Human Activity Recognition using Deep Learning Approach

Submitted in partial fulfillment of the requirements of the degree of

Bachelor of Science in Computer Science and Engineering

Developed By

Habibullah	18192103080
Pallab Majumdar	18192103050
Mafuja Akter Mitu	18192103068
Joy Adhikary	18192103062
Al Ahad Sufian	18192103056

Supervised By

M. M. Fazle Rabbi

Assistant Professor

Dept. of CSE, BUBT

Bangladesh University of Business and Technology - BUBT

May 2023

Declaration

We do hereby declare that the research works presented in this thesis entitled,**"Human Action Recognition using Deep Learning"** are the results of our own works. We further declare that the thesis has been compiled and written by us and no part of this thesis has been submitted elsewhere for the requirements of any degree, award or diploma or any other purposes except for publications. The materials that are obtained from other sources are duly acknowledged in this thesis.

Signature of Authors

Habibullah
ID: 18192103080

Pallab Majumdar
ID: 18192103050

Mafuja Akter Mitu
ID: 18192103068

Joy Adhikary
ID: 18192103062

Al Ahad Sufian
ID: 18192103056

Certification

This is to certify that Habibullah - 18192103080, Pallab Majumdar - 18192103050, Mafuja Akter Mitu - 18192103068, Joy Adhikary - 18192103062, Al Ahad Sufian - 18192103056, student of **Bangladesh University of Business and Technology**, has completed a thesis work on **Human Activity Recognition using Deep Learning Approach** under my supervision. The thesis work involved the development of a machine-learning algorithm for the Recognition of Human Activity from video footage data. The students conducted a comprehensive review of the literature on the subject, and proposed a novel approach for the detection of Human Activity Recognition using the Deep Learning Approach.

M. M. Fazle Rabbi

Assistant Professor & Supervisor

Department of Computer Science & Engineering

Bangladesh University of Business and Technology

Dhaka, Bangladesh

Approval

I do hereby declare that the research works presented in this thesis entitled,**"Human Action Recognition using Deep Learning"**, are the outcome of the original works carried out by Habibullah, Pallab Majumdar, Mafuja Akter Mitu, Al Ahad Sufian, Joy Adhikary under my supervision. I further declare that no part of this thesis has been submitted elsewhere for the requirements of any degree, award or diploma or any other purposes except for publications. I further certify that the dissertation meets the requirements and standard for the degree in Computer Science and Engineering.

M. M. Fazle Rabbi

Assistant Professor & Supervisor

Department of Computer Science & Engineering

Bangladesh University of Business and Technology

Dhaka, Bangladesh

Md Saifur Rahman

Assistant Professor & Chairman

Department of Computer Science & Engineering

Bangladesh University of Business and Technology

Dhaka, Bangladesh

Dedication

We would like to dedicate this research to our loving parents. Words cannot express how grateful we are for your love, support, and encouragement throughout our academic journey. Your unwavering faith in us has been our source of strength, and we dedicate this thesis work to you as a token of our love and appreciation. Thank you for always being there, for your unwavering love and support, and for instilling in us the values of perseverance, dedication, and hard work. We hope to make you proud and repay you for all that you have done for us.

Acknowledgement

At first we would like to thank Almighty Allah for his countless blessings and mercy upon us. Without his guidance, we would not have been able to achieve what we have today. We are also deeply thankful to Bangladesh University of Business and Technology - BUBT for providing us such a wonderful environment to peruse our research.

We would like to express our sincere gratitude to M. M Fazle Rabbi, Assistant Professor and Supervisor, Department of CSE, BUBT. We have completed our research with his help. We found the research area, topic, and problem with his suggestions. He guided us with our study, and supplied us many research papers and academic resources in this area. He is patient and responsible. When we had questions and needed his help, he would always find time to meet and discuss with us no matter how busy he was.

We also want to give thanks to Md Saifur Rahman, Assistant Professor & Chairman, Department of CSE, BUBT. We would also like to acknowledge our team members for supporting each other and be grateful to our university for providing this opportunity for us. Lastly special thanks to Google scholar for collecting so many papers.

Abstract

Human action detection and recognition is a topic that is constantly being researched in machine learning and deep learning. We can predict the human action recognition by using machine learning and deep learning. On this research work, both machine learning and deep learning are used in the research work. There are separate data sets for deep learning and machine learning. With the help of deep learning algorithms, the model can predict the amount of activity. How algorithms are working and their accuracy will be compared. Machine learning models work with classifier algorithms. Classifier Algorithms Model Algorithms used are Logistic Regression, Naive Bayes, SVM, Decision Tree. We will compare algorithmic Accuracy and facilitate data visualization to better understand the data set.

