CHAPTER-2

BACKGROUND AND PREREQUISITES

2.1 Types Of Human Activity Recognition

There are three types of HAR techniques that are commonly used and are illustrated in **Figure 2.1**. These are:

- **Sensor-based HAR:** Sensor-based HAR involves using data collected from wearable sensors such as accelerometers, gyroscopes, or magnetometers to recognize human activities. The data from these sensors is typically analyzed using machine learning algorithms to identify patterns that correspond to specific activities.
- **Vision-based HAR:** Vision-based HAR involves using computer vision algorithms to analyze visual data captured by cameras to recognize human activities. This type of HAR can be used for applications such as surveillance, safety monitoring, or robotics.
- **Fusion-based HAR:** Fusion-based HAR involves combining data from multiple sources such as wearable sensors, cameras, or other sensors to improve the accuracy of activity recognition. For example, combining data from both wearable sensors and cameras can provide a more complete understanding of a person's activities.

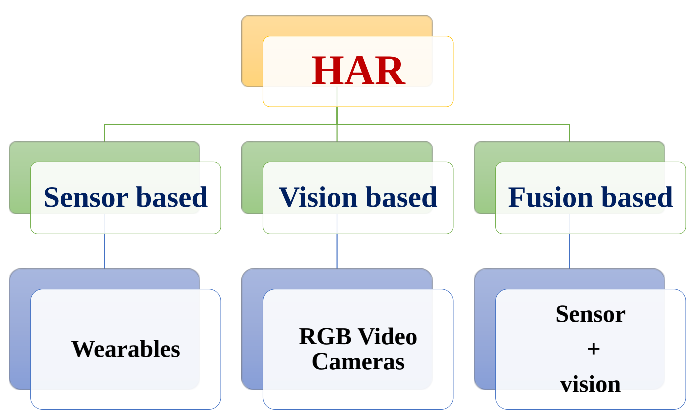

Figure 2.1 Types of HAR.

2.1.1. Human Activity Recognition with Wearable Sensors

Sensor-based HAR is a technique that involves using data collected from wearable sensors to recognize human activities. These sensors may include accelerometers, gyroscopes, magnetometers, or even heart rate monitors, and the data collected from them is typically analyzed using machine learning algorithms to identify patterns that correspond to specific activities. The process typically involves collecting sensor data from a group of people performing various activities, then using this data to train a machine learning model to recognize those activities. Once the model has been trained, it can be used to identify activities in real-time by analyzing the sensor data from a wearable device worn by a person. Sensor-based HAR has a wide range of applications, including healthcare, sports performance monitoring, and even in fields like law enforcement and defense. In healthcare, it can be used to monitor the physical activity of patients and help healthcare professionals track their progress. In sports performance monitoring, it can be used to measure the physical activity of athletes during training and competition to help optimize their performance. In law enforcement and defense, it can be used to monitor the physical activity of personnel and detect potential threats.

Sensor-based HAR is a powerful technique that has been shown to be effective in many applications. However, it requires careful consideration of the type of sensors to use, the placement of those sensors, and the machine learning algorithms used to analyze the data. Overall, sensor-based HAR is an exciting area of research with many potential applications, and it is likely to become increasingly important as wearable technology continues to evolve and become more sophisticated.

2.1.2. Human Activity Recognition with Vision-based HAR

Vision-based human activity recognition (HAR) involves using computer vision algorithms to analyze visual data captured by cameras to recognize human activities. This type of HAR can be used for applications such as surveillance, safety monitoring, or robotics.

The process typically involves collecting video data of people performing various activities, then using this data to train a machine learning model to recognize those activities. Once the model has been trained, it can be used to identify activities in real-time by analyzing the video data from a camera or a series of cameras.

Vision-based HAR can be achieved through several techniques, including:

- **Object detection:** Object detection involves identifying and localizing specific objects in the video frame, such as a person, a vehicle, or a ball. Once the objects have been detected, machine learning algorithms can be used to recognize the specific activities being performed by those objects.
- **Pose estimation:** Pose estimation involves estimating the 2D or 3D position of a person in the video frame and identifying the specific activities they are performing. This technique is useful for applications such as sports performance monitoring or rehabilitation.
- **Action recognition:** Action recognition involves directly recognizing the specific actions being performed by people in the video frame. This technique is useful for applications such as surveillance or human-robot interaction.

Vision-based HAR has a wide range of applications, including surveillance, safety monitoring, robotics, and even gaming. It can be used to detect and prevent criminal activity, detect potential safety hazards in industrial environments, enable robots to interact with humans more naturally, and more. However, vision-based HAR requires careful consideration of the cameras to use, the quality of the visual data, and the machine learning algorithms used to analyze the data.

2.1.3. Fusion-based HAR

Fusion-based HAR involves combining data from multiple sources, such as wearable sensors, cameras, or other sensors, to improve the accuracy of activity recognition. The idea behind fusion-based HAR is to take advantage of the complementary information that can be obtained from different sensors, which can help to reduce the impact of noise and improve the overall accuracy of the recognition. Fusion based HAR can be achieved through several techniques, including:

- **Sensor fusion:** Sensor fusion involves combining data from multiple wearable sensors, such as accelerometers, gyroscopes, or magnetometers, to improve the accuracy of activity recognition. This technique can be useful for applications such as health monitoring or sports performance analysis.
- **Vision-sensor fusion:** Vision-sensor fusion involves combining data from cameras and wearable sensors to improve the accuracy of activity recognition. For example, combining data from a camera and an accelerometer can help to more accurately recognize the activity of a person walking or running.
- **Feature fusion:** Feature fusion involves combining different types of features extracted from the sensor data to improve the accuracy of activity recognition. For example, combining features extracted from both accelerometer data and heart rate data can help to recognize the activity of a person jogging more

accurately.

