Bullet points on Literature review on Urban Human mobility - Data collection:

On literature collection:

The **main scope of the literature study** is to find *how the Wearable sensor data is collected, processed (pre-processing, feature selection/engineering/extraction)) and predicted (Classifier - Machine Learning) according to current literature? to apply it on an offline largely unlabelled data to accurately classify vertical mobility activities and derive the vertical displacement*. In the pursuit of this, literature is spread through many interconnected and sometimes disparate fields of research. They are also subdivided into application in medical, transportation, etc.

Fields of study - Human Activity recognition; Transport mode identification; Indoor navigation, detection, and localization; Urban human mobility; vertical transportation.

Papers are found majorly through **google scholar,** and other major journals such as SENSOR, IEEE, Science direct, etc. In google scholar, most cited and relevant papers for all time, since 2015 and since 2018 are used as filters. All the keywords used are documented.

Human activity recognition is a large field of research concerning the usage of wearable and ambient sensors to detect different activities including gestures and transportation. Transportation mode identification is a part of transportation research that includes detecting motorized and non-motorized modes such as walking, cycling, etc. Urban human mobility includes both indoor activity recognition and transportation mode recognition. Vertical transportation research concerns the use of stairs, elevators, and escalators in both private and public places to study energy consumption, mode of choice, etc. Indoor - outdoor navigation and localization also uses a barometer to detect the path of travel, vertical mobility mode, floors, vertical height traversed and localize them inside buildings, train stations, etc.

A note on the use of barometer for activity/mobility recognition:

In total, 127 highly cited papers were collected in the field of Human activity recognition. Of which **54 papers used vertical displacement**, in the form of climbing upstairs/downstairs and elevator as an activity class to be classified.

Of the 54 papers, **only 2 papers used barometer as a sensor**. None of both used them to derive the magnitude of a vertical jump or estimate altitude but to simply help classify whether a vertical mobility event took place.

On researching for more papers on the specific **use of barometer as a sensor** to identify activity recognition, **53 papers** were found that has sufficient citation to be deemed important.

These papers are not the top cited (most papers are since 2015) or most conventional in their usage of barometer for activity/mobility recognition.

The most **common sensor** used to identify vertical mobility is 3-axis accelerometer, after which, gyroscope and magnetometer are used in much lesser frequency.

In the 53 papers, **the barometer is used for** activity recognition (including vertical displacement), transport mode classification, elderly or sick fall detection, floor identification, indoor mapping, indoor localization, vehicle tracking, etc.

More detailed review on the use of barometer to classify vertical displacement event and other activity/mobility will follow.

On literature closer to our works:

There is no paper that comprises the exact work we are trying to achieve with the resources at our disposal and the application of our result. Some papers are relevant in understanding methodologies, others come closer to the type of activities we are trying to recognize (vertical mobility), type of data (offline, largely unlabelled, low frequency, diverse population, activities of daily living, etc.), type of sensor (wearable, barometer, etc.). The papers identified to be of closer to our work is tagged 'closest works' and found in Mendeley. This collection also includes the most representative papers that strongly contribute to our current works.

On Data collection and annotation techniques in activity recognition:

AR: Activity recognition

Data collection and annotation techniques are studied mindful of the sensor used, feature extraction, recognition techniques, and their validation techniques, as all stages of the HAR inform each other.

The **sampling frequency** in AR, generally varies between 10 Hz to 100 Hz, with few exceptions of using even higher (250 Hz). No article in this survey topic uses less than 1 Hz. In transportation research, which uses WiFI, GPS, Cellular data, the sampling frequency varies between 1Hz (1 sec) to 0.008 Hz (2 mins). Our sampling frequency is 0.0625 Hz (16 sec).

Public datasets are very commonly used and often repeated. Of the 69 papers in AR that provide accurate information about their data, 25 papers used public datasets. They are generally done in controlled environments, although they claim to be realistic, semi-naturalistic, etc.

The **nature of the datasets** collected in the literature is generally in a controlled environment. It varies between, 'told what to do and how', 'told what to do but not how', 'implicit instructions on what to do but hiding the true nature', 'Stealth observation of subjects', 'observation through video'. Truly realistic data borders on the latter called as the 'free-living conditions', where a subject's ground truth is annotated from videos taken from camera subtly attached to their watch or packets, while they go about their activities of daily living (ADL) for a 24 hour period.

The process of **annotation techniques/ground truth** recording shapes the research done in HAR. The different ways to do that are subject self-report, direct observation (researcher in the lab, researcher following the subject in the real world), indirect observation (video, GPS traces). The difficulty of obtaining ground truth puts a cap on the many literature's research output.

The problem of **class imbalance** in a dataset comes often, as it is found in both public and private datasets. Data which is tiring or difficult to obtain always forms the underrepresented portion of the data contributing to data imbalance. The problem with class imbalance can be offset by oversampling the minority class or undersampling the majority class, using SMOTE algorithm, etc. Many researchers tackled the problem with class imbalance by simply using F1-score to report classification accuracy that takes into factor the different sizes of classes.

The **generalization of the data** to new users, devices, etc is widely studied and have been taken into consideration in the majority of the literature. This is generally solved by including diverse physical characteristics of subjects and literature often reports the mean and standard deviation of age, BMI, height, weight, etc. along with physical ailments.

Size of the data set used for training and testing varies between 15 mins to 2.5 million hours for 1 to 100,000 people while the average data is 6 hours long with 10 subjects. The number of data points might be significantly high on average because of the high sampling frequency generally used. The closest data size to our work contains data from 100,000 people for over 24 hours each.

Validation procedures are the bridge that connects the training data to test data and provides convincing evidence on the workability of a method on the intended data set. To prove that a particular recognition method is user-independent, device-independent, orientation-independent, location-independent, etc., one or more of the parameter (such as the user) is left out of training and used as test data. The resulting validation proves that the model works even on users that the device was not trained on.

Recommendations for our work on data collection and annotation techniques:

Our dataset consists of **activities of daily living**, meaning a continuous time series data of people in their natural daily living. When we collect segmented (Each continuous experiment contains different activities or only a subset of all activities) data for training, it is difficult to prove that such training data is representative of the test data (NSE database). This can be solved by using segmented data for training and continuous data of daily living (at least over 8 hours for different people) for testing, as done in some literature.

NSE database consists of data from school students of age 7 to 21. Classification of data with different characteristics can be proved to work by collecting validation data from people of different physical characteristics through the validation technique leave-one-user/characteristic-out.

Validation procedure hence provides a practical way to prove the working of our model on unlabeled NSE data by only collecting small but representative validation data, which usually forms 20 to 30% of the total training data.