Contents

List of Figures

Chapter 1

1 Introduction

1.1 Introduction

Artificial intelligence is improving day by day[1]. Today people are trying to improve artificial intelligence in various ways. There are many topics that people are working on it to improve these sectors. Human action is one of the topics from these. Human action recognition or Human activity recognition works very well in creating interactions with the help of humans. A lot is known about a person through his activities. By identifying people activities machine can predict their personality and next step. That's why human action recognition or predict people activities now become one of the best topics to research on computer vision and machine learning. Activity reconstruction is being done in many ways such as video surveillance systems, human-computer interaction, and robotics for human behavior characterization.

1.2 Problem Statement

Now we see that many people use different types of devices such as running jogging etc. These devices work by human activity recognition. The world is now depended on data to capture data for people we need to recognize the activity of people. We can identify many accelerometers data where we can predict people running time. For example, people are running and count how many minutes they are running and how many minutes they are jogging, this prediction is based on human action recognition. Action recognition is working various way.

1.3 Problem Background

First, activity recognition development began with sensors and detectors. To identify communication, efficiency, system flexibility activity recognition is developing. Deep learning is helping a lot to improving this sector.[2] To developing the automated system activity, recognize system is very important. Modern science is trying to working with robotics. For future world, scientist is dreaming to work with robotics. Action recognition system is also very important to this sector.

1.4 Research Objectives

The objectives of our research work are as follows:

- Importance of activity recognition system.
- Identify the data from different types of data set.
- Working with data set to fit the model
- Visualize the data
- Generate sequential model to identify the data
- Identify the activity recognition

1.5 Contribution

- Working with different types of data set like video and accelerometers data.
- Trying to find out the recognize activities with different types.
- Visualize accelerometers data in different way.
- Verify all the model and trying to improve more accuracy.

1.6 Motivation

For Modern world we need to build up a system where robot can identify all types of data from human activities. There are many companies who are trying to identify activities from their employee. This sector is developing day by day so on our research work we are trying to research and trying a little contribution to developing computer vision and machine learning.

1.7 Flow of the Research

The research work is developing into several steps. First, we have analysed the research topics and then studied the basic theory action activity recognition. Then we identify problem statement and vulnerability. Then we finalized the dataset. There are two types of data. Figure 1.1 illustrates the overall steps to the research procedure in the following diagram.

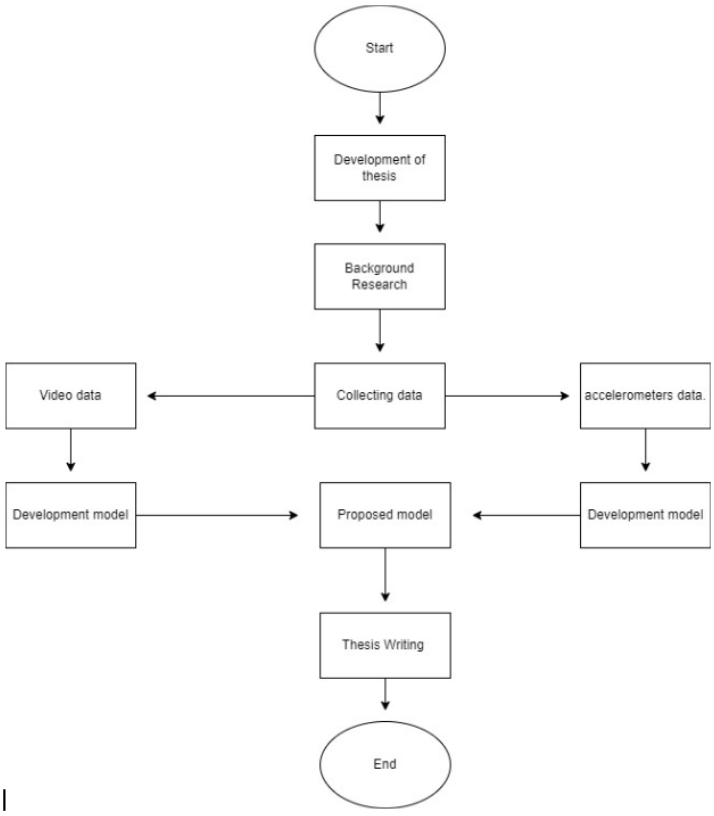


Figure 1.1: The figure illustrates the flow of the thesis work.