Fusion-based HAR has a wide range of applications, including healthcare, sports performance monitoring, and robotics. It can be used to monitor the physical activity of patients, improve the accuracy of sports performance analysis, and enable more sophisticated human-robot interaction more accurately. However, fusion based HAR requires careful consideration of the type of sensors to use, the quality of the data, and the machine learning algorithms used to analyze the data.

2.2 Sensor modality

Although HAR approaches can be applied universally across various sensor modalities, the majority are tailored to specific sensor types. As outlined in [1], these sensor modalities can be broadly categorized into three main aspects: body-worn sensors, object sensors, and ambient sensors. A summary of these modalities [13] is presented in Table 1 below.

Table 2.1 Sensor modalities for HAR tasks.

Aspect	Sensor Modality	Description
Body-Worn Sensors	Accelerometer	Measures acceleration in three axes (X, Y, Z).
	Gyroscope	Measures angular velocity in three axes.
	Magnetometer	Measures the Earth's magnetic field.
	Inertial Measurement Unit	Combines accelerometer, gyroscope, and often magnetometer data.
	Heart Rate Monitor	Measures heart rate and provides physiological data.
	Skin Conductance Sensor	Measures changes in skin conductivity.
	Respiration Sensor	Monitors respiratory rate and patterns.
	Electroencephalogram	Records electrical activity in the brain.
	Electromyogram	Records muscle activity and contractions.
	Pedometer	Counts steps taken by the user.

Object Sensors	Camera	Captures visual data, often used for gesture recognition.
	Microphone	Captures audio data, useful for speech recognition.
	RFID Reader	Reads data from RFID tags for tracking purposes.
Ambient Sensors	Light Sensor	Measures ambient light levels.
	Temperature Sensor	Measures ambient temperature.
	Pressure Sensor	Measures atmospheric pressure.
	Proximity Sensor	Detects the presence of nearby objects.
	Global Positioning System	Provide location data (latitude, longitude, altitude).

2.3 Human Activity Recognition Process

The procedure involved in recognizing human activities are quite similar to those of a general-purpose pattern recognition system which involves data gathering to action recognition. With the goal of developing an efficient classification model HAR, a series of transformations are applied to the raw data that is retrieved from sensors. The HAR approach for smartphones with inertial sensors can be classified into two categories depending on machine learning approaches: deep algorithms (e.g., CNN, RNN, RBM, SAE, DFN, and DBM) and shallow algorithms (e.g., SVM, KNN, and decision tree). The primary difference between the two methods lies in the method of feature extraction, specifically, whether the extraction process is manual or automated [13]. This difference is primarily emphasized due to the fundamental constraints of human expertise in the standard feature extraction procedure [14]. The features extracted from the data collected from the inbuilt smartphone sensors can be categorized into two main domain: (1) time domain and (2) frequency domain

[15]. The drawback of this traditional methodology is that, at times, human expertise may not consistently identify the optimal feature set for varying scenarios. Another disadvantage is that this method can produce redundant features, necessitating the application of dimensionality reduction techniques, such as feature selection, to minimize the impact of irrelevant features on the performance of classification algorithms.

In order to address these limitations, DL models provide an advantage in the process of feature extraction due to their ability to generate their own features. The algorithms possess the capacity to generate complicated and complex attributes that comprehensively contain the data, resulting in extremely effective classification models for many applications. As a result, deep learning techniques are widely recognized as being at the forefront of advances in fields such as computer vision and natural language processing [16].

The difference between the traditional machine learning approach and the deep learning approach is illustrated in **Figure 2.2** and **Table 2.2**. The segmentation phase is an integral part of the data preparation process, wherein the data is partitioned into segments referred to as time windows. In the traditional approach, these time windows are used for feature extraction. Conversely, the deep learning approach eliminates the need for time windows as it processes the raw data directly. However, some studies [17],[18],[19] introduce an initial preprocessing step for raw data to mitigate noise stemming from environmental conditions, movements, and shifts in user behavior during data collection. Common techniques employed for noise reduction include Lowpass filters, moving average filters, and Kalman filters.

