Classical Machine Learning Approach for Human Activity Recognition Using Location Data

Safaeid Hossain Arib
University of Dhaka
Dhaka, Bangladesh
safaeid48@gmail.com

Rabeya Akter
University of Dhaka
Dhaka, Bangladesh
rabeyaakter231023@gmail.com

Omar Shahid
University of Dhaka
Dhaka, Bangladesh
omarshahid232@gmail.com

Md Atiqur Rahman Ahad
University of Dhaka
Dhaka, Bangladesh
atiqahad@du.ac.bd

ABSTRACT

The Sussex-Huawei Locomotion-Transportation (SHL) recognition Challenge 2021 was a competition to classify 8 different activities and modes of locomotion performed by three individual users. There were four different modalities of data (Location, GPS, WiFi, and Cells) which were recorded from the phones of the users in their hip position. The train set came from user-1 and the validation set and test set were from user-2 and user-3. Our team **'GPU Kaj Kore Na'** used only location modality to give our predictions in test set of this year's competition as location data was giving more accurate predictions and the rest of the modalities were too noisy as well as not contributing much to increase the accuracy. In our method, we used statistical feature set for feature extraction and Random Forest classifier to give prediction. We got **validation accuracy of 78.138%** and a **weighted F1 score of 78.28%** on the SHL Validation Set 2021.

CCS CONCEPTS

• **Activity Recognition**; • **Machine Learning**; • **Feature Extraction**; • **Classifier**;

KEYWORDS

Feature extraction, classical approach, classifier, SHL recognition challenge 2021

ACM Reference Format:
Safaeid Hossain Arib, Rabeya Akter, Omar Shahid, and Md Atiqur Rahman Ahad. 2021. Classical Machine Learning Approach for Human Activity Recognition Using Location Data. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers (UbiComp-ISWC '21 Adjunct), September 21–26, 2021, Virtual, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3460418.3479376

UbiComp-ISWC '21 Adjunct, September 21–26, 2021, Virtual, USA

ACM ISBN 978-1-4503-8461-2/21/09…$15.00
https://doi.org/10.1145/3460418.3479376

1 INTRODUCTION

Human Activity Recognition (HAR) is the task of identifying human motion and locomotion. There have been quite a few works over the years in this field. Most of the works were based on the body-worn sensor (gyroscope, accelerometer)-related data [3]. Due to significant improvements in technology like smartphones and smartwatches, these data can be collected and used to recognize various human motions or activities (walking, running, swimming, cycling, etc) through the inbuilt app [6]. People can track their activity during the day with the technology they have on themselves easily. The companies providing these services have a huge collection of data some of them are publicly available. There was not much previous work on activity recognition using Location, WiFi, Cells, and GPS data. Farhan et al. [1] had reviewed previous works on different modalities of data for activity recognition. The main purpose is to predict the human activities from the sensors' time-series data, which were recorded at a specific polling rate and labeled to make proper prediction.

Mobile sensor data-based human activity recognition has a lot of use cases. It can discover activity patterns and the variables determining the patterns, get information about a person- their personality and psychological state. It can help in the medical sector by monitoring health by analyzing a person's activity and designing individualized exercise tables to improve well-being [2, 4]. Moreover, it might be very helpful in case of user experience analysis or daily life monitoring by avoiding extra use of device as it could predict activity through only mobile sensors. Sussex-Huawei Locomotion-Transportation (SHL) recognition challenge 2021 is the fourth edition of the SHL challenge that had a dataset- very different from the last three years. In this paper, we developed a activity recognition method for SHL Challenge 2021. We used a traditional machine learning approach to fit the given data for predicting the activities. Figure 1 describes the main skeleton of our method.

We described the specifications of this year's dataset in Section 2 In Section 3, we demonstrated our method in details containing feature extraction method and classifier. Section 4 contained our result and a brief discussion on result and its possibilities. Finally, the conclusion was drawn in Section 5.

diagram.pdf

Figure 1: Main skeleton of our method

2 DATASET OVERVIEW

In Sussex-Huawei Locomotion-Transportation (SHL) recognition Challenge 2021, the goal was to recognize eight modes of locomotion and transportation- 1) Still, 2) Walking, 3) Run, 4) Bike, 5) Car, 6) Bus, 7) Train, 8) Subway - by using radio data that included GPS reception, GPS location, WiFi reception, and GSM cells tower scans. SHL dataset [5, 8] 2021 was divided into three parts- SHL Train Set 2021, SHL Validation Set 2021, and SHL Test Set 2021. SHL Train Set 2021 contained data from a phone located at the hips position of user-1 only for 59 days. SHL-Validation Set 2021 contained data from a phone as well and located at the hips position of user-2 and user-3 for 4 days. On the contrary, SHL-Test Set 2021 comprised of data from user-2 and user-3 for 39 days through a phone at same body position. In Figure 2, we showed the percentage of the timestamps count of the Label files from the train set, validation set, and test set to get an idea about the distribution of the dataset.

