CHAPTER II

BACKGROUND STUDY

One of the most well-liked subfields in computer vision research is the detection of human activity in video footage. It uses a variety of Deep Learning techniques, including convolutional neural networks, recurrent neural networks, skeleton-based recognition, and graph convolutional networks for detecting the human actions. This chapter provides a literature review on the of HAR. Limitations of HAR and the approaches used are also discussed.

2.1 REVIEWS OF RELATED WORKS

2.1.1. Skeleton Based Human Activity Recognition Methods

In a novel approach, Jiang et al. (2015) offered skeletal context to assess the level of similarities across postures, and then utilize this information to enhance action recognition. The similarity is obtained by taking the multiple scales paired positional distributions at each relevant joint and dividing it by the number of informative joints. Following that, the feature sets are assessed using a bag-of-words technique by applying the linear Conditional Random Field (LCRF).

Nikolov et al. (2018) created a novel human activity recognition system that utilizes Convolutional Neural Networks (CNN) for classification and establishes correspondences between chosen feature points across frames next to one another to represent human actions. An iterative principal component analysis (PCA) technique was used in order to simplify the data complexity. At the conclusion of each iteration, the pipeline is supplied with the finalized matrix either from the preceding iteration or from another user doing the same action. Nevertheless, the task of recognizing human actions in real time poses significant challenges and is characterized by its complexity and noise interference.

By utilizing skeleton joints coordinates that exploit spatial form structures of the skeletons and RGB Dynamic Images (DIs), (Dhiman et al. 2019) offer a novel view-invariant human action recognition system. Features based on skeletal coordinates aid CNN models in locating humans and their shifting positions and orientations throughout an action, all while cutting down on background noise. In contrast, DIs use

the Average Rank pooling concept to hide the temporal dynamics of action. The InceptionV3 architecture has undergone refinement specifically for two multi-view human action datasets. Through the use of transfer learning, it has the ability to project a hybrid representation of human qualities into a higher dimensional environment.

Hristov et al. (2020) proposed a method for classifying human behavior using data collected from 3D models of their skeletons. A parallel convolutional and dense layer neural network is then fed this altered input. By re-propagating the training data, the network's output is inferred at the next-to-last layer after training. This output is fed into a SVM learning environment. Bayesian Optimization was used to find all hyperparameters for the PKU-MMD dataset. However, because to its intricacy, this may increase the entire training time.

Using information from 3D skeletal scans, (Agahian et al. 2020) created a novel model for HAR. The model's fundamental components were the representation and encoding of poses. Pose descriptors have three components, all of which pertain to human behaviors in terms of space and time. The real joint data is standardized in the first element's coordinates. The second component stores information about shifts in the fixed sequential equalizer, while the third was utilized to store information about shifts in the sequential firmness over time. Human behavior was also classified using a classifier developed by Extreme Learning Machine (ELM). However, it must improve the sequential links between the activity and discriminatory postures in order to distinguish between similar positions.

Using Graph Convolutional Networks (GCNs), (Liu et al. 2020) demonstrated the first instance of conflict in skeleton-based HAR. An activity structure's joint positions were jumbled up to create a fighting formation, while the skeletons' sequential consistency, 3D reliability, and anthropomorphic probability were all preserved. This was accomplished by maintaining a great deal of physical restraint, employing 3D skeleton relocations for the unnerved skeletons, and standardizing the aggressive skeletons' reproductive systems. However, there were setbacks wherein the preferred strategy led to ambiguity regarding several essential tasks.

In order to determine the exact pixel position of important body joints, (Pham et al. 2020) created a real-time 2D posture recognizer. Joints in two dimensions were converted into three dimensional poses with the help of a DNN. Photos were used to characterize the spatial and temporal evolution of the projected 3D postures via an intermediary interpretation and HAR after an effective neural structure exploration technique was employed to find the best suitable DNN structure. However, it is unable to reconstruct 3D postures from 2D failures.

To extract the gesture information (Wang et al. 2021) developed a unique HAR approach they termed Skeleton Edge Motion Networks (SEMN). Skeletal border position shifts and associated joint trajectories also accounted for the SEM. The SEMN was then developed by fusing several spatiotemporal slices to better understand skeletal frameworks. To further safeguard the sequential imperative data, a novel advanced rank error was implemented. However, a grainy skeletal image made it difficult to distinguish between the various actions.

(Weiyao et al. 2021) created a Bilinear Pooling and Attention Network (BPAN)-based multi-modal HAR framework. The RGB and skeletal data were preprocessed first. After that, an RGB film and skeleton structure fusion network was developed. The BPAN was used to filter out the RGB and skeletal features and project them into latent subspace so that the fused features could be acquired. In addition, a fully connected 3-unit perceptron was used to make the ultimate classification call. However, the weight value in the loss function affected the overall accuracy, and the training database was small.

With the ability to train HAR using a hierarchical system, skeletal motion contour and postural structure (Li & Sun 2021) developed a new CNN fusion framework based on a 2-stream structure with 3 Skeletal Pose Image (SPI) structures and 3 Skeletal Trajectory Shape Images (STSIs). In addition, a translation technique for 3D skeletons was created, which uses gray-coded subdivision into 3D SPI and STSI structures on 3 orthogonal surfaces to define spatiotemporal data. Additional methods for discovering spatial and temporal characteristics included spatial and temporal pooling. However, for improved performance in human-object activity recognition, it must incorporate characteristics surrounding 3D skeletal information.

Using four classifier models (RF, SVM, MLP, and KNN), (Ramirez et al. 2021) proposed a camera-vision based fall detection and activity identification system. It is proposed to characterize RGB images in terms of sets of human skeletons using posture estimation as a feature extraction procedure, with the most prominent skeleton in each image being taken into account. The method's efficacy is measured with the UP-FALL public dataset's single-camera RGB modality evaluation.

(Ryu et al. 2022) created a HAR model that can extrapolate coordinates and angles from the provided data. To further classify the data, the collected features were combined and input into an ensemble of multiple ELM classifiers. However, in order to maximize HAR's efficacy in real-world applications, adequate training data is required to handle each possible action variation.