1.8 Significance of the Research

The central significance of this research topic is as follows-

- Activity recognition is one of the most important topics for modern world.
- As a result of this research, many applications, including video surveillance systems, human-computer interaction, and robotics for human behavior characterization, require a multiple activity recognition system.[3]
- This research topics include health care systems, activities monitoring in smart homes, Autonomous Vehicles and Driver Assistance Systems

1.9 Thesis Organization

The thesis work is organised as follows. Chapter 2 highlights the background and literature review on the field of the human activity recognition. Chapter 3 contains the human activity recognition system's proposed architecture and a detailed walk-through of the overall procedures. Chapter 4 includes the details of the tests and evaluations performed to evaluate our proposed architecture. Finally, Chapter 5 contains the overall conclusion of our thesis work.

1.10 Summary

This chapter includes a board overview of the problem that we aimed explicitly at our research work's objectives, the background, and the research work's motivation. This chapter also illustrates the overall steps on which we carried out ore research work.

Chapter 2

2 Literature Review

2.1 Introduction

Human activity recognition (HAR) is an important subfield of computer vision and machine learning that has the potential to help with a wide range of practical applications, such as surveillance, sports analysis, and healthcare.[4] Using data from numerous sensors, HAR attempts to identify and categorise human behaviours. Deep learning methods have made great strides in HAR in recent years, allowing for more precise recognition of human actions. In this survey, we provide a synopsis of current works in HAR that have employed deep learning methods. Methods by Kwon et al., Xu et al., Yang et al., Wang et al., Bhattacharya et al., Vijayaprabakaran et al., and Abdellaoui et al. have been specifically discussed. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and deep belief networks (DBNs) are just some of the deep learning designs and methodologies on display here.[5] These models have been tested on a variety of benchmark datasets, including UCF-101, HMDB-51, KTH, and Weizmann, and shown to perform well. According to the existing literature, some of the difficulties in HAR, such as dealing with complex environments, enormous datasets, and real-time applications, may be overcome by using models informed by deep learning. More work is required in this area, however, to produce models that are more precise and efficient at recognising human behaviours.

2.2 Literature Review

In their study, Kwon, Tong, and Haresamudram[6] provide a solution to the lack of large-scale, labeled datasets that currently plagues sensor-based human activity identification (HAR). The authors propose a method for retrieving virtual sensor data from free video sharing platforms like YouTube for the purpose of training HAR models. The method entails visual tracking information being translated into virtual motion sensors (IMUs), and then activity recognizers being derived from the produced virtual sensor data. The 3D human motion is estimated from a video segment using standard pose tracking and 3D scene understanding techniques. The authors show that their method works on three standard datasets and imply that it may pave the way for the human activity recognition community to tackle more difficult recognition tasks.

The novel strategy taken by the authors may be the key to overcoming the issue of insufficient data in sensor-based HAR. The authors' approach to extracting virtual sensor data from films is technically sound, and their use of publicly available video data is an ingenious technique to work around the lack of large-scale, annotated datasets. However, the authors recognise that there are many obstacles to overcome in order to make the vast amounts of existing video data usable for sensor-based activity recognition. To get around the limitations of learning from video-sourced motion information for eventual IMU-based inference, new forms of features and activity recognition models will need to be designed.

The studies performed by the authors demonstrate the effectiveness of their approach to the problem of sparse data in sensor-based HAR, and the research as a whole provides an innovative and promising solution to this problem. To learn the limitations of the method and develop new features and activity recognition models that can be used to train sensor-based activity recognizers using video data, however, more study is required.

In their recent study, Cheng Xu, Duo Chai, and Jie[7] He suggest a model for

HAR that uses deep learning, namely the combination of an inception neural network and a recurrent neural network, to identify different types of human behavior. (GRU). In order to extract multi-dimensional characteristics, the model uses kernel-based convolution layers to process data from a variety of sensor channels. In order to extract spatial and temporal data, it employs four modules that are quite similar to Google's Inception module. Python's Keras 2 was used to create the model since it has a high-level API for neural networks that can be used with either Tensorflow or Theano as a backend. The experimental findings across all three available datasets demonstrate that the proposed InnoHAR model outperforms the state-of-the-art models CNN and DeepConvLSTM.

