. Stenneth, O. Wolfson, P. S. Yu, B. Xu. Transportation mode detection using mobile phones and GIS information. Proc. ACM SIGSPATIAL Int. Conf. Advances in Geographic Information Systems, 2011, 54-63.

[29] X. Yang, K. Stewart, L. Tang, et al. A review of GPS trajectories classification based on transportation mode. Sensors, 18(11): 3741, 2018.

[30] L. Gong, T. Morikawa, T. Yamamoto, H. Sato. Deriving personal trip data from GPS data: A literature review on the existing methodologies. Procedia-Social and Behavioral Sciences, 138: 557-565, 2014.

[31] S. Dabiri, K. Heaslip. Inferring transportation modes from GPS trajectories using a convolutional neural network. Transportation Research Part C: Emerging Technologies, 86: 360-371, 2018.

[32] M. Guo, S. Liang, L. Zhao, P. Wang. Transportation mode recognition with deep forest based on GPS data. IEEE Access, 8: 150891-150901, 2020.

[33] L. Wang, D. Roggen. Sound-based transportation mode recognition with smartphones. Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2019, 930-934.

[34] S. Richoz, M. Ciliberto, L. Wang, et al. 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, 67-72.

[35] S. Richoz, L. Wang, P. Birch, D. Roggen. Transportation mode recognition fusing wearable motion, sound and vision sensors. IEEE Sensors Journal, 20(16): 9314-9328, 2020.

Table 8: Summary of the SHL recognition challenge 2023.

App.	Rank	Team	Classifier	Decision window	Post processing	Input	Sensor modality	Fusion scheme	Performance		Computational resource		Time		Implementation		Model size (MB)	Ref
									Predict	Test	CPU	GPU	Train [h]	Test [s]	Lang.	Library		
ML	4	Fighting_zsn	XGBoost + RF + LGBM	1 s	150 s	Features	AGMLR	early+late	80.0%	92.4%	32-core@2.1GHz RAM-173G	/	1.3	/	Python	Scikit-learn	/	[4]
	5	MUSIC	XGBoost	40 s	480 s	Features	AGMLR	early	84.0%	91.2%	10-core@3.70GHz RAM-128G	/	2	14	Python	Scikit-learn	9.2	[5]
	7	WinGPT	XGBoost	60 s	300 s	Features	AGMLR	early	90.9%	85.6%	/	/	0.1	12	Python	Scikit-learn	18.7	[7]
	9	ZZL	XGBoost + RF + LGBM	1 s	225 s	Features	AGMLR	early+late	80.0%	71.0%	32-core@2.5GHz RAM-128G	/	4	300	Python	Scikit-learn	30.4	[9]
	11	KDDI	XGBoost	9 s	N	Features	AGMLR	early	75.0%	69.1%	20-core@2.20GHz RAM-4226G	/	0.25	20	Python	Scikit-learn	1.58	[11]
	12	DataScience	RF	1 s	adaptive	Features	AGMLR	early	0.6564	0.6024	16-core@2.5GHz	/	0.5	1200	Python	Scikit-learn	20	[12]
DL	1	HELP	FC+RF	5 s	240 s	Features	AGMLR	early+late	0.9	0.9599	8-core@3.5GHz RAM-128GB	RTX 3090	4	3600	Python	Scikit-learn + Pytorch	127	[1]
	2	HYU-CSE	CNN	60 s	trip segment	Features + Raw data	AGML	middle +late	94.1%	93.7%	8-core@3.8GHz RAM-32G	RTX 3080	1.6	2840	Python	Scikit-learn + Pytorch	4.2	[2]
	3	Juliet	CNN-RNN-Transformer + XGBoost + LGBM + CatBoost	5 s	300 s	Features + Raw data	AGMLR	early + late	91.4%	92.7%	16-core@2.10GHz RAM 188G	Titan V	10	11	Python	Scikit-learn + Pytorch	14	[3]
	6	TDU_BSA	CNN + XGBoost	5 s	480 s	Features + Raw data	AGML	early + middle + late	90.0%	85.9%	16-core@3.2GHz RAM-64G	RTX 3090	1	60	Python	Scikit-learn + Tensorflow	186	[6]
	8	Ds	CNN	110 s	No	Raw data	AGM	late	85.0%	83.6%	/	Google Colab Pro	1	30	Python	Pytorch	5	[8]
	10	we-can-fly	CNN + Transformer	5 s	No	Raw data	AGMLR	middle	70.0%	70.6%	i5-7500 CPU @ 3.40GHz	RTX 3070	8	600	Python	Pytorch	173	[10]
	13	TeamX	CNN + XGBoost + RF	60 s	N	Features	AGMLR	early	60.0%	55.6%	i7-10510U CPU @ 1.80GHz 2.30 GHz RAM-16GB	/	/	120	Python	Scikit-learn + Keras	/	[13]
	14	Yummy MacMuffin	CNN	5 s	N	Raw data	AGMLR	early	54.4%	37.2%	16-core @ 2.50GHz RAM-128G	Tesla P100	12	600	Python	Pytorch	2.2	[14]
	15	TeamY	LSTM	50 S	N	Features	A	early	88.0%	19.3%	8-core, RAM-32G	?	8	7200	Python	Scikit-learn + Keras	10	[15]
Sensor modalitiy: A - Accelerometer; G - Gyroscope; M - Magnetometer; L - GPS location; R - GPS reception.																		