Validation Test Data Distribution.pdf

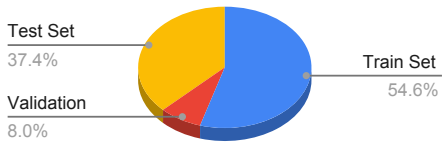


Figure 2: Distribution of the Dataset

SHL Train Set 2021, SHL Validation Set 2021, SHL Test Set 2021- all three of the sets have data regarding - Location, GPS, WiFi, and Cells. Among them, GPS, WiFi, and Cells have variable numbers of data in each timestamp. The Location has a fixed number of data in every timestamp. While collecting data for the challenge, all the sensors were asynchronously sampled with a sampling rate of roughly 1 Hz. But for each sensor, the sampling rate was time-varying. Also, depending on the condition of the satellite and cell, one sensor may not receive any signal at a certain interval and thus no data could be found for that interval. For this reason, Location, GPS, WiFi, and Cells contained the different numbers of total timestamps. In Figure 3 and Figure 4, class-wise data distribution is shown for the train set and validation set respectively.

Data Distribution of Train set.pdf

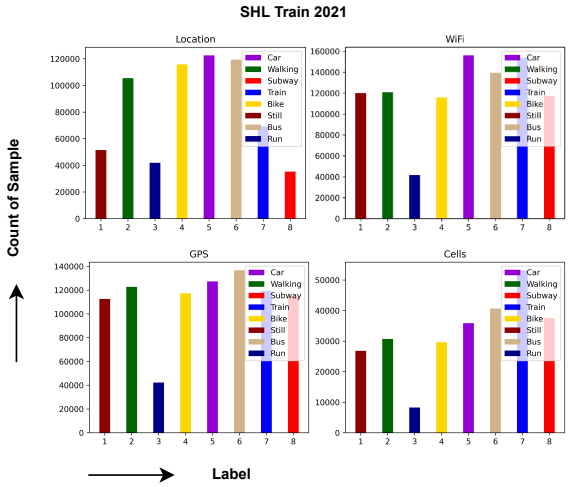


Figure 3: Class-wise Data Distribution of Train Set

Data Distribution of Validation set.pdf

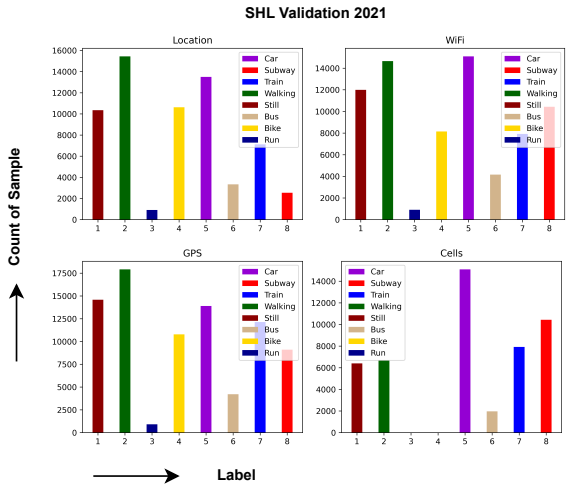


Figure 4: Class-wise Data Distribution of Validation Set

3 METHODOLOGY

SHL 2021 dataset contained sensor data- Location, GPS, WiFi, and Cells. We matched the labels of the separate sensor data of timestamps and trained different models individually as well as combination of different modalities. Initially after label matching, we tried sliding window method to organize the data. Then, we extracted some meaningful features from the given Location data. We have got our best result while applying a traditional machine learning algorithm to the extracted feature. For prediction purpose, we used data interpolation if there was no location data for that instance.