Three-dimensional connections between the bodies in RGB videos were created by (Cha et al. 2022) for HAR. Rebuilt 3D interconnections were analyzed using the transformer architecture to yield helpful skeleton interpretation. However, training the bones and skeleton joints together was challenging because to poor memory efficiency.

For skeletal-based HAR, (Yadav et al. 2022) developed a novel Convolutional Long Short-Term Memory (ConvLSTM) network. Human identification and pose estimation were utilized initially to precompute skeletal coordinates from the image/video sequences. The trained ConvLSTM ensemble then created the new reference attributes based on the geometric and kinematic aspects of the actual skeleton coordinates. In addition, a connected unit categorizer head was used for HAR. However, it did not take into account the geometric features of edges and surfaces that can improve HAR performance.

Table 2.1 Comparison of Skeleton- based methods for Human Activity Recognition

Author and Year	Methods	Merits	Demerits	Accuracy%
Jiang et al. (2015)	LCRFs, Dataset: UTKinect dataset, MSRAction3D dataset.	It avoids the label bias problem	The training phase of the method is computationally intensive.	This method achieved accuracy 97.08%
Nikolov et al. (2018)	CNN and PCA algorithm, Dataset: NTU RGB+D, Kinect dataset	PCA analysis improves when a variety of users perform the same task.	It is difficult and noisy to recognize human actions in real time.	This method not provide a predicted accuracy.
Dhiman et al. (2019)	CNN models, Dataset: NUCLA Multi-View Dataset, NTU RGBD Dataset	It projects higher dimensional space	It takes lots of time to predict the HAR	It achieved 89.6% accuracy.
Hristov et al. (2020)	Neural network with parallel convolutional and dense layers, Dataset: UTDMHAD dataset	Increasing the predictive ability of a model	Because to its intricacy, this may increase the entire training time.	The system was 92.4% accurate.
Agahian et al. (2020)	ELM, Dataset: UTKinect Action Dataset, Florence 3D Action Dataset, UTD-MHAD Dataset	It has a better generalization ability	It needs to differentiate similar poses by improving the sequential relationships of the activity and discriminatory postures	This method achieved 97.0% accuracy.
Liu et al. (2020)	GCNs, Dataset: NTU and Kinetics data set	It makes easily understand the human action and gives good accuracy	There were a few failures wherein the desired method was prone to confusion on a few skeleton activities	This method achieved 88.3%
Pham et al. (2020)	Deep Neural Network (DNN), Kinect Interaction datasets	The computational overhead for training and inference is low using this approach.	It can't rebuild 3D postures from 2D unsuccessful outcomes.	This method achieved accuracy of 97.98%

Wang et al. (2021)	Skeleton Edge Motion Networks (SEMN), Dataset: Penn Action, UTD-MHAD, NTU RGB+D, NTU RGB+D 120, and CSL datasest	Advanced rank error was applied to preserve sequential imperative data	It was hard to differentiate the individual activities from a granular skeleton image.	It achieved a promising accuracy of 93.05%
Weiyao et al. (2021)	Bilinear Pooling and Attention Network (BPAN), Dataset: NTU-RGB+D, NTU-RGB+D 120, UTD-MHAD dataset.	It needs to capture the traits surrounding 3D skeletal information to enhance the efficiency of recognizing human-object activities	The training database was limited, and the total accuracy was influenced by the weight value in the loss function	It achieved the accuracy of 95.07%
Li & Sun (2021)	Skeletal Pose Image (SPI), Skeletal Trajectory Shape Images (STSIs), Dataset: Florence3D-Action, Toyota Smarthome, and NTU RGB + D dataset.	It easily predicts the action of the human	It needs to capture the traits surrounding 3D skeletal information to enhance the efficiency of recognizing human-object activities.	It achieved 98.2% accuracy.
Ramirez et al. (2021)	RF, SVM, MLP and KNN, Dataset: UP-FALL public dataset.	Using only a video image and a person's posture, it can identify human activities.	The process of determining the desired skeleton from a given series of frames is lengthy.	The metrics accuracy has been increased by 65.25–71.85%.
Ryu et al. (2022)	Extreme Machine Learning (ELM), dataset: UTD-MHAD dataset, DBC dataset,	The cross-subject protocol showed that the recognition system using the angle data was more accurate.	In order to improve HAR's efficacy in real-world applications, it requires adequate training data to manage all possible action variations.	It achieved 91.67% of accuracy.

Cha et al. (2022)	3D-CNN, Dataset: SYSU and UTD-MHAD datasets	Reduction of design time and costs	Due to inefficient memory, it was difficult to teach the bone and skeleton joint locations in tandem.	Improves accuracy by roughly 6% on average compared to a similar SGN action classifier trained on sensor-captured skeletons.
Yadav et al. (2022)	ConvLSTM network, Dataset: KinectHAR dataset	They are much better at handling long-term dependencies.	It did not consider the edges and surface-related geometric traits to enhance HAR efficiency.	98.89% accuracy

2.1.2. Human activity recognition using deep leaning methods

To identify both normal and abnormal human behavior, (Arifoglu et al. 2017) presented a recurrent neural network. Vanilla RNNs (VRNN), Long Short Term Memories (LSTM), and Gated Recurrent Unit RNNs (GRU) are the three varieties of RNNs that being investigated. In this setup, activity identification is viewed as a sequence labeling issue, and out-of-the-ordinary actions are detected by comparing them to the norm. Researchers compare RNNs against other state-of-the-art methods including SVMs, Naive Bayes (NB), Hidden Markov Models (HMMs), Hidden Semi-Markov Models (HSMMs), and Conditional Random Fields (CRFs) to provide a thorough assessment of RNNs' performance in this setting.

Deep recurrent neural networks (DRNNs) are recommended by (Murad et al., 2017) for use in the development of recognition models that can capture long-range relationships in input sequences of variable durations. Unidirectional, bidirectional, and cascaded LSTM-based architectures are introduced, and their performance is compared and contrasted over a number of benchmark datasets. Furthermore, the suggested models outperform competing deep learning approaches, including DBNs and CNNs. These models use deep layers in a task-dependent and holistic way,

allowing them to extract more discriminative features. However, it is not practical for low-power gadgets.