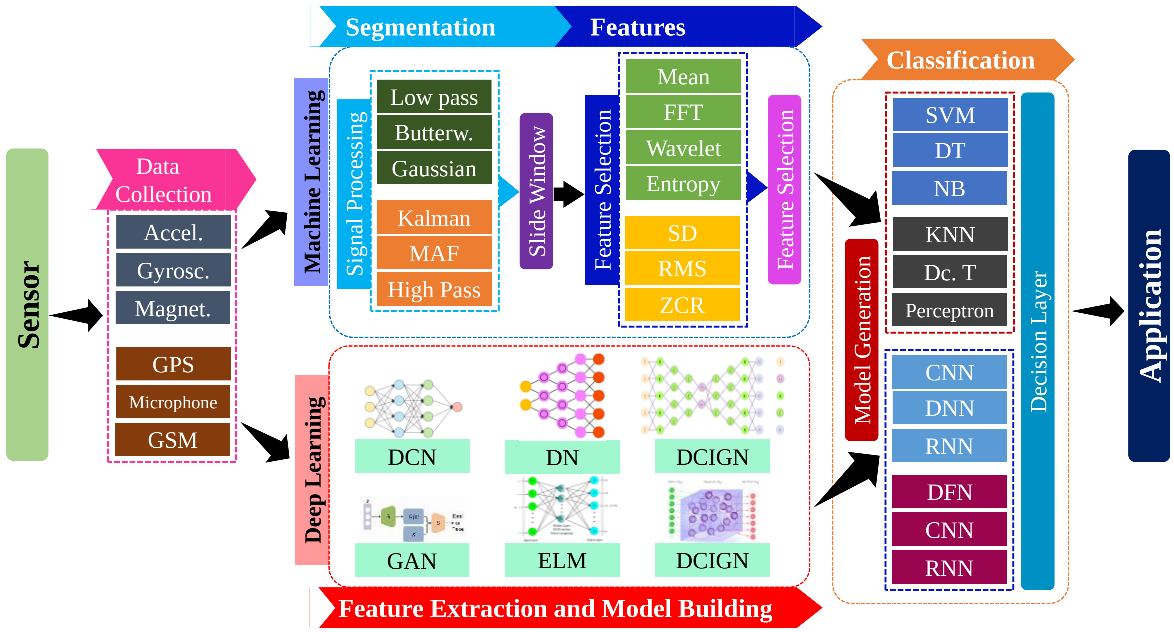

Figure 2.2 A brief overview of Machine learning and Deep learning approach.

The initial step in data collection studies typically involves gathering raw data from smartphone sensors, such as accelerometers and gyroscopes. During this process, several parameters such as data collection type, timing, frequency, smartphone position, and orientation on the user's body must be carefully considered. Smartphones commonly used for data collection are equipped with operating systems like Android, iOS, and Symbian.

The final step revolves around constructing classification models to infer human activities. These classification models are developed based on either shallow or deep machine learning algorithms.

Table 2.2 Basic difference between ML and DL-based HAR characteristics.

Characteristic	ML-based HAR	DL-based HAR
Data Requirement and Feature Engineering	• In traditional ML-based HAR, domain knowledge is crucial for selecting relevant features from sensor data. • Data preprocessing typically involves segmenting time-series data into fixed-sized windows and extracting features from these windows. • Feature engineering can include statistical measures, time-domain features, and frequency domain features.	• Deep learning models can automatically learn relevant features from raw sensor data, eliminating the need for manual feature engineering. • Features are learned through layers of neural networks, such as convolutional layers for spatial patterns and recurrent layers for temporal dependencies.
Model Complexity	Typically involves simpler models, such as SVM, Random Forest, or k-NN.	Utilizes complex models, including CNNs, RNNs, LSTMs, or hybrid architectures.
Performance and Generalization	Performance highly dependent on handcrafted features and domain knowledge. May struggle with generalization.	Often outperforms ML-based HAR due to its ability to capture intricate patterns and generalize across activities.
Real-time Processing	Well-suited for real-time processing due to simpler models and lower computational requirements.	May require more computational resources, making real-time processing challenging on resource-constrained devices.
Transfer Learning	Less common due to the specificity of handcrafted features.	More feasible, with pre-trained models being fine-tuned for specific HAR tasks.
Interpretability	More interpretable, as the relationships between features and activities can be well understood.	Often considered as "black boxes," making it challenging to interpret the learned representations.
Deployment on Edge Devices	More feasible due to lower model complexity and resource requirements.	Challenging due to higher computational demands, but lightweight architectures and optimizations can help.
Handling Multimodal Data	Capable of handling multimodal data by incorporating various handcrafted features from different sensor types.	Well-suited for multimodal data fusion, as it can process data from multiple sensors simultaneously.
Scalability	May struggle with scalability, especially with numerous activities or large datasets.	Can scale efficiently to accommodate more activities and larger datasets.

2.4 Existing Method

A lot of sensor-based [20]–[26] and vision-based [27]–[30] research is done field of HAR. Accelerometers and gyroscopes have been used in a large amount of research to classify actions. These research includes a wide range of activities. For example, fall detection techniques that use multiple sensor integration or setups based on a single accelerometer sensor[31]–[33]. While some systems focus on mobile activity recognition [34], [35], others are more concerned with posture recognition [36]. Many strategies, such using statistical models or keeping an eye on times of inactivity, have been put forth to reduce false alerts [37]–[39]. Using only accelerometer data, several algorithms have been developed for mobile applications for HAR. These algorithms have been implemented in a variety of instances, such as identifying between cycling, walking, running, and driving. Accelerometers have also been used to help in training statistics collection and analysis by classifying physical activities such as workouts and sports. Depending on the body motion, HAR system can be categorized into two major types- global body motion type and local interaction motion. However, this thesis mainly focuses on global motion type which is further discussed in the later section.

2.4.1. Global Body Motion Activity Classification

Activity classification based on global body motion involves categorizing various activities according to the overall movements of the body. This approach utilizes sensors such as accelerometers and gyroscopes to detect and classify activities based on the patterns of motion exhibited by the entire body. It encompasses a wide range of activities, including but not limited to walking, running, cycling, and driving. Studies focusing on global body motion activity classification often employ algorithms and techniques tailored to analyze and interpret the motion data collected from these sensors.