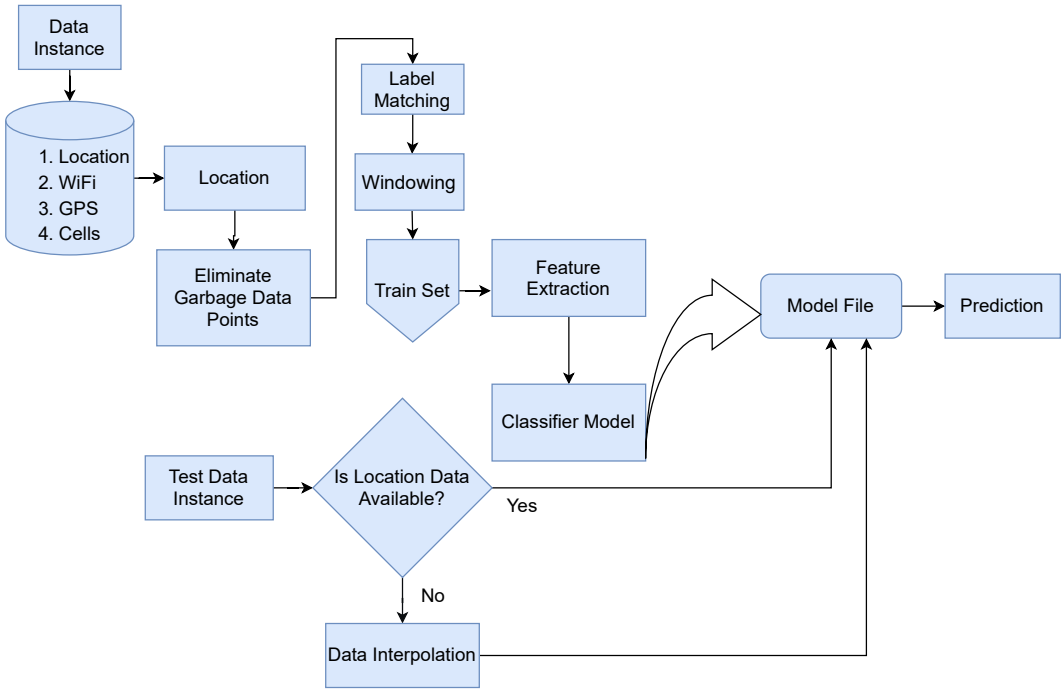


Figure 5: Block Diagram of Our Method

3.1 DATA PRE-PROCESSING

The dataset contained sensor data files and label files separately. So we had to label the sensor data files. We pre-processed the data in the following step:

Label Matching: We matched the label depending on the Epoch time [ms] feature in the files. As the sensors were asynchronously sampled, the Epoch time[ms] columns were not identical in the files. In the Label file, in between every timestamp(t) and the next one of that timestamp(t+1), if we found any timestamp

$$t <= s < t + 1 \tag{1}$$

in the GPS, WiFi, and Cells files; we labeled that timestamp (s) of GPS, WiFi and Cells file with the given Label of the Label file's considered timestamp (t). The timestamp that was unlabeled while following the technique was dropped.

3.2 FEATURE EXTRACTION

We have exploited two features: Haversine distance and average speed. All the statistical features mentioned in Table 1 below were extracted using the window selection method as a part of feature extraction.

Haversine Distance: Using the latitude and longitude between two consecutive timestamps, we calculated the Haversine distance. The first timestamp had a null value due to the method of calculation. All the null value was filled using the mean of the column containing Haversine distance. The formula used for calculating Haversine distance:

$$d = 2r \arcsin \sqrt{(\sin^2(\frac{\alpha 2 - \alpha 1}{2}) + \cos(\alpha 1)\cos(\alpha 2)\sin^2(\frac{\beta 2 - \beta 1}{2})} \tag{2}$$

Here, $\alpha 1$ = Latitude of point 1 (in radians), $\alpha 2$= Latitude of point 2 (in radians), $\beta 1$= Longitude of point 1 (in radians), $\beta 2$= Longitude of point 2 (in radians)

Speed: Using the Haversian distance between the two timestamps and the time delta average speed between two consecutive timestamps was calculated. The formula for average speed:

$$Speed = \frac{d}{t} \tag{3}$$

Here, d = Haversian distance and t = Time difference between two points

Window Selection:

All the features in the dataset and the added features were segmented into the sliding window of an empirically chosen size of l samples with overlapping of r. Form M data samples in the training set, we get N windows.

$$N = \frac{M - 1}{l \times (r - 1)} + 1 \tag{4}$$

Window length of 30 data points and overlapping of 0% were used to extract the following statistical features mentioned in the table below from all the features of the dataset except the timestamp or "Epoch time[Ms]". For only the feature "Epoch time[ms]" or the time stamp we only extracted the minimum, maximum, standard deviation, mean, and variance was extracted using a window length of 30 data points and 0% overlapping.

Table 1: The name of the features and the number of each feature extracted from a window.

Features	Numbers
Minimum	7
Maximum	7
Standard Deviation	7
Average	7
Variance	7
Peak to Peak Range	6
Max Rate of Change	6
Average Rate of Change	6
Standard Deviation of Rate of Change	6
Mean Absolute Deviation	6
Inter-Quartile Range	6
Autocorrelation	6
Mean Crossing Rate	6
Linear Velocity	6

3.3 Classifier

We used the Random Forest Classifier to fit the training data using these hyper parameters: n estimators=300, min samples split=2, verbose=0, alpha=0.