An RNN model was presented for human activity recognition by (Singh et al., 2017). On the publicly available Benchmark datasets, the LSTM is used to categorize human activities including cooking, bathing, and sleeping. As can be seen from these experimental data, the proposed strategy significantly outperforms the state-of-the-art alternatives.

When considering a wide variety of parameters, the LSTM-RNN-based model seems to be a promising approach to simulating emotional states (Son et al., 2017). This includes a two-layer LSTM network model with two dimensions of emotional state: One output layer that calculates the outcome as the mean of the hidden values and one recurrent states layer made up of m-memory cells to compute the hidden values.

The accelerometers in mobile phones are used to gather data for HAR in a study by (Alsarhan et al. 2019), who employ the bidirectional gated recurrent units (GRU) technique. The bidirectional GRU model is used to effectively identify human behaviors from time series data. However, it struggles to perform difficult tasks.

Ding et al. (2019) suggested a deep recurrent neural network (HARNN) method for human activity recognition based on WiFi CSI. The use of the suggested Recurrent Neural Network (RNN) model is employed for the purpose of recognizing various human activities that are essential for the implementation of the Hierarchical Attention Recurrent Neural Network (HARNN). It may also substitute these generated representative characteristics for raw WiFi CSI data when doing activity recognition, which is a critically important use case. Because of this, the impact of indoor-sourced random noise is greatly diminished.

Wearable body multi-sensor data was used to develop a deep SRUs-GRUs based activity recognition system (Gumaei et al., 2019). This research done using multimodal body sensing data with a deep learning model composed of simple recurrent units (SRUs) and gated recurrent units (GRUs). They process sequences of data from multiple sensors using the fast processing power and enormous memory capacity of

deep SRUs networks. However, it fails when dealing with a huge or complicated data set.

This study proposes a novel approach to identify and extract data segments related to an expected sub-activity within a separate activity by integrating outlier detection with Human Activity Recognition (HAR) techniques (Munoz-Organero 2019). A DRNN-based method for identifying HAR outliers was proposed in this study. Two distinct datasets are used to verify the outcomes of both within-subject along with between-subject examinations, conduct outlier identification and sub-activity recognition. In addition, the architecture has undergone extensive optimization to cut down on computational costs and hardware prerequisites.

In the context of a human activity recognition challenge, the architecture of a convolutional neural network (CNN) is shown (Zebin et al., 2019). Hyper-parameters like the number of filters and filter size can greatly affect a CNN's performance, and in this case, the layers of the CNN learn features automatically to adapt for this. In terms of robustness and the ability to recognize actions with a temporal dependence, the suggested CNN beat statistical machine learning techniques. In order to be implemented on edge devices like smartphones and wearables, it has undergone extensive empirical inquiry on memory utilization and execution time for its activities. Classification accuracy drops with this pre-trained model during transitions between activities and when sensors are moved.

The InnoHAR model suggested by (Xu et al. 2019) combines convolution kernels of varying scales with max-pooling layers to identify human activities from data gathered by wearable sensors. High generalization performance is achieved by the suggested method, and it regularly outperforms the three most popular public datasets. Extensive testing on a MinnowBoard Turbot Dual Core Board proved experimentally that our unique structure excels in real-time applications. However, the network is deep, and there isn't enough training data to do it right.

(Vahora et al., 2019) offer a novel unified framework based on deep neural networks to recognize group behaviors in video surveillance. Understanding the temporal dynamics at the group level requires a multi-layer deep architecture, such as a convolutional neural network (CNN) trained on human action-pose data followed by a

recurrent neural network (RNN) model. To further disentangle environmental factors from otherwise puzzling group actions, scene information is recorded as a scene feature and scene level semantics are learned via a scene CNN. The LSTM model and the GRU model are used to tackle the long-range dependency problem of the simple RNN model, and their performances are compared across a variety of learning parameters.

This study introduces a Deep Learning model that is designed to be lightweight and suitable for Human Activity Recognition (HAR) tasks specifically on edge devices (Agarwal et al., 2020). The algorithms for Long Short-Term Memory (LSTM) and Shallow Recurrent Neural Networks (RNN) were used in the development of this model, which was created utilizing deep learning methods. During the model training and testing processes, six HAR activities are used to establish the best setup for a low-resource edge device, such as a Raspberry Pi3. The WISDM dataset serves as the experimental platform, and it contains sensor data from 29 people as they jog, walk, stand, sit, ascend, and descend stairs. However, because it was built with only a single tri-axial accelerometer, this system is unable to accommodate data from many sensors.

Channel State Information (CSI) and deep learning (DL) are offered as a device-free method for human identity authentication (Ding et al. 2020) present a passive identity identification method based on WiFi CSI that makes use of a recurrent neural network. The viability of the proposed WiFi technique in everyday interior situations is tested by real-world studies. First, to describe the performance in the two scenarios based on the real-world datasets were obtained. Next, to observe the differences between the suggested Wihi method and the currently used identification techniques.

Due to its capacity to distinguish characteristics via encoding and decoding and to evaluate a time series with a recurrent layer, the RNN architecture developed by (Paydarfar et al. 2020) was ideal for classifying the noisy piezo resistor data. This study demonstrates that modest amounts of data gathered by low-cost, uncalibrated sensors are comparable in accuracy to that obtained using IMU-based techniques.

Human activity recognition using mHealth data led to the development of a residual bidirectional LSTM deep learning solution to the problem of imbalanced classes (Malki et al., 2020). To enhance the abilities of the minority class, a weighted-class technique was used in conjunction with a residual bidirectional LSTM. The

findings from the experiment indicate that the accuracy of predicting human activity on datasets with a large number of dimensions may be improved by integrating residual bi-directional LSTM alongside the weighted-class approach.

Using motion tracking and the extraction of spatial attributes, (Jaouedi et al., 2020) recommended an integrated deep learning development for video motion recognition. The system uses Gated Recurrent Neural Networks (GRNN) to collect per-frame data useful for human action prediction, as well as the Gaussian Mixture Model (GMM) and Kalman Filter (KF) to identify and segment the moving subject. However, its effectiveness on complicated datasets was limited, and the time required for video categorization was extensive.