This paper talks about the problems with the human activity recognition system that is currently in use. These problems include the interference of complex and changing backgrounds, the difficulty of positioning and recognising multiple active subjects, and the need for strict environmental conditions for light, brightness, and contrast. It also talks about the benefits of using deep learning technology in HAR, which can automatically extract and categorize complex information.

F1-score mean comparisons confirmed the results and showed that repeating techniques work well enough to be thought of for future innovative (real-time) applications in HAR. There are still ways to improve and learn more about the InnoHAR model, but it has shown promise and could be used in real-time applications. The author's technique and experimental results shed light on human activity recognition and should be of interest to those working on the topic.

This paper proposes two approaches[8] for real-time human action recognition (HAR) from raw depth video sequences using ConvLSTM. One approach uses a stateless model with a video-length adaptive input data generator, while the other approach explores the stateful ability of general recurrent neural

networks. Both models only use depth information to preserve privacy. The models are trained and tested on the large-scale NTU RGB+D dataset, and experimental results show that the proposed models achieve competitive recognition accuracies with lower computational cost compared to state-of-the-art methods. The stateful model significantly improves accuracy compared to the standard mode, achieving recognition accuracies of 80.43% and 79.91% with an average time consumption per video of 0.89 s. The stateless model achieves recognition accuracies of 75.26% and 75.45% with an average time consumption per video of 0.21 s .

The NTU RGB+D dataset has also been used for the test phase of the proposed methods. The authors of this dataset suggest two different evaluations: cross-subject (CS), where 40 320 samples recorded with 20 subjects are dedicated for training and 16 560 samples with 20 different subjects for test; and cross-view (CV), where 37 920 videos were recorded with 2 cameras from different viewpoints and 18 960 videos from a third different viewpoint for test.

This paper presents two novel approaches using ConvLSTMs for human action recognition, which aim to boost time efficiency performance while keeping competitive accuracy rates. The authors leveraged the stateful capability of LSTMs and proposed an input data generator to learn long-term characteristics of videos of variable lengths. The results on the NTU RGB+D dataset showed that both proposed models reach competitive accuracy rates with very low computational cost compared with state-of-the-art methods. The stateful ConvLSTM achieved better accuracy rates than the stateless ConvLSTM, proving the effectiveness of this uncommon methodology for videos.

The paper "Vision-based human action recognition: An overview and real world challenges"[9] provides an overview of vision-based human action recognition and the challenges in applying these techniques in real-world scenarios. The paper highlights the importance of human action recognition in various applications and discusses different approaches to vision-based human action

recognition, such as template-based methods and model-based methods. The authors also discuss the challenges in applying these methods in real-world scenarios, such as variations in illumination and occlusion.

The paper mentions several datasets that are commonly used in vision-based human action recognition research. The Hollywood (2008) dataset contains short video clips of actors performing different actions, while the UCF Sports (2014) dataset contains video clips of people performing sports-related actions. The HMDB51 dataset contains video clips of people performing everyday actions, while the PHAV dataset contains videos of people performing household activities.

The paper highlights the importance of continued research in this field and the need for novel techniques to address the challenges faced in real-world scenarios. The authors suggest that the development of more robust and efficient algorithms, the use of multiple modalities, and the integration of deep learning techniques could help to improve the accuracy and applicability of vision-based human action recognition in the future.

The experiments by Bevilacqua[10] identify a deep learning approach based on CNNs, but are based on exercises rather than common activities. Their dataset was derived from specially placed sensors on the lower half of the body, which makes it less applicable to a large user base, as most people just have smartphones or smart watches

The experiments by Weiss[11] use five distinct algorithms: Random Forests, J48 decision trees, B3 instance-based learning, Naive Bayes, and multi-layer perceptron. Using these algorithms, Weiss was able to obtain an overall accuracy rate of 25.3% with the phone accelerometer data, and 64% accuracy with the watch accelerometer data..

Reining et al.[12] performed a systematic literature review of HAR for produc-

tion and logistics. This survey presents a detailed overview of state-of-the-art HAR approaches along with statistical pattern recognition and deep architectures. This study is beneficial for industrial applications.

Beddiar et al.[13] surveyed vision-based human action recognition and categorized the entire study into the following fields: Handcrafted-feature and feature learning-based approach, where authors discussed the various techniques, including their implementation details. The authors also highlight related literature based on human activity types – Elementary human actions, Gestures, Behaviors, interactions, group actions, and events, which advocate HAR approaches at the minute level..