2.4.1.1. Waist-Mounted

Many research studies prefer to set accelerometer-based sensors at the abdominal region because it to be a more stable location as compared to the extremities for monitoring motion throughout the body. This is due to the fact that the extremities don't always move in synchronization with the body. Waist installation also makes it possible to conveniently position the sensor on a belt [40]. Some other earlier works did not even use a classification algorithm, but rather employed empirically determined thresholds for classification [34], [41] and [42]. Subsequent studies further developed this approach by employing a hierarchical approach to refine the initial classification of broader activity categories. For instance, the threshold would initially distinguish between motion-based activities like walking or running and motionless activities such as standing or sitting. Subsequently, another method would often be employed to further classify these activities into sub-categories such as walking, running, sitting, or standing. In [43], a rule-based approach is implemented to distinguish between activities that are in motion and those that are not. Following that, a SVM is trained on 3-axis acceleration data is used to make the final decision of classification. Weng et al. [44] utilize an SVM to differentiate between static and motion activities, as well as for the final classification within subcategories. Ataya et al. [45] employ a hierarchically structured approach, initially discerning static from dynamic activities. They evaluate multiple classifiers for the final classification decision, concluding that Random Forests outperform SVMs and other techniques. Notably, they also investigate the temporal coherence of activities over time, employing a Hidden Markov Model (HMM) to determine the plausibility of transitions such as lying down after running. Another significant trend in this field, apart from hierarchical methods, involved the comparison of various classifiers in the development of systems. As computational capacity increased, more recent efforts have focused on enhancing feature sets rather than classifiers [46]–[48]. Some of these works also included feature selection, meaning that the enhancements in accuracy were attributed not to a specific classifier, but rather to the input provided to the classifier. Additionally, some researchers focused on innovations beyond improving activity classification accuracy through feature sets. For instance, Alshurafa et al. [49] concentrated on distinguishing intensity levels of different activities, while [50] and [51] aimed to maintain classification accuracy while reducing energy consumption.

2.4.1.2. Mounted in other locations:

Since 2006, the utilization of accelerometer-based sensors in single systems expanded in terms of both placement and number. This trend led to an increase in the average number of classified activities, with many studies focusing on sets of approximately five activities initially. However, with variations in sensor placement and increased sensor numbers, subsequent works were able to classify a wider range of activities, with some covering up to 20 activities, averaging around seven or eight. While some studies continued to focus on central body locations such as the chest or lower back, others explored extremity placements, particularly with the emergence of arm-band mounted smartphones and wrist-based hardware. These latter works tended to employ more complex learning algorithms compared to threshold-based techniques used for central body locations, often utilizing supervised learning techniques. For instance, thigh-mounted deployments of single 3D accelerometers were employed by both [52] and [53], utilizing multiple hierarchically structured algorithms for classification decisions.

Similarly, wrist placements were also explored by various studies [54]–[56]. In [54] comparing different feature combinations and learning algorithms. Other works, [57], [58], and [59], discriminated between a larger set of activities using information from two sensor locations. Multi-sensor approaches extended beyond pairs of sensors, with some studies using data from sensors at three [60]–[62] or more locations [63]–[69]. Studies not focused on multi-sensor combinations or placements could be distinguished by their hardware implementations. While earlier works often developed specialized hardware, later studies attempted to utilize commercially available platforms such as mobile platforms or smartphones [70]–[76] . These efforts aimed to monitor mobility patterns over time or accommodate different placements of mobile platforms to ensure ease of use for individuals.

2.5 ML Algorithms for HAR

ML algorithms designed for time-series classification challenges in the domain of HAR can be divided into three primary categories [77]: discriminative, generative, and hybrid. Extensive research efforts have been dedicated to HAR using smartphones and smartwatches over the years. The principal ones are elaborated upon in the following sections.

2.5.1. Discriminative methods

The conventional methods for classifying HAR usually depend on algorithms like Random Forests, Decision Trees, Support Vector Machines (SVM), XG Boost, Logistic Regression, Artificial Neural Networks (ANN), and related methods.

In a study [78] on classification of different activities from accelerometer sensor data of smartphones and smartwatches three different algorithms Random Forest, k Neighbors, and Decision Tree were explored for the classification of 18 different activities of the WISDM dataset. They achieved an average accuracy of around 91% for each of the different activities. Another study used SVM [79] for classification of six ADL activities like Walking, Sitting, Laying, Standing, Upstairs, Downstairs from Smart Phone sensor data. They used 248 handcrafted features and obtained an accuracy of 89.5%.

The authors in [80] explored the frequency and wavelet domain feature and used the XG boost classifier for recognizing five different activities with an accuracy of 84.19% which outperformed the others. The a study on different accelerometer data from watch, phone and smartwatch gyroscopes [81], the authors explored five different classifying techniques, namely Naïve Bayes, J48 decision trees, RF, MLP and B3 instance-based learning obtaining an overall accuracy of 64%. A hybrid approach [82] for optimal feature selection was adopted and SVM was explored to with optimized and unoptimized features. The optimized features achieved an accuracy of 96% which is greater than 6% in the case unoptimized features. These feature driven methods are highly dependent on manual feature extraction. This is constrained by the need for domain-specific knowledge, time-intensive procedures, and resource-heavy requirements. To address these issues, DL methods are extensively used. Recent developments in HAR using sensor data have demonstrated the effectiveness of DL techniques, specifically CNNs and RNNs. These methods have achieved remarkable performance on complex HAR tasks, without the need for extensive manual feature engineering or laborious manual feature extraction from raw data. One of the key advantages associated with the utilization of DL techniques lies in the significant impact of automatic feature learning. This involves the extraction of temporal information through the implementation of DNN [83]. RNN, LSTM & GRU [84] are a few of the DL techniques that have been used for extraction of temporal features. CNN is one of the widely used DL technique for temporal features extraction. CNN eliminates the need for sliding windows to segment time series data. It directly applies convolution operations with small kernels along the temporal dimension of sensor signals, enabling the capture of local temporal dependencies [85]. Researchers have effectively employed both 1D and 2D CNNs for the recognition of human activities in HAR. For instance, the authors [86] implemented LSTM based approach to achieve a remarkable 92.1% accuracy in recognizing human activities from tri-axial accelerometer data in the WISDM dataset. In a similar vein, [87] employed attention-based CNN for HAR from weakly labelled data. One of the benefits of an attention-based framework is its capacity to effectively focus on labelled actions within long sequences, while efficiently ignoring background signals.