Two enhanced deep CNN models for sensor-based HAR were suggested by (Xu et al., 2020), which used the GAF approach for time-series processing. The a single-dimensional time series of the sensor is first converted into two-dimensional images by the use of the Gramian angular field (GAF) approach, and then the various human actions are classified using an enhanced deep convolutional neural network (CNN). To create a more accurate feature map, the multi-dilated kernel residual (Mdk-ResNet) captures features from sampling sites at varying time intervals. Also, the multi-sensor data fusion network Fusion-MdkResNet is proposed. Data from multiple sensors can be processed automatically by this network and merged into a unified whole.

End-to-End Deep Learning Framework (E2EDLF) is provided by (Alazrai et al., 2020) to detect HHIs from CSI signals. A full-stack deep learning architecture is provided by the authors, which includes three stages. The first stage encompasses the conversion of unprocessed Channel State Information (CSI) data into visual representations that could be utilized for temporal and spatial event monitoring. In the feature extraction phase, a novel convolutional neural network (CNN) is devised. Subsequently, the gathered characteristics are used during the recognition stage, whereby every CSI image is allocated to one of many HHI categories.

A DL model using LSTM and 1D-CNNs been proposed (Zhu et al., 2020). The experimental results show that the proposed model is effective in extracting spatio-temporal features from the radar data, which in turn leads to improved recognition

accuracy and simpler implementation in 2D-CNN approaches. The simplest parameter set of this network significantly increases its efficacy.

Convolutional neural networks (CNNs) and Bidirectional long short-term memories (BLSTMs) are used in the deep learning multi-channel architecture described by (Ihianle et al., 2020). This setup is preferable because the CNN layers may abstractly represent raw sensor inputs for feature extraction and perform direct mapping between resolutions. In order to greatly enhance the retrieved characteristics for activity recognition, the BLSTM layer makes use of both forward and backward sequences.

(Maragathavalli et al. 2021) proposed a modified Bidirectional Recurrent Neural Network (BRNN) with Long Short-Term Memory (LSTM), a Deep Learning (DL) technique, as an improved human activity identification model for fall detection. The approach under consideration is based upon data acquired from the sensors included inside the smartphone (such as the accelerometer and gyroscope). Next, the information is divided and preprocessed. Then, six different human actions are added to the system's training and evaluation.

A novel hybrid deep learning model, CNN-GRU, was presented to classify complicated human actions (Gupta et al., 2021). The raw sensor data from the WISDM dataset was used for this investigation. Data specific to wearables and smartphones was extracted from the full collection. Sliding window transformations were applied to the data during preprocessing. In this research, no human-driven feature engineering was used. The usage of AutoML, a port of the open-source McFly software, considerably facilitated the development of these intricate DNN models. This allowed for the creation of baseline models like DeepConvLSTM and InceptionTime.

(Nouriani et al. 2022) a strategy that use deep learning techniques to accurately identify and classify everyday tasks. A nonlinear observer that utilizes inertial sensors to assess tilt angles of body segments and sensor bias factors provides the inputs to the deep learning-based approach. The current gold standard for activity recognition is a deep learning-trained convolutional neural network with a long short-ter m memory (CNN-LSTM). Using raw inertial sensor data and nonlinear observer-determined angle inputs, they give experimental results that assess the CNN-LSTM network's performance.

For Human Activity Recognition with Wearable Sensors, (Khatun et al. 2022) presented the deep CNN-LSTM with self-attention model. The human activity identification framework integrates a deep learning model of self-attention with data from a variety of smartphone sensors. While additional sensors like a Global Positioning System (GPS) or a pressure sensor might improve recognition accuracy, the suggested technique just requires a three-axis accelerometer, gyroscope, and linear acceleration sensor.

An augmented multi-channel convolutional neural network (AMC-CNN) is proposed for HAR (Shi et al. After the feature window is built using time series sliding windows, the data is transformed and augmented. Second, a lightweight and efficient multi-channel convolutional neural network was developed. In order to explore deep connections among sensor data, researchers use suitable convolution kernel configurations to perform multiple channels of communication convolutions, which retrieve simultaneous multi-scale attributes, and training the model for the purpose of recognizing human behaviors. However, it is important to note that.

To achieve optimal performance in HAR classification, (Raziani et al. 2022) suggested a series of methods for autonomously tweaking the hyperparameters of 1-D CNN. The suggested method uses metaheuristic algorithms to randomly scan the search space of each hyperparameter in order to discover the ideal values for increasing CNN performance. The outcomes showed that the recommended procedure had better classification accuracy than the alternatives. Setting weights and biases at random from the outset increases the possibility of local maxima forming, which might slow down the convergence speed of DL or ML algorithms.

To develop thorough classification strategies for HAR utilizing data from wearable sensors, (Athota et al. 2022) developed a number of Hybrid Learning Algorithms (HLA). The objective of this research is to employ the The process of convolution Memories Fusing Algorithms (CMFA) as well as the The process of convolution Gated Fusing Algorithms (CGFA) for the analysis of sequential data. The aim is to develop a model that can effectively acquire knowledge pertaining to both local qualities and relationships that span both long-term and gated-term periods. The use of various filter sizes may be attributed to the advancements in feature extraction.

This put into place because of their usefulness in capturing various local temporal dependencies.

Using a convolutional neural network (CNN) and a bidirectional-gated recurrent unit (Bi-GRU), (Ahmad et al. 2023) created a system for recognizing human actions from video data. CNN is used to extract deep features from a framesequence of human-activity films, and then the most essential features are selected from the deep appearances in order to increase speed and decrease computing complexity. Second, using the critical information the researcher gathered from the frames' sequence of people's behaviors, we will develop Bi-GRU, a neural network that can learn forward and backward temporal dynamics at each time step. It has various restrictions, such as the fact that it can only recognize human behaviors when running on GPU and can't predict human activity on IOT-based devices.