Similarly, Zhu et al.[14] also examined both handcrafted and learning-based approaches for action recognition. Unlike, the authors first evaluated the limitation of the handcrafted method then shows the rise of deep learning techniques of HAR in brief, till 2016.

In Tushar D. et al.[15] proposed a deep learningbased technique using binary motion image (BMI), the authors used Gaussian Mixture Models (GMM) to subtract binary backgrounds used to create BMIs (Fig. 1), and three (3) CNN layers to extract features and classify activities. BMIs are extracted from the frames, which make the use of this approach impossible for multi-human recognition.

Moez B. et al.[16] proposed in , a two-steps neural recognition method (Fig. 2) using an extension of convolutional neural network to 3D to learn spatialtemporal features. The authors proposed to extract the features using 10 layers of CNN (input layer, two combinations of convolution/rectification/sub-sampling layers, a third convolution layer and two neuron layers). For the recognition step, they proposed to use a recurrent neural network classifier (Long Short-Term Memory (LSTM) classifier) .

In Pichao Wang et al.[17] used a weighted hierarchical depth motion map (WHDMM) and three channel deep CNN for human activity recognition. Here, the authors proposed to feed three separate ConvNets using WHDMMs constructed by the projection the 3D points of depth images to three orthogonal planes, the final classification decision is obtained by the fusion of the three ConvNets.

In Zuxuan W. et al.[18] constructed a hybrid method for video classification by extracting two types of features from spatial frames (raw frames) and short-term stacked motion optical flows using convolutional neural network (Fig. 3). These features are used to feed two separate LSTM networks for fusion and classification.

The authors in [19] proposed a human activity recognition approach using depth images, from which they proposed to extract three derived images (Fig. 4): Average depth image (ADI), Motion history image (MHI) and depth difference image and used a deep belief network (DBN) using a Restricted Boltzmann Machine (RBM) .

Andrej K. et al. in [20], proposed a multi-resolution convolutional neural network approach (Fig. 5), here the authors used two ConvNet channels, the first channel is fed by a context stream representing a lowresolution image; the second one is fed by a fovea stream representing a high resolution centred image. The two channels converge towards two fully connected layers.

In Simonyan et al.[21] presented a two-stream architecture for video classification (Fig. 6), the authors proposed to use a spatial stream ConvNet using raw video frames to carry information about the objects and the general spatial information in the scenes, and a Temporal stream ConvNet using optical flow from multiple input video frames. Classification is done using two fusion methods: the average of the two stream scores or by using a multiclass SVM

on Softmax scores.

Limin Wang et al.[22] propose a novel architecture called Temporal Segment Networks (TSN) for video-based action recognition, which incorporates a segment-based sampling strategy to improve temporal modeling in deep neural networks. They used a 2D CNN, features extraction standard stochastic gradient descent methods including the segment-based sampling strategy, the consensus function, and the use of multi-scale and multi-modal inputs, and show that TSNs achieved an accuracy of 94.5%. On Something-Something V1 and V2, TSNs also achieved an accuracy of 49.7

Sukrit Bhattacharya et al.[23] perform an analysis of human activity recognition based on the video that uses both spatial and temporal information from video frames to make predictions. They proposed a method, called SV-NET, which consists of two separate streams: a spatial stream that processes individual frames using the VGG-16 network and a temporal stream that processes stacked optical flow images using a 3D CNN, and the output is fed into a softmax layer for final classification. After training and testing the proposed model, they got an accuracy of 96.3% on UCF101, 75.3% on HMDB51, and 80.7% on Hollywood2.

Vijayaprabakaran K, Sathiyamurthy K, and Ponniamma M[24] Proposed a deep learning approach for video-based human activity recognition for the elderly, with a focus on recognizing activities of daily living. They used CNN-based architecture for video-based human activity recognition, Specifically, the VGG-16 network as a feature extractor and data augmentation techniques such as random cropping and flipping to improve the robustness of the model. They got an overall accuracy of 97.5%. for recognizing six different activities of daily living: walking, sitting, standing, lying down, bending, and turning.