2.5.2. Generative methods

In the field of activity recognition through sensor data Naive Bayes, Hidden Markov Models (HMM) are the traditional generative models. HMM [88] was used to classify 10 different human actions by capturing the smoothness and temporal regularities in the human actions. It achieved an accuracy of 95%. In retrospective study, [89] a novel generative model emerged in the shape of a two-layer stacked auto-encoder facilitating the extraction of features from sensor data. This approach was applied to classify six different human body movements performed by five different individuals using the Leave-one-out method, resulting in a F-score value of 0.77. Furthermore, in a different context, Deep Belief Networks (DBN) [90] was implemented with other methods like CNN, LSTM etc for the classification of different activities. The explored ten different available public datasets using this method. DBNs are composed of a vertical arrangement of Restricted Boltzmann Machines (RBMs). The inherent challenge of incorporating discriminative elements from time-series data into these architectural models, along with their high processing costs, renders them outdated in contemporary contexts. However, it is worth noting that despite their obsolescence, these models can still yield superior outcomes compared to conventional methods.

2.5.3. Hybrid methods

Hybrid approach in the field of HAR endeavour to produce efficient attributes by integrating many discriminative, generative models together. In a study, the authors[91] have proposed a novel approach that combines the Stacked Denoising Autoencoder (SDAE) with LightGBM to effectively categorize diverse human activities. This classification task was performed on four different datasets, each comprising a combination of multiple sensor modalities. The combination of LGB and SDAE achieved a much greater accuracy of 98.23% on the HSBD dataset compared to the use of individual models such as LGB or SDAE for action recognition. CNNs are the predominant models that have been successfully combined in hybrid approach with various generative and discriminative models. The authors [92] employed a hybrid approach that integrated HMM and DNN to perform action recognition on the Acceleration dataset. Additionally, the combination of GMM and RF was also used with DNN. The Comparative analysis revealed that the HMM-DNN model exhibited superior accuracy (93.52%) and precision (93.37%) compared to the HMM-RF and HMM-GMM models. Another approach [93] used the RBM and CNN in an hybrid method for classification.

A combination of LSTM-CNN was used for the purpose of HAR [94]. The LSTM has the tendency to extract temporal features from inertial sensor data. CNNs were employed to extract the relevant information. For the purpose of model tuning, Batch Normalization and other hyperparameters were used which achieved and F1- score of 0.9578 on the UCI-HAR dataset, surpassing the performance of other baseline models. Another CNN LSTM [95] based approach was for the purpose of recognizing egocentric activities. These activities are classified into many categories, including daily activities, exercise, ambulation, and officerelated tasks. This study employed a multimodal sensor fusion approach, integrating egocentric videos and accelerometer data through the utilization of a combined CNN and LSTM model. Nevertheless, the outcomes achieved were suboptimal in comparison to the reference model owing to the limited availability of training data.

2.6 Concept of eXplainability

Explainability in HAR is a critical aspect that involves providing clear, interpretable explanations for the decisions and predictions made by HAR systems. HAR itself is a field of research and technology that aims to automatically identify and classify human activities using various sensors, such as accelerometers and gyroscopes. These sensors capture data from human motion and activities; next, machine learning and artificial intelligence algorithms are commonly used to process this data and activity classification is performed. The widespread adoption of DL techniques has brought to the forefront the challenge of interpretability. DL models are often considered black-box systems due to their complex architectures. In addition to models that are inherently interpretable, such as linear models or decision trees, there exist black-box models, such as DNN, which do not yield explanations in a straightforward manner. Therefore, contemporary approaches often employ post-hoc explainability techniques to analyze input samples and derive explanations for the model's reasoning. There are three distinct categories in which post-hoc approaches are classified. First, gradient-based methods [96], were developed. As a result, significant efforts have been devoted to experiment and conduct scientific research on methods and tools for making these models more interpretable. The initial focus to study an image deeply led to the rise in the concept of Saliency [97]. Furthermore, a gradient-based attribution method termed as

Gradients was initially proposed in [96], in accordance with this principle, each gradient measures the extent to which a modification in each input dimension would influence the predictions within a limited vicinity around the input. The advancement in the saliency and gradients led to further improvements in the explainability [98]. A new Krizhevsky network [99] was proposed to outperform the SOTA saliency architectures by 67%. Moreover, in [100], a customized pre-training approach was introduced to tailor multi-context modeling specifically for the purpose of saliency detection. With further advancement in the gradient techniques, Integrated Gradients [101] were introduced. It is a gradient-based attribution technique used to explain the predictions made by deep neural networks. It attributes these predictions to the input features of the network. This method is a variation of the Gradients approach, calculating the gradient of the prediction output concerning input features. It outlines two crucial axioms for attribution methods: sensitivity and implementation invariance. In response, [101] introduce the Integrated Gradients method as a straightforward approach that offers exceptional interpretability results. In a related work, presented in [102], attributions are employed to identify the shortcomings of three question-answer models more effectively than traditional methods. Simultaneously, these attributions enhance workflow efficiency [103]. Backpropagation variants are, for example, SmoothGrad [104] creates neighbor input samples by adding Gaussian noise to the features and averages the gradients computed on each of them.

Furthermore, perturbation-based techniques involve the introduction of disturbances to input data by incorporating a baseline value. Subsequently, these disturbances are assessed based on the performance degradation of the evaluation metric. The initial method employing this approach was Occlusion [105], which perturbed inputs through the use of square patches.