3.4 Prediction

From the test data, we checked if there is any location data present. We took the available location data points and preprocessed the test set in a similar way as the training set. Using the window length of 30 data points, we predicted the label of those data points. After generating the prediction, we used the predicted label for each data points to fill the next 30 data points using the same label to match the length of the test set before preprocessing the test set using the window length of 30 data points. We had to predict the labels for the given epoch times. There were data missing for a lot of these epoch times that we had to predict the labels for. So we used the model of the predicted label to fill the epoch times for which we could not get any data points to give a prediction.

4 RESULTS AND DISCUSSION

To give the final prediction on test data, we used the result which we got using location data only. Table 2 demonstrate all of our experimented results.

Data Distribution of Test set.pdf

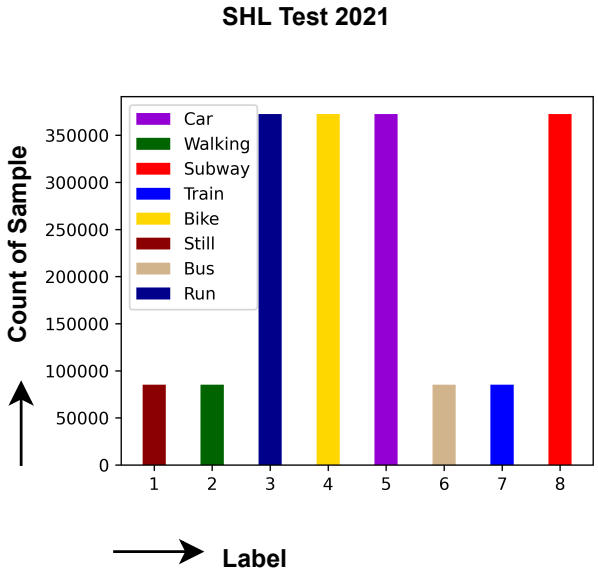


Figure 6: Predicted Class-wise Data Distribution of Test Data

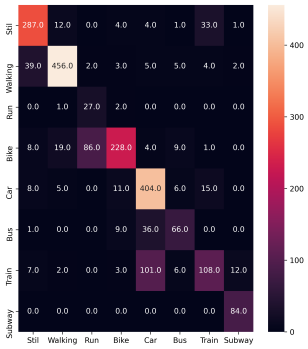
Table 2: Final Result

Modalities	Accuracy
GPS	32.97%
Wifi	30.99%
Location	**78.14%**
GPS + Location	75.56%

The dataset had a lot of missing labels and noise. The data was collected at 1 Hz rate but for a single timestamp, not all the sensor data were registered together, some were missing. That caused the data to have sudden skips in timestamps. These issues made this dataset challenging. The test data had some timestamps where we were asked to make a prediction but no sensor data was recorded for that timestamp making this dataset even more challenging. We have cleaned the data and extracted the features with a window length of 30 data point with no overlapping.

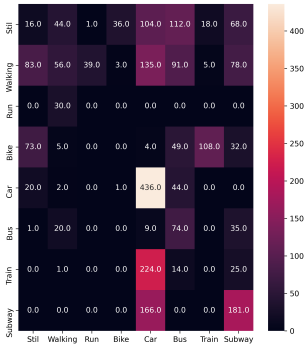
The training set and validation set had a class imbalance, so our final model was biased for some activity. We used up-sampling to see if it reduced the bias but it didn't help much. Label 3 had the lowest amount of training data and seeing the confusion matrix we can see that its F1 score is 36 percent which means it's being predicted poorly. Label 6 and 7 is being poorly recognized too. The rest of the labels had a high F1 score meaning the model is predicting them well. From the figure below we can see that label 5 has the highest number of instances and label 3 having the lowest matching up with the training and validation set instances.

Figure 7 is the confusion matrix using location data that we used in final prediction of test data. Figure 8, 9, and 10 represents the confusion matrix of Wifi, GPS, and combination of Wifi and GPS modality respectively.