[36] T. Sohn, A. Varshavsky, A. LaMarca, et al. Mobility detection using everyday gsm traces. Proc. Int. Conf. Ubiquitous Computing, 2006, 212-224,

[37] T. Feng and H. J. Timmermans, Transportation mode recognition using GPS and accelerometer data. Transportation Research Part C: Emerging Technologies, 37: 118-130, 2013.

[38] B. D. Martin, V. Addona, J. Wolfson, G. Adomavicius, and Y. Fan. Methods for real-time prediction of the mode of travel using smartphone-based GPS and accelerometer data. Sensors 17(9): 2058, 2017.

[39] H. Wang, F. Calabrese, G. D. Lorenzo, C. Ratti. Transportation mode inference from anonymized and aggregated mobile phone call detail records. Proc. Int. Conf. Intelligent Transportation Systems, 2010, 318-323.

[40] V. C. Coroama, C. Turk, F. Mattern. Exploring the usefulness of bluetooth and wifi proximity for transportation mode recognition. Adjunct Proc. 2019 ACM Int. Joint Conf. Pervasive and Ubiquitous Computing and Proc. 2019 ACM Int. Symp. Wearable Computers, 2019, 37-40.

[41] L. Wang, H. Gjoreski, K. Murao, T. Okita, D. Roggen. Summary of the Sussex-Huawei locomotion-transportation recognition challenge. Proc. 2018 ACM Int. Joint Conf. and 2018 Int. Symp. Pervasive and Ubiquitous Computing and Wearable Computers, 2018, 1521-1530.

[42] L. Wang, H. Gjoreski, M. Ciliberto, P. Lago, K. Murao, T. Okita, D Roggen. Summary of the Sussex-Huawei locomotion-transportation recognition challenge 2019. Adjunct Proc. 2019 ACM Int. Joint Conf. Pervasive and Ubiquitous Computing and Proc. 2019 ACM Int. Symp. Wearable Computers, 2019, 849-856.

[43] L. Wang, H. Gjoreski, M. Ciliberto, P. Lago, K. Murao, T. Okita, D Roggen. Summary of the Sussex-Huawei locomotion-transportation recognition challenge 2020. Adjunct Proc. 2020 ACM Int. Joint Conf. Pervasive and Ubiquitous Computing and Proc. 2020 ACM Int. Symp. Wearable Computers, 2020, 351-358.

[44] L. Wang, M. Ciliberto, H. Gjoreski, P. Lago, K. Murao, T. Okita, D Roggen. Locomotion and transportation Mode Recognition from GPS and radio signals: Summary of SHL Challenge 2021. Adjunct Proc. 2021 ACM Int. Joint Conf. Pervasive and Ubiquitous Computing and Proc. 2021 ACM Int. Symp. Wearable Computers, 2021, 412-422.