Table 2.2 Comparison of human activity recognition using deep leaning methods

Author & Year	**Methods & datasets**	**Merits**	**Demerits**	**Accuracy(%)**
Arifoglu et al. (2017)	Vanilla RNNs (VRNN), LSTM-RNN and Gated Recurrent Unit RNNs (GRU-RNN); Dataset: dataset collected by Van Kasteren	It gives better performance while using LSTM-RNN	Due of the difficulty in collecting real-world data, they propose simulating some of the behaviors of people with dementia.	LSTM has the highest accuracy, at 96.7%.
Murad et al. (2017)	Deep recurrent neural networks (DRNNs), dataset: UCI-HAD, USC-HAD, Opportunity, Daphnet FOG, Skoda	When compared to CNNs and Deep Belief Networks (DBNs), the results are superior.	It doesn’t implement for low power devices	The UCI-HAD performed best when using a unidirectional DRNN model, achieving a success rate of 96.7%.

Singh et al. (2017)	RNN with LSTM Classifier, dataset: Opportunity	It gives high accuracy	Different hyperparameters in the LSTM model dramatically alter the model's performance.	When compared to the proposed work, it increases accuracy by 40%.
Son (2017)	LSTM-RNN, suggested framework, an integrated app called Collection Information was developed and published on the Google Play Store so that users could easily acquire the appropriate dataset.	Superior sensitivity to human brain processes	The Thayer model appears to have empirical support for its suggestion that the mechanism of mood is where psychology and biology meet.	This method not predict any accuracy.
Alsarhan et al. (2019)	Bidirectional gated recurrent units (GRU), dataset: UniMiB SHAR, WISDM, UCI-HAR Dataset.	It has less memory and faster than compared to LSTM	It is not effective in complex tasks	When the modified parameters were applied to the AF-17 problem, accuracy increased to 93.9%.
Ding et al. (2019)	RNN, data from two indoor settings (a conference room and a laundromat) that they themselves collected are used.	It reduces the influence of the random noise derived	It is difficult to process longer sequence	The system was able to attain an overall recognition accuracy of 96% across all tasks.
Gumaei et al. (2019)	Simple Recurrent Units (SRUs) and Gated Recurrent Units (GRUs), dataset: MHEALTH	The issues of a vanishing gradient and the existence of fluctuations are simply resolved.	It doesn’t work on complex or large data set.	This model achieved 99.80% of accuracy for recognizing the activities

Munoz-Organero (2019)	DRNN, Dataset: Realworld (HAR), opportunity dataset	It extend the powerful pixel neighbourhood.	The architecture has been highly optimized to minimize both computational expenses and hardware prerequisites.	It achieved better F1 score is 0.97
Zebin et al. (2019)	CNN Architecture Dataset: Opportunity, Pamap2, and UCI HAR dataset.	Strengthening through enhanced activity detection	Transitions between activities, as well as sensor displacement, reduce classification accuracy.	This method achieved accuracy rate 96.4%.
Xu et al. (2019)	Convolution kernels, Gated Recurrent Unit (GRU). Dataset: Opportunity, PAMAP2, Smartphone Dataset.	They used less memory and is faster than LSTM	The network is too deep, and there isn't enough data to properly train the structure.	Classification accuracy as a whole should not be used as a measure of performance. The highest F-measures (results) of 0.946 were found across all combinations of models and the Opportunity Dataset.
Vahora et al. (2019)	LSTM, RNN, GRU. Dataset: Collective Activity Dataset	It can recall temporal dependencies over lengthy periods of time without suffering from the issues of vanishing or ballooning gradients.	They are more difficult to implement and need more data for training than regular RNNs.	GRU RNN methods achieved 83.45%.
Agarwal et al. (2020)	Recurrent Neural Network (RNN) combined with Long Short Term Memory (LSTM),	A single tri-axial accelerometer was used in the development of this system.	It cannot support multi-sensor data.	The lightweight RNN-LSTM was able to get a 99% accuracy rating for running and walking.

	dataset: WISDM dataset.			Upstairs action is predicted with a minimum accuracy of 81%.
Ding et al. (2020)	RNN, dataset: The indoor office and laboratory settings are used to obtain data on human gaits.	The effect of random noise created by indoor surroundings can be greatly diminished.	Multiple target identification, footpaths, and a limited testing range are only some of the problems with the proposed WiFi system.	In this method accuracies which vary from 88% to 95%
Paydarfar et al. (2020)	RNN architecture Dataset: data acquired in raw form from shoes fitted with sensors	It is inexpensive	The accuracy is low	After 20 iterations, the validation accuracy was 84.8 2%, and the testing accuracy was 87.0 8.9%.
Malki et al. (2020)	Bidirectional LSTM Deep Learning method, dataset: mHelath dataset.	It's a useful method for predicting human behavior from complex datasets.	Required more time for training	Unprocessed data gathered using specialized, instrumented footwear.
Jaouedi et al. (2020)	GMM and KF, dataset: UCF Sports, UCF101 and KTH datasets	Analysis and feature extraction from every moment and every video frame	It took a long time to classify videos and was not very accurate, especially with more complicated datasets.	Accuracy = 89.30%
Xu et al. (2020)	GAF Algorithm, Dataset: WISDM, UCI HAR and Opportunity dataset.	It has good convergence speed	Its computational cost is high	GAF + Mdk-ResNet has a 96.83% success rate.
Alazrai et al. (2020)	E2EDLF, Dataset: Channel state information (CSI) dataset.	It accurately predicts the Human-to-Human Interaction.	This take some time to predict the interaction without using CSI	It achieved 86.3%