Mehrez Abdellaoui et al, Ali Douik et al [25] propose a deep learning approach

for human action recognition in video sequences using Deep Belief Networks (DBNs). The authors proposed a method of deep neural networks that uses a two-stage approach for action recognition, pre-trained DBNs, and unsupervised learning, a Support Vector Machine (SVM) classifier. The authors show that their proposed method achieves high accuracy of 95.6% on the KTH dataset, and 90.2% on the Weizmann dataset.

Abdellaoui, M., Douik, A. et al[25] proposed a brand-new DBN-based HAR technique, attempts to increase the accuracy of human activity categorization. Segment the video clips from the human activity dataset into frames as a first step. The result is then converted to binary frames, and the resulting frames are subjected to a series of morphological filtering techniques in order to improve their quality. Then, generate an input matrix containing the training data, the testing data, and their labels by converting the new frames into a binary vector. The input data for our DBN architecture are represented by this matrix. The DBN classifier will be trained using the training data matrix in the last phase, and the classification outcome will be extracted.

In this paper,[26] the authors propose a new deep-learning model that combines 3DCNN with ConvLSTM layers to optimize traditional 3DCNN for human activity recognition. Their experimental results demonstrate the superiority of the 3DCNN + ConvLSTM combination for recognizing human activities using the LoDVP Abnormal Activities dataset, UCF50 dataset, and MOD20 dataset. The proposed model is well-suited for real-time human activity recognition applications and can be further enhanced by incorporating additional sensor data.

Overall, this work shows promise for improving the accuracy and efficiency of human activity recognition tasks using deep learning models. The combination of 3DCNN and ConvLSTM layers can capture both spatial and temporal features, enabling accurate classification of videos with high precision. Further

research in this area could lead to more effective surveillance systems that can enhance safety and security in various settings.

The paper proposes[27] a novel model based on a modified capsule network (MCN) for accurate human activity recognition (HAR). The model consists of a convolution block and a capsule block, which preserves the spatial relationship among features through a dynamic routing process. The effectiveness of the proposed model is validated on a human activity dataset collected through an inertial measurement unit (IMU) and achieves an accuracy of 96.08%, outperforming the traditional convolutional neural network (CNN) with an accuracy of 91.62%. The proposed model is also evaluated on two public datasets, WISDM and UCI-HAR, achieving accuracies of 98.21% and 95.28%, respectively, which surpass the reported results obtained from benchmark algorithms like CNN.

This paper presents[28] a novel approach for human activity recognition using a self-attention-based neural network architecture. The dataset used for this study was collected using smartphones, which are equipped with sensors such as accelerometers and gyroscopes to collect data related to human activities. The proposed architecture was compared to two strong baseline models, namely, convolutional neural network (CNN) and long-short term network (LSTM). Various components of the self-attention model, such as dropout, scaling factor, and positional encoding, were also investigated to find the best model. The results showed that the proposed model achieved a test accuracy of 91.75%, which is comparable to the baseline models. The application of this approach can be useful in assisted living technology for elderly people. The paper also provides a review of the previous works in human activity recognition, including the use of SVM, RF, CNN, RNN, and their variants. The paper concludes by highlighting the importance of combining different machine learning algorithms and techniques to improve the accuracy of human activity

recognition.

This paper presents a novel methodology for remote health monitoring through Human Activity Recognition (HAR) using wearable sensors.[29] The proposed method extracts power spectra to reduce feature extraction complexity using the HuGaDB dataset, where sliding windows techniques and a QRS algorithm are utilized to determine the dominant spectrum amplitude. The bandwidth algorithm is then used to remove redundant dimensions and improve feature extraction, achieving an accuracy rate of 95.1% in HAR with 70% of bandwidth. This approach outperforms others in human activity recognition accuracy, making it promising for use in rehabilitation and physical therapy for the detection of gait disorders.

The paper describes an efficient and lightweight deep learning-based approach for human activity recognition (HAR) using radar signals.[30] The approach decouples the Doppler and temporal features of radar preprocessed signals according to the feature representation of human activity in the time-frequency domain. The Doppler feature representation is obtained in sequence using the one-dimensional convolutional neural network (1D CNN) following the sliding window. Then, HAR is realized by inputting the Doppler features into the attention-mechanism-based long short-term memory (LSTM) as a time sequence. The proposed method achieves an accuracy close to 96.9% on two human activity datasets, and it has a more lightweight network structure compared to algorithms with similar recognition accuracy. The method is expected to have great potential for real-time embedded applications of HAR. The paper also describes traditional methods for MDF categorization, which are limited by classification complexity, and previous approaches to deep learning for radar-based HAR using CNN and RNN.