One prominent contemporary method is RISE [106], which generates masks via an up-sampling process using randomly filled binary masks and subsequently constructs a heatmap by evaluating the relevance of each mask in relation to the prediction. These perturbation-based methods tend to be computationally more intensive when compared to most of the post-hoc techniques, as they necessitate the replacement of input values and model computation for each iteration. Additionally, Guided Backpropagation [107], also known as guided saliency, represents a variant of the deconvolution technique [108] used to visualize features learned by CNN. This method is versatile and it can be applied across a wide spectrum of network architectures. Notably, it questions the utility of max-pooling layers in small image-based CNNs and suggests replacing them with convolutional layers featuring increased stride, all while preserving accuracy in multiple image recognition benchmarks.

Lastly, the CAM-based techniques are employed with CNNs to achieve the activation patterns from the convolutional layers, often the final one, to generate saliency maps. Among these methods GradCAM [109] is the very popular. It assigns weights to the activation maps based on the gradients obtained through backpropagation from the output neuron of a selected class to the corresponding convolutional layer. Various adaptations and enhancements of this approach are available. For instance, Layer-CAM [110] combines activations from different layers, Poly-CAM [111] introduces more complexity, and CAMERAS [112] aggregates outcomes at different scales for the input image. Another method, Score-CAM [113], employs the activation maps as masks to predict their significance.

2.7 Attention Mechanism

The attention mechanism in HAR is a computational component that allows a model to focus on specific parts of the input data, typically sensor data, when making predictions about human activities. It plays a very crucial role in enhancing the model's ability to understand and classify complex activities. In recent years, attention mechanism has been explored for deep HAR models. Attention can be categorized into different types depending on this use:

1. **Temporal Attention:** HAR often involves time-series data, and temporal attention mechanisms can help the model to pay more attention to specific time steps or segments of the data. This is particularly useful when certain parts of the sensor data are more informative for activity recognition. By assigning different attention weights to different time steps, the model can focus on the most relevant information. It can be implemented using RNNs, specifically LSTM cells and GRUs with an attention mechanism. These cells learn to give more attention to time steps that are more informative for recognizing specific activities.
2. **Spatial Attention:** In some cases, the spatial distribution of sensors on a wearable device or smartphone can vary. Spatial attention mechanisms can be used to give more importance to sensors that are more relevant for the current activity. For example, if you're recognizing activities related to hand movements, sensors on the wrist might be more important than those on the ankle.
3. **Multi-Modal Attention:** HAR can involve multiple sources of sensor data, such as accelerometer, gyroscope, and magnetometer data. Multi-modal attention mechanisms can help the model weigh the importance of each sensor type differently for different activities. This allows the model to learn which sensors are more informative for specific activities. For example, when identifying jogging, the model might focus more on the accelerometer data, whereas for activities involving precise

movements, it might give more attention to gyroscope data.

4. **Self-Attention:** Self-attention mechanisms, often used in natural language processing, can also be adapted for HAR. They allow the model to learn relationships between different elements of the input data. In the context of HAR, self-attention can capture dependencies between different sensor data streams or time steps, which can improve activity recognition accuracy.

In this research work, we have used self-attention-based approach for the classification of different activities.

2.8 Dataset and Framework

2.8.1. Dataset description

In this thesis, we use three datasets, including Sanitation dataset [114], UCI HAR dataset [115], and UCI HAPT dataset [116]. A brief overview of all three dataset is illustrated in **Table 2.3**

Table 2.3 The brief introduction of three datasets for HAR.

Datasets	**Sanitation** [114]	**UCI-HAR** [115]	**UCI-HAPT** [116]
Year	2019	2012	2012
Classes	7	6	12
No. of participants	50	30	30
Sampling rate	25	50	50
Sensor used	Accelerometer and gyroscope	Accelerometer and gyroscope	Accelerometer and gyroscope
Number of Samples	266555	10299	10299

2.8.1.1. Sanitation Dataset

This dataset [114] contains information on various activities performed by sanitation personnel, such as sweeping, sweeping with a big broom (Bsweep), walking, running, cleaning, dumping, and everyday routines (like sitting and smoking) collected in the open environment. A triaxial accelerometer worn in a wrist smart watch is used to collect seven types of daily work activity data of sanitation workers. The data includes X, Y, and Z acceleration values as well as labels for the corresponding sampling points. The size of the whole dataset is 266555 × 3, which contains 266555 samples. The dataset also includes a preprocessed version that has been split into windows using sliding window segmentation and features have been extracted from the time-domain as well as frequency-domain characteristics. The proportion of various types of activity samples is shown in **Figure 2.3**

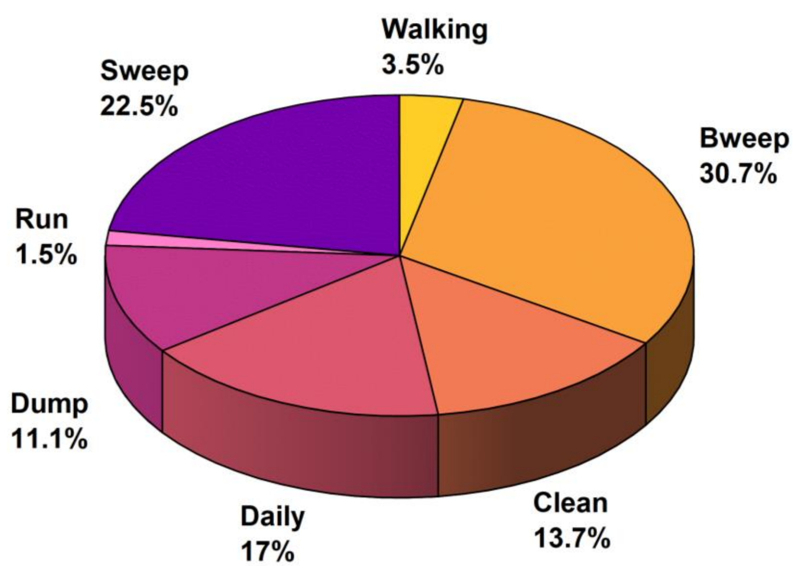

Figure 2.3 Percentage of activity in Sanitation dataset.