Zhu et al. (2020)	1D-CNNs and Long Short-Term memory (LSTM), Dataset: this work is trained and tested on a seven-class HAR data set	It can be used for monitoring and fault identification in real time.	It's a type of limited-parameter models known for their efficiency.	The result was an accuracy of 98.28%.
Ihianle et al. (2020)	CNN and Bi-LSTM, Dataset: WISDM and MHEALTH datasets	As a result, it vastly enhances activity recognition's derived features.	To learn efficiently, more training data was needed.	Accuracy = 98.6 ± 0.682
Maragathavalli et al. (2021)	Bidirectional Recurrent Neural Network (BRNN), dataset: Mobifall & MobiAct	It gives good accuracy than compared to other methods	It doesn't predict more complex activity	The accuracy is significantly increased by nearly 4% with the existing work.
Gupta (2021)	CNN-GRU, Dataset: WISDM Dataset.	It greatly simplified the process of developing these intricate models of deep neural networks.	It has slow convergence and low learning efficiency.	Accuracy 96.54%
Nouriani et al. (2022)	CNN-LSTM, Dataset: A dataset is collected from the 3 IMU sensors attached to the Body.	Using real-time orientation data from observers reduces the computational load on the deep learning network.	It requires a large Dataset to process and train the neural network	This method achieved 96.47%
Khatun et al. (2022)	CNN-LSTM, Dataset: MHEALTH and UCI-HAR databases	It can perform timing analysis while extracting abstract features	It significantly slower due to an operation such as maxpool.	It achieved an accuracy of 99.93%
Shi et al. (2022)	Multichannel Convolutional Neural Network With Data	It improves data quality in samples and provides better	The computational complexity is high.	The overall accuracy of AMC-CNN's recognition was

	Augmentation (AMC-CNN), dataset: WISDM and MHEALTH	feature discrimination.		95.53% on average.
Raziani et al. (2022)	Metaheuristic Algorithms with random search, Dataset: UCI HAR dataset.	It finds good solutions with less computational effort	If the optimisation problem is complex and/or very large	It doesn't predict the better accuracy.
Athota et al. (2022)	CMFA and CGFA, dataset: WISDM dataset	It gives high accurate prediction	They have a hard time classifying images with different positions.	It scored 97.76% on the CMFA's smartwatch and smartphone test, 94.98% on the CGFA's test, and 84.35% on the CGFA's smartphone test.
Ahmad et al. (2023)	CNN and Bi-GRU, Dataset: YouTube11, HMDB51, UCF101 Dataset	It uses less memory	It can't foresee how people will use devices connected to the internet of things (IoT).	It achieved high prediction of accuracy of 93.38% while using selected feature of YouTube11 Dataset.

2.1.3. Human Activity Recognition Using Spatio and Temporal Features

Ji et al. (2012) presented a unique 3D Convolutional Neural Network (3D CNN) model for human action recognition to tackle this problem. Both spatial and temporal features were extracted using 3D convolutions in this model. Therefore, they were able to record the motion data included in the sequence of neighboring frames. The model that has been created produces numerous streams containing information from the given input frames. These channels are then merged to create the feature representation. In order to enhance the performance of 3D CNN models, auxiliary outputs were included as high-level motion characteristics. This was achieved by using a method of model fusion and regularization. The supervised training approach is necessary for the 3D CNN model because to its dependence on a substantial quantity of labeled data.

A new dataset of videos from YouTube that fall into 487 classifications was used to propose CNN for large-scale video recognition (Karpathy et al., 2014). Multiple strategies were examined to enhance the connectivity of a Convolutional Neural Network (CNN) in the temporal domain, with the objective of leveraging local spatio-temporal information. Additionally, the adoption of a multi-resolution approach was recommended to expedite the training procedure. Nevertheless, a more robust methodology was necessary to integrate predictions at the clip level into overarching video-level predictions, so augmenting the process of action identification.

Complex human behaviors in video were first proposed to be recognized using a discriminative hierarchical model (Lillo et al., 2014) that operates on three semantic levels. Pose information for the body was encoded at the base level. At the intermediate level, encoded poses cover the ground covered by the composition of elementary human actions. This model captures, at a high level, how actions are composed in time and space to form complex human activities. In addition, the human activity classifier uses a discriminative modeling strategy to simultaneously model the appearance and composition of important body components and the action of interest. Using a max-margin framework for model learning, the proposed method not only provides useful intermediate-level semantic annotations, but also conducts effective multiclass discrimination. This model, however, proved less accurate.

Recently, a new hybrid architecture was introduced (Tompson et al., 2014) which involves the integration of a deep Convolution Networks with a Markov Random Field (MRF). The proposed approach utilizes a multi-resolution representation of features that is constructed by using overlapping receptive fields. Additionally, this model has the capability to do back-propagation and training using an approximate representation of Markov Random Field (MRF) loopy belief propagation, similar to the Part-Detector. However, there is a need to improve its performance effectiveness.

Cao et al. (2015) proposed hierarchical model, they first tackled the issue of complex activity recognition. Later, the spatial component was added to the TriCRF model. In this model, complicated interdependencies between action labels and activity labels were modeled concurrently. It has to be improved, though, by adding a stage for learning stance representations in concert with other actions.

(Park et al., 2016) employed the deep learning technique of a Recurrent Neural Network (RNN) to propose an innovative HAR system. A spatiotemporal feature matrix (i.e., a set of joint angles over time for different body joints) is used. Researchers use these extracted features to train and evaluate the RNN for HAR.

Nikolov et al. (2018) created This study proposes a novel approach for human activity recognition, which utilizes a Convolutional Neural Network (CNN) for classification and establishes correspondences between chosen feature points across adjacent frames to represent various human actions. An iterative principal component analysis (PCA) technique was used in order to simplify the data complexity. The pipeline receives the fully processed matrix from the preceding attempt or from another user doing the exact same function at every iteration. The real-time identification of human behaviors, however, presents inherent complexities and uncertainties.

(Zhao et al., 2018) suggested the Res-Bidir-LSTM network architecture. The deep network can enhance learning ability early in training to hasten the learning process. The new network is advantageous in several ways, one of which is that it supports bidirectional connections that can link both forward and backward in time. The second way that the gradient vanishing problem can be circumvented is by the use of residual connections between stacked cells. The proposed network improves recognition accuracy in two ways: temporally (through bidirectional cells) and spatially (by stacked residual connections).

For the purpose of action recognition in depth videos, (Liu et al., 2018) developed a multi-view hierarchical recurrent neural network (BRNN). As a result, the corresponding action recognition model needs to be able to identify human behavior from a wide range of perspectives inside the image. After being projected into three different planes, the 3D human action data is then input into a multi-view hierarchical BRNN that can account for the specifics of the 3D human action space. This model offers a practical approach to the modeling of temporal dynamics in multi-level video sequences. But the conventional RNN training procedure has issues with error explosion and vanishing gradient.