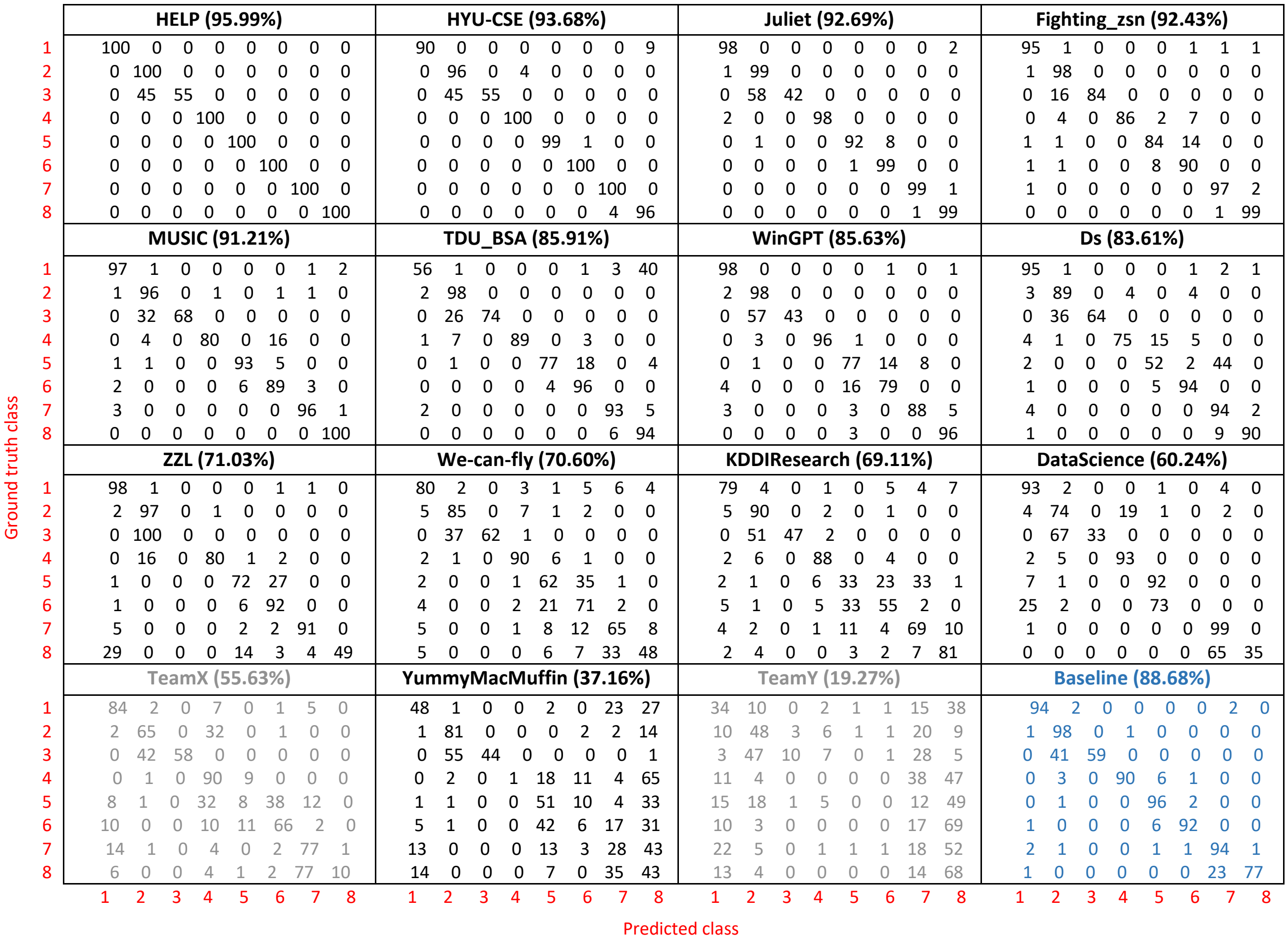
[45] L. Wang, H. Gjoreski, M. Ciliberto, S. Mekki, S. Valentin, D. Roggen. Benchmarking the SHL recognition challenge with classical and deep-learning pipelines. Proc. 2018 ACM Int. Joint Conf. and 2018 Int. Symp. Pervasive and Ubiquitous Computing and Wearable Computers, 2018, 1626-1635.

[46] L. Wang, H. Gjoreski, M. Ciliberto, P. Lago, K. Murao, T. Okita, D Roggen. Three-year Review of the 2018-2020 SHL Challenge on Transportation and Locomotion Mode Recognition from Mobile Sensors. Frontiers in Computer Science, 3 (713719), 1-24, 2021.

[47] 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.

[48] 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.

Table 9: 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

	HELP (95.99%)								HYU-CSE (93.68%)								Juliet (92.69%)								Fighting_zsn (92.43%)							
1	100	0	0	0	0	0	0	0	90	0	0	0	0	0	0	9	98	0	0	0	0	0	0	2	95	1	0	0	0	1	1	1
2	0	100	0	0	0	0	0	0	0	96	0	4	0	0	0	0	1	99	0	0	0	0	0	0	1	98	0	0	0	0	0	0
3	0	45	55	0	0	0	0	0	0	45	55	0	0	0	0	0	0	58	42	0	0	0	0	0	0	16	84	0	0	0	0	0
4	0	0	0	100	0	0	0	0	0	0	0	100	0	0	0	0	2	0	0	98	0	0	0	0	0	4	0	86	2	7	0	0
5	0	0	0	0	100	0	0	0	0	0	0	0	99	1	0	0	0	1	0	0	92	8	0	0	1	1	0	0	84	14	0	0
6	0	0	0	0	0	100	0	0	0	0	0	0	0	100	0	0	0	0	0	0	1	99	0	0	1	1	0	0	8	90	0	0
7	0	0	0	0	0	0	100	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	99	1	1	0	0	0	0	0	97	2
8	0	0	0	0	0	0	0	100	0	0	0	0	0	0	4	96	0	0	0	0	0	0	1	99	0	0	0	0	0	0	1	99