2.8.1.2. University of California HAR Data Set (UCI-HAR)

UCI HAR dataset is one of the most famous open datasets in the field of HAR, provided by University of California Irvine. This dataset [115] includes data from 30 individuals volunteers aged between 19 and 48, using a Samsung Galaxy SII smartphone attached to their waist. The data set contains information on 3-D acceleration and gyroscope measurements and covers a total of 10,299 instances of 6 different activities – 3 dynamic (upstairs, downstairs, walking) and 3 static (standing, sitting, and laying). This data has been recorded at a sampling frequency of 50 Hz and segmented into 2.56-second windows using a 50% overlap sliding window technique. For the purposes of experimentation, this data has been divided into a training set consisting of 7,352 samples and a test set of 2,947 instances. The proportion of various types of activity samples is shown in **Figure 2.4**

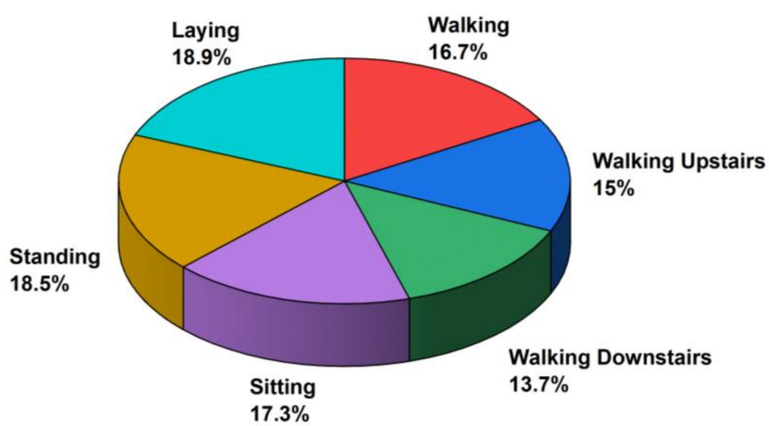

Figure 2.4 Percentage of activity in UCI-HAR dataset.

2.8.1.3. University of California Human Activities and Postural Transitions (UCI-HAPT) Data Set

The UCI-HAPT dataset is an extension of the UCI-HAR dataset. The distribution is illustrated in **Figure 2.5**. In addition to six basic activities provided in UCI-HAR dataset, this dataset [116] consists of additional 6 postural transitions: stand-to-sit, sit-to-stand, sit-to-lay, lay-to-sit, stand-to-lay, and lay-to-stand. We use processed dataset as the sampling datapoints termed as feature vectors. The on-body sensor position for all three datasets is shown in **Figure 2.6**

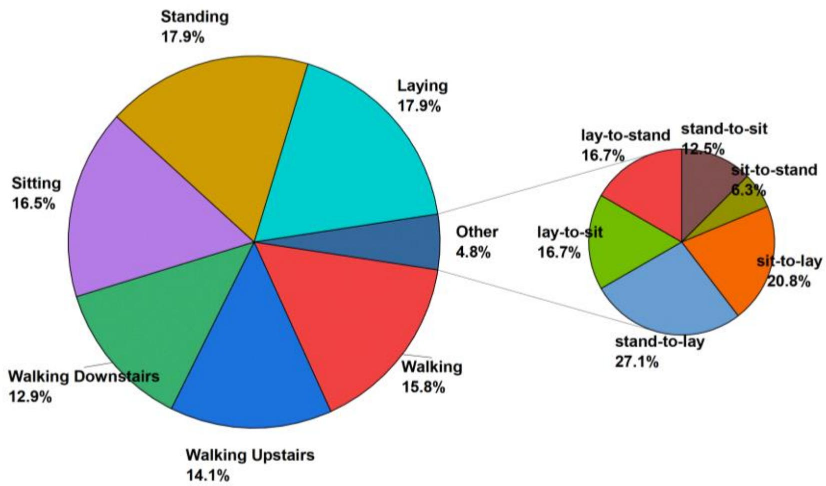

Figure 2.5 Percentage of activity in UCI-HAPT dataset.

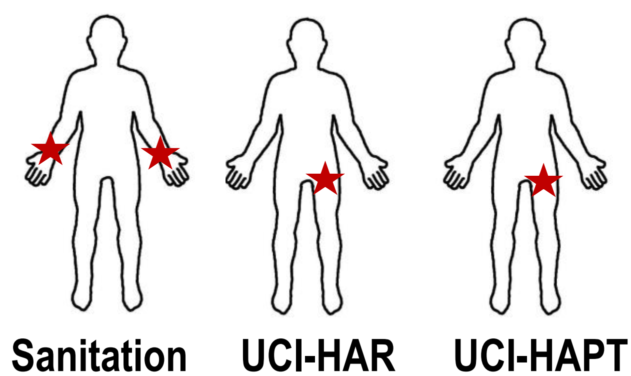

Figure 2.6 On-body sensor placement.