To identify human activity in recorded footage, (Hao et al. 2019) developed a Spatio-Temporal Distilled Dense-Connectivity Network (STDDCN). Knowledge

distillation and dense-connectivity were used in this approach. The impediment was created with the hope of discovering interaction laws between the shape and movement points of various designs. At the feature representation layer, extensive interactions between form and motion circuits permitted spatiotemporal interaction. Knowledge distillation between the two streams led to their eventual integration, which facilitated interaction between both points at the higher levels. It was also possible to derive useful hierarchical spatiotemporal characteristics. However, its precision proved insufficient for use with more intricate data sets.

To distinguish human activities using motion data, geographical features, and temporal relationships, (Majd et al. 2020) offer a Correlational Convolutional LSTM (C2LSTM) network. At first, the video sequence's spatial and motion data were credited using convolution and correlation methods. The proposed units were then used to develop a deep network for human-action recognition. However, its performance suffers on more intricate data sets.

For the purpose of extracting spatial and temporal information from a picture sequence, (Sánchez-Caballero et al. 2023a) describe The proposed model utilizes ConvLSTM, a form of Long Short-Term Memory (LSTM) that incorporates the convolution operation. Various methodologies based on deep learning concept, like the rate of learning ranging evaluation, cyclic training track, and batches normalization, have been used by researchers to optimize the efficacy of such structures. The proposed approach for performing HAR protects the privacy of the individuals in the scene by using just depth information. The neural networks were trained and evaluated using NTU's huge RGB+D dataset.

Table 2.3 Comparison of Human Activity Recognition Using Spatio and Temporal Features

Author and Year	Methods & datasets	Merits	Demerits	Accuracy (%)
Nikolov et al. (2018)	CNN and PCA algorithm, Dataset: NTU RGB+D, Kinect dataset	PCA analysis improves when a variety of users perform the same task.	Recognizing human behaviors in real time is challenging and noisy.	This method not provide a predicted accuracy.
Ji et al. (2012)	3DCNN Architecture, Dataset: TRECVID data	It has the ability to handle the large dataset.	Since it uses a supervised technique to train the 3D CNN model, it needs a sizable amount of labeled examples.	It achieved an overall accuracy of 90.2%
Karpathy et al. (2014)	CNN, Dataset: UCF-101 dataset, Sports-1M dataset	A data augmentation were used to reduce the overfitting problem.	Integrating clip-level calculations with global video projections requires a more rigorous strategy.	It achieved 65.4% accuracy.
Lillo et al. (2014)	CNN, Dataset: MSR Action3D Dataset, CAD120 Dataset	It minimize the computation cost	The accuracy of this model was less	It achieved 89.5% accuracy.
Tompson et al. (2014)	ConvNet and MRF, Dataset: FLIC Dataset, extended-LSP	They can be used in situations where the dependencies between variables do not have a clear direction.	It requires further improvement on performance effectiveness	It achieved 74.8 % accuracy
Cao et al. (2015)	TriCRF, Dataset: Composable Activity dataset	It reduces the computation cost	For progress to be made, it is vital to add the extra layer of learning posture representations in parallel with actions and activity.	It achieved the accuracy of 56.6%

Park et al. (2016)	RNN, dataset : MSRC-12 activity dataset	Ideal for deciphering time-based, sequential material like texts and films.	It is slow and complex training problem	Accuracy achieved 99.55%
Zhao et al. (2018)	Res-Bidir-LSTM, dataset: UCI, Public Domain UCI Dataset.	It effectively avoiding the gradient vanishing problem.	The grid search is not working	This method achieved accuracy of 93.6%.
Liu et al. (2018)	Multi-view hierarchical Bi-RNN, dataset: MSR, Action3D, DHA Dataset.	It's a great tool for simulating long-term temporal dynamics in videos.	Error explosion and vanishing gradient are two issues that plague the standard method of RNN training.	Accuracy with 94.15%
Hao et al. (2019)	Spatio-Temporal Distilled Dense-Connectivity Network (STDDCN), dataset: UCF101 and HMDB51 Dataset	It allows interactions between appearance and motion streams.	It requires more data during training	Model yielded the best results for the UCF101 with 96.7%.
Majd & Safabakhsh (2020)	Correlational Convolutional LSTM (C^2LSTM) network, Dataset: UCF101 and HMDB51 dataset.	This unit in enhancing LSTM units with motion information.	For more intricate datasets, its accuracy decreases.	Accuracy = 93.6%.
Sánchez-Caballero et al. (2023)	ConvLSTM, Dataset: NTU RGB+D dataset	It has lower computational cost	Continuous recognition of actions techniques in actual environments need a long-term spatiotemporal patterns recognition.	At an average of 0.21 seconds per video, the stateless model achieves a recognition accuracy of 79.91%(CV) to do the same task.

2.1.4. Human Activity Recognition Using Graph Convolutional Network

By starting with a skeletaltemporal graph and applying an existing transform, like the graph Fourier transform (GFT), to the graph signal defined on the resulting graph, (Kao et al. 2019) establish graph-based motion representations. It introduces an innovative method of using graph topologies to express data about human mobility. Skeletal-temporal graphs are used to model the human body, with the tracked joints serving as vertices and motion data as the signal. Time complexity for computing the feature representation is greatly reduced using the proposed method.

Using structure-based graph pooling (SGP) and joint-wise channel attention (JCA) modules, (Chen et al., 2020) present a unique graph convolutional network. Using what is previously known about the body's typology, the SGP scheme classifies the human skeletal graph into distinct categories. The model's one-of-a-kind attention mechanism helps it better distinguish between seemingly similar actions. However, this method drastically reduces both the number of parameters and the required computing power.

To address the issue of skeleton-based action recognition, Peng et al. (2021) introduced ST-GGN, a space-efficient and time-efficient spatial temporal global graph network. The ST-GGN cleverly offers a method of gathering data about global graphs using tensor rotation. Using this method, a topology matrix is not necessary to learn the graph embedding. Graph data extraction is also a breeze in Euclidean space. As a result, not only are fewer parameters required for the model, but also the topology of the graph can be learned more quickly. To build a dynamic topology for the skeleton, however, requires a computationally intensive technique.