	MUSIC (91.21%)								TDU_BSA (85.91%)								WinGPT (85.63%)								Ds (83.61%)							
1	97	1	0	0	0	0	1	2	56	1	0	0	0	1	3	40	98	0	0	0	0	1	0	1	95	1	0	0	0	1	2	1
2	1	96	0	1	0	1	1	0	2	98	0	0	0	0	0	0	2	98	0	0	0	0	0	0	3	89	0	4	0	4	0	0
3	0	32	68	0	0	0	0	0	0	26	74	0	0	0	0	0	0	57	43	0	0	0	0	0	0	36	64	0	0	0	0	0
4	0	4	0	80	0	16	0	0	1	7	0	89	0	3	0	0	0	3	0	96	1	0	0	0	4	1	0	75	15	5	0	0
5	1	1	0	0	93	5	0	0	0	1	0	0	77	18	0	4	0	1	0	0	77	14	8	0	2	0	0	0	52	2	44	0
6	2	0	0	0	6	89	3	0	0	0	0	0	4	96	0	0	4	0	0	0	16	79	0	0	1	0	0	0	5	94	0	0
7	3	0	0	0	0	0	96	1	2	0	0	0	0	0	93	5	3	0	0	0	3	0	88	5	4	0	0	0	0	0	94	2
8	0	0	0	0	0	0	0	100	0	0	0	0	0	0	6	94	0	0	0	0	3	0	0	96	1	0	0	0	0	0	9	90

	ZZL (71.03%)								We-can-fly (70.60%)								KDDIResearch (69.11%)								DataScience (60.24%)							
1	98	1	0	0	0	1	1	0	80	2	0	3	1	5	6	4	79	4	0	1	0	5	4	7	93	2	0	0	1	0	4	0
2	2	97	0	1	0	0	0	0	5	85	0	7	1	2	0	0	5	90	0	2	0	1	0	0	4	74	0	19	1	0	2	0
3	0	100	0	0	0	0	0	0	0	37	62	1	0	0	0	0	0	51	47	2	0	0	0	0	0	67	33	0	0	0	0	0
4	0	16	0	80	1	2	0	0	2	1	0	90	6	1	0	0	2	6	0	88	0	4	0	0	2	5	0	93	0	0	0	0
5	1	0	0	0	72	27	0	0	2	0	0	1	62	35	1	0	2	1	0	6	33	23	33	1	7	1	0	0	92	0	0	0
6	1	0	0	0	6	92	0	0	4	0	0	2	21	71	2	0	5	1	0	5	33	55	2	0	25	2	0	0	73	0	0	0
7	5	0	0	0	2	2	91	0	5	0	0	1	8	12	65	8	4	2	0	1	11	4	69	10	1	0	0	0	0	0	99	0
8	29	0	0	0	14	3	4	49	5	0	0	0	6	7	33	48	2	4	0	0	3	2	7	81	0	0	0	0	0	0	65	35

	TeamX (55.63%)								YummyMacMuffin (37.16%)								TeamY (19.27%)								Baseline (88.68%)							
1	84	2	0	7	0	1	5	0	48	1	0	0	2	0	23	27	34	10	0	2	1	1	15	38	94	2	0	0	0	0	2	0
2	2	65	0	32	0	1	0	0	1	81	0	0	0	2	2	14	10	48	3	6	1	1	20	9	1	98	0	1	0	0	0	0
3	0	42	58	0	0	0	0	0	0	55	44	0	0	0	0	1	3	47	10	7	0	1	28	5	0	41	59	0	0	0	0	0
4	0	1	0	90	9	0	0	0	0	2	0	1	18	11	4	65	11	4	0	0	0	0	38	47	0	3	0	90	6	1	0	0
5	8	1	0	32	8	38	12	0	1	1	0	0	51	10	4	33	15	18	1	5	0	0	12	49	0	1	0	0	96	2	0	0
6	10	0	0	10	11	66	2	0	5	1	0	0	42	6	17	31	10	3	0	0	0	0	17	69	1	0	0	0	6	92	0	0
7	14	1	0	4	0	2	77	1	13	0	0	0	13	3	28	43	22	5	0	1	1	1	18	52	2	1	0	0	1	1	94	1
8	6	0	0	4	1	2	77	10	14	0	0	0	7	0	35	43	13	4	0	0	0	0	14	68	1	0	0	0	0	0	23	77
	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

Predicted class