2.8.2. Frameworks

Several libraries and tools were used for implementing HAR system, making it easier to work with sensor data and develop HAR models. All libraries and tools used for HAR in Python are illustrated in **Figure 2.7**. All the python library version are as follows:

- ❑ Python → 3.9

- Keras → 2.6.0
- Pandas → 1.3.5
- Tensorflow → 2.9.0
- Matplotlib → 3.5.3
- Numpy → 1.21.6
- Scikit-learn → 1.1.1

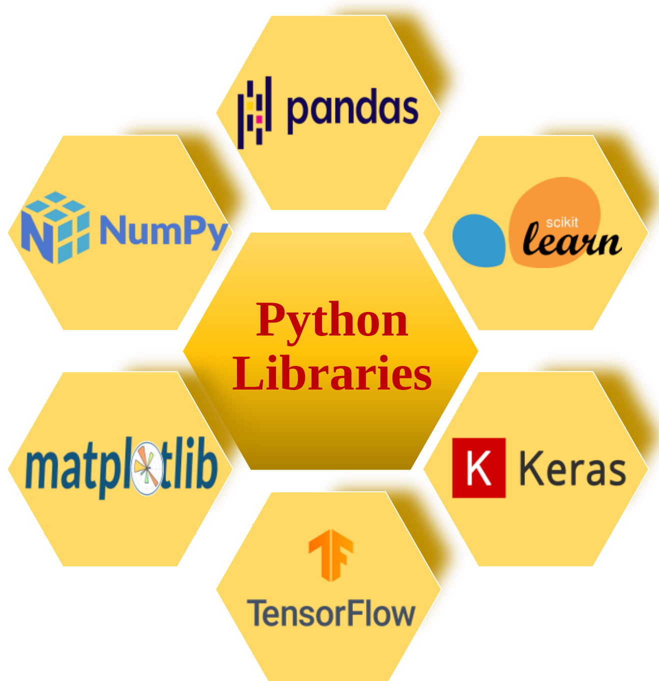

Figure 2.7 Python tools and libraries.

2.9 Evaluation Metrics

For the evaluation purpose, six evaluation metrics were considered to assess the overall performance. These metrics include true positive (α_{TP},) for correctly identified positive activities, true negative (α_{TN}) for correctly identified negative activities, false negative (α_{FN}) for misclassified positive activities, and false positive (α_{FP}) for misclassified negative activities. These parameters allow for a comprehensive evaluation of the model's ability to accurately classify activities based on whether they are positive or negative. By analyzing these metrics, the performance of the proposed models can be effectively evaluated.

[1] Precision: Precision is the ratio of true positives and anything that is predicted as a positive.

$$Precision = \frac{\alpha_{TP}}{\alpha_{TP} + \alpha_{FP}} \tag{8}$$

[2] Recall: Recall is the ratio of true positives and anything that should have been predicted as positive.

$$Recall = \frac{\alpha_{TP}}{\alpha_{TP} + \alpha_{FN}} \tag{9}$$

[3] Accuracy: The ratio of accurately predicted samples to all predicted samples is known as accuracy. The effectiveness of a model on test data can be best demonstrated by accuracy.

$$Accuracy = \frac{\alpha_{TP} + \alpha_{TN}}{\alpha_{TP} + \alpha_{TN} + \alpha_{FP} + \alpha_{FN}} \tag{10}$$

[4] Weighted F1-score: The harmonic mean of recall and precision is F1. Accuracy, on the other hand, is unable to precisely measure a multi-classification model's performance due to the issue of unbalanced data in the dataset. Consequently, in addition to accuracy, we also utilize the weighted F1-score as a statistic to assess the model's performance.

$$F1_{\,weighted} = 2 * \sum_{0}^{i} w_i \; \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{11}$$

[5] Kappa score: The Kappa score is utilized to assess classification effectiveness by measuring the agreement between classifier-assigned classes and ground truth data. Its value ranges from -1 to 1, with +1 representing perfect prediction, 0 indicating average random prediction, and -1 denoting inverse prediction.

$$Cohen\ Kappa = \frac{2.[(\alpha_{TP}\times\alpha_{TN})-(\alpha_{FN}\times\alpha_{FP})]}{(\alpha_{TP}+\alpha_{FP}).(\alpha_{TP}+\alpha_{FN}).(\alpha_{TN}+\alpha_{FP}).(\alpha_{TN}+\alpha_{FN})} \tag{12}$$

[6] Matthew's correlation coefficient (MCC): Accuracy and F1 score, commonly used metrics, can yield misleading results on imbalanced datasets as they overlook the positive-to-negative variable ratio [117]. MCC (Matthews Correlation Coefficient) is a valuable indicator that considers this ratio, with a range of -1 to 1. Equation (13) provides the formula to calculate MCC, where +1 signifies perfect prediction, 0 indicates average random prediction, and -1 represents inverse prediction.

$$MCC = \frac{(\alpha_{TP}\times\alpha_{FP})-(\alpha_{TP}\times\alpha_{FP})}{\sqrt{(\alpha_{TP}+\alpha_{FP}).(\alpha_{TP}+\alpha_{FN}).(\alpha_{TN}+\alpha_{FP}).(\alpha_{TN}+\alpha_{FN})}} \tag{13}$$

2.10 Conclusion

In this chapter, a brief history of HAR has been presented. It outlines the various types of HAR and their associated sensor modalities. Following this, a general process for HAR is introduced, followed by a discussion of existing literature in the field. The chapter also discussed the concept of explainability, addressing the inherent black box nature of deep learning models. Additionally, various types of attention mechanisms are introduced as part of this exploration. Furthermore, the datasets, framework and different evaluation metrics used in this research are also presented.