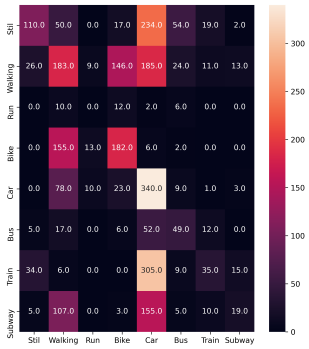
matrix-Location.pdf

Figure 7: Confusion matrix using Location



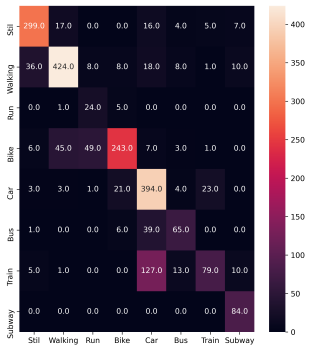
matrix-WIFI.pdf

Figure 8: Confusion matrix using WiFi



matrix-GPS.pdf

Figure 9: Confusion matrix using GPS



matrix-GPS+Location.pdf

Figure 10: Confusion matrix using GPS and Location

5 CONCLUSION

A straightforward machine learning approach has been proposed in this paper to recognize the 8 activities of SHL Challenge 2021. Our final proposed model only used the location data to classify the activities which resulted in a decent score. The rest of the data modalities (WiFi, Cells, GPS) could be used to improve the accuracy of prediction further through using more smart and generic features. Deep learning can also be used to generalize and improve prediction accuracy. Our model is performing decently on the validation set, and we expect expect that it might work better in the SHL Test Set 2021. Though things may not work as per as our plan, because we have used data interpolation method to predict timestamps where it do not have any location data, but we are still hopeful with our results. The recognition result for the testing dataset will be presented in the summary paper of the challenge [7].

REFERENCES

[1] Farhan Fuad Abir, Md Ahasan Atick Faisal, Omar Shahid, and Mosabber Uddin Ahmed. 2021. Contactless Human Activity Analysis: An Overview of Different Modalities. *Contactless Human Activity Analysis* 200 (2021), 83.

[2] Md Atiqur Rahman Ahad, Anindya Das Antar, and Omar Shahid. 2019. Vision-based Action Understanding for Assistive Healthcare: A Short Review.. In *CVPR Workshops*. 1–11.
[3] Anindya Das Antar, Masud Ahmed, and Md Atiqur Rahman Ahad. 2019. Challenges in Sensor-based Human Activity Recognition and a Comparative Analysis of Benchmark Datasets: A Review. In *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. IEEE, 134–139.
[4] Md Ahasan Atick Faisal, Md Sadman Siraj, Md Tahmeed Abdullah, Omar Shahid, Farhan Fuad Abir, and MAR Ahad. 2020. A pragmatic signal processing approach for nurse care activity recognition using classical machine learning. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 396–401.
[5] H. Gjoreski, M. Ciliberto, L. Wang, F.J.O. Morales, S. Mekki, S. Valentin, and D. Roggen. 2018. The University of Sussex-Huawei locomotion and transportation dataset for multimodal analytics with mobile devices. *IEEE Access* 6 (2018), 42592—-42604.
[6] Md Sadman Siraj, Md Ahasan Atick Faisal, Omar Shahid, Farhan Fuad Abir, Tahera Hossain, Sozo Inoue, and Md Atiqur Rahman Ahad. 2020. UPIC: user and position independent classical approach for locomotion and transportation modes recognition. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 340–345.
[7] L. Wang, M. Ciliberto, H. Gjoreski, P. Lago, K. Murao, T. Okita, and D. Roggen. 2021. Locomotion and transportation Mode Recognition from GPS and radio signals: Summary of SHL Challenge 2021. In *Proceedings of the 2021 ACM International Joint Conference and 2021 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*.
[8] L. Wang, H. Gjoreski, M. Ciliberto, S. Mekki, S. Valentin, and D. Roggen. 2019. Enabling reproducible research in sensor-based transportation mode recognition with the Sussex-Huawei dataset. *IEEE Access* 7 (2019), 10870—-10891.

A APPENDIX

A.1 Features Used

Minimum, Maximum, Standard Deviation, Average, Variance, Peak To Peak Range, Max Rate of Change, Average Rate of Change, Standard Deviation of Rate of Change, Mean Absolute Deviation, Inter-Quartile Range, Auto-correlation, Mean Crossing Rate, Linear Velocity

A.2 Programming Language and Libraries Programming language

Programming language: Python
Python Libraries: Numpy, Pandas, Matplotlib, Scikit-learn

A.3 Window size and Post processing

Window size: 30 data points
Post-processing: N/A

A.4 Used Notebook Specification

CPU Model Name: Intel(R) Xeon(R)
CPU Freq.: 2.30GHz
No. CPU Cores: 2
CPU Family: Haswell
RAM: 25.46GB
GPU: N/A
Disk Space: 25GB

A.5 Training and Testing Time

Training: 30 minutes
Testing: 10 minutes