Yuan et al. (2021) introduced a multiple filter dynamically graph convolution networks inspired by Inception. The proposed technique enhances the flexibility and accuracy of the model by using an evolving skeletal schematic architecture to produce distinct body links for various activities, rather than depending on a preset set of connections established by the source point.

The Manifold Regularized Dynamic Graph Convolutional Network (MRDGCN) is a network proposed by (Liu et al. 2021) that uses a dynamic graph

convolutional network that is regularized on a manifold. Manifold regularization is used routinely in the proposed MRDGCN to update the structural data until the model is fitted. In order to acquire high-quality structural data, it was necessary to construct an optimum convolution layer formulation. Therefore, MRDGCN is capable of automatically learning high-level sample characteristics to enhance data representation learning performance.

The study of (Liu et al. 2021) presents the geographic structure and temporal dynamics of a skeleton series. In addition, a view transformation module that can efficiently identify optimal perspectives for improved recognition is developed.

In order to represent skeleton data for HAE tasks like abnormality detection and quality rating, a two-task graph convolutional network (2T-GCN) was proposed (Bruce et al., 2021). Anomaly detection in skeletal exercise data can benefit from the proposed strategy. The 2T-GCN evaluation score can be used to assess the efficacy of behavioral interventions for people of advanced age.

The Hyper-Graph Neural network (Hyper-GNN) was introduced by (Hao et al. 2021) to extract spatiotemporal data and high-order correlations for HAR. Before applying the convolution technique to the hypergraph, the underlying skeleton graph was extended such that the hyperedge structure could describe the high-level correlations. The enhanced residual unit was then used to induce the spatial co-occurrence trait and incorporate the time-based correlation in order to capture more sophisticated features. In addition, the various aspects were fused together, and actions were identified, by employing a dynamic fusion of the 3-stream model. When the number of hyperedges, which could cause noise, was increased, however, accuracy dropped.

Using graph-based operations to learn action patterns, (Li et al. 2021) developed Symbiotic GNN (Sybio-GNN) to tackle HAR and motion prediction at the same time. A base, an activity-detection head, and a motion-estimation head make up this system. Similar complementing properties for HAR were trained using twin bone-based graphs and nets. However, it learned only the long-range joint linkages and ignored the short-range ones, as well as temporal details.

Table 2.4 Comparison of Human Activity Recognition Methods Using Graph Convolutional Network

Author and Year	Methods	Merits	Demerits	Accuracy(%)
Kao et al. (2019)	Graph Fourier Transform (GFT), dataset: MSR-Action3D and UTKinect-Action3D dataset	It allows to view signal in different domain	The computation time for the feature representation is much reduced.	It achieved better accuracy above 90%
Chen et al. (2020)	SGP scheme and JCA modules, SYSU-3D dataset	This technique can efficiently extract elements that reflect regional variations and identify murky behaviors.	It significantly reduces both the number of parameters and the computing time and effort.	It achieved accuracy with 79.2%
Peng et al. (2021)	ST-GGN, dataset: NTU-RGB+D and NTU-RGB+D 120	This not only makes learning network topology more efficient, but it also greatly minimizes the model's reliance on a large number of parameters.	The approach it provides to build the skeleton's dynamic topology is computationally intensive.	It achieved the accuracy with 88.2%
Yuan et al. (2021)	Multi-filter dynamic graph convolutional neural network, dataset: Kinetics-keleton.	Adaptability and accuracy were improved	Time and space complexity are higher.	Top-1 model accuracy was 32.8%, and top-5 model accuracy reached 54.8%.
Liu et al. (2021)	MRDGCN, dataset: CAS-YNU-MHAD, Unstructured social activity attribute (USAA) dataset, 2moons database, Citeseer dataset, Cora dataset	It helps data representation learning function better.	Multi-sets mean-pooling will result in the same average, hence it is not injective	It provides a recognition accuracy of between 30% and 10% on average.

Liu et al. (2021)	Adaptive multi-view graph convolutional networks, dataset: NTU RGB+D 60, NTU RGB+D 120, Northwestern-UCLA and UTD-MHAD.	It gives efficient result	It requires a large Dataset	Using UTD-MHAD dataset, it achieved high accuracy with 95.11%
Bruce et al. (2021)	Two-task graph convolutional network, dataset: UI-PRMD, Elderly home exercise dataset (EHE),	It has a high prediction rate	It has high computational cost	The outcome is achieved by training with a 100% accuracy across the board employing all skeletal joint characteristics.
Hao et al. (2021)	Hyper-Graph Neural network (Hyper-GNN), Skeleton dataset	It is effective on dealing with multi-modal data/features.	The accuracy was degraded while increasing the number of hyperedges, which may produce noise.	It achieved accuracy with 98.1%
Li et al. (2021)	Symbiotic GNN (Sybio-GNN), dataset: NTU-RGB+D, Kinetics, Human3.6M, and CMU Mocap	It is used to trained complex data	It used only the long-range joint relations	CMU Mocap dataset achieved 100% accuracy.

2.2 RESEARCH GAP

The HAR are processed using these prediction model for a given dataset. The expenses of managing massive data sets, however, may be prohibitive. Some approaches necessitate configuring appropriate parameters, while others necessitate extensive manual work prior to training a classifier. Recognizing and making use of pertinent features in the dataset is crucial to resolving this issue. HAR is difficult since a wrong recognition can have devastating consequences. Geometric relations among

joints often failed on large-scale datasets due to the limited representation ability of engineered features. These problems were ignored by the proposed approaches for HAR prediction.

2.3 CHAPTER SUMMARY

An in-depth comparison of the various methods for detecting signs of human activity is presented in this section. Considering these findings, the demerits of the all existing work are listed above. Although the results are improved by using these techniques, many obstacles have been discovered. These include computational complexity, accuracy, and the difficulty of predicting human actions. This study helps to identify the problem clearly and provide a road map for solving the HAR problem in this research.