This is a survey paper that discusses the application of transfer learning in

vision sensor-based human activity recognition (HAR).[31] HAR is essential for industrial and institutional operations, but traditional machine learning algorithms assume that training, validation, and testing data come from the same domain, which is not always true in the real world. Transfer learning can compensate for outdated data, reduce the need for recollecting training data, and improve the accuracy of testing data. The paper reviews around 350 related research articles from 2011 to 2021 and selects approximately 150 significant ones that give insights into various activity levels, classification techniques, performance measures, challenges, and future directions related to transfer learning enhanced vision sensor-based HAR. The survey is the first to link machine learning, transfer learning, and vision sensor-based activity recognition under one roof.

The article[32] discusses the challenges of deploying activity recognition models in real-world scenarios, particularly the privacy concerns associated with collecting large amounts of sensor data and the divergence in the distribution of sensory data among individuals. To address these issues, the authors propose a diversity-aware activity recognition framework called DivAR, which uses a federated meta-learning architecture. The proposed framework clusters individuals based on their behavioral patterns and social factors and uses a centralized embedding network and individual-specific features to extract general sensory features shared among individuals. The authors conduct two data collection experiments and survey social characteristics, including personality, mental health state, and behavior patterns. The empirical results show that DivAR has a better generalization ability and achieves competitive performance on multi-individual activity recognition tasks. The article also discusses existing approaches to activity recognition, including domain adaptation techniques and FL-based approaches, and highlights the limitations of these approaches in addressing the heterogeneity challenge in activity recognition.

This is a research paper that reviews various research works on human action recognition.[33] The authors develop artificial intelligence models to recognize human activity from a dataset provided by UCI's online repository. They use various machine learning classification techniques such as Random Forest, kNN, Neural Network, Logistic Regression, Stochastic Gradient Descent, and Naïve Bayes to analyze human activity. The dimension of the dataset is reduced by the feature selection process, and precision and recall values are calculated, and a confusion matrix is created for each model. The experiment results show that neural network and logistic regression provide better accuracy for human activity recognition compared to other classifiers, although they require higher computing time and memory resources. The authors propose a system architecture and a methodology that includes Python and machine learning. The primary objective is to develop an efficient human action recognition system using multiple views, and the secondary objective is to understand human behavior models using probabilistic action graphs.

The automatic detection of human behaviors in surveillance footage is explored by Ding, Chunhui Fan, et al.[34] The majority of currently used methods rely their classifiers on very complex characteristics that are computed from the raw inputs. In this study, a brand-new 3D CNN model for action recognition is developed. The developed model generates multiple channels of information from the input frames, and the final feature representation combines information from channels. further, boost the performance and propose regularizing the outputs with high-level features and combining the predictions of a variety of different models.

Convolutional Neural Networks (CNNs), a strong group of models for image recognition issues, have been proposed by Andrej Karpathy et al,[35] Encouraged by these findings, present a thorough empirical analysis of CNNs' performance on large-scale video classification . Investigate several ways to

strengthen a CNN's connection in the time domain so that it can take use of local spatiotemporal data, and suggest a multiresolution, foveated architecture as a potential fasttracking option. By retraining ,the generalization performance of our best model (63.3 percent up from 43.9 percent).

Yu-Wei Chao et al,[36] proposed a better method for locating temporal activity in video, and was modelled around the Faster RCNN object identification framework. By using a multi-scale architecture that can accommodate extreme variation in action durations, TAL-Net can improve receptive field alignment, extend receptive fields for proposal generation and action classification. The model achieves state-of-the-art performance for both action suggestion and localization, as well as competitive performance on the Activity Net challenge.

In addition, Vrigkas et al.[37] categorized human activity recognition methods into two main categories including "unimodal" and "multimodal".Then, they reviewed classification methods for each of these two categories.

Presti et al. [38] provided a survey of human action recognition based on 3D skeletons, summarizing the main technologies, including both hardware and software for solving the problem of action classification inferred from time series of 3D skeletons.

2.3 Summary

This chapter investigate and reviewed the latest techniques of brain tumor detection and segmentation. The thesis's target is to eliminate the imperfection as much as possible and introduced a new way to detect brain tumor.

Chapter 3

3 Proposed Model

3.1 Introduction

The objective of this study is to prediction the human action. When people are walking, running or doing anything the model will predict the activity of those people. We have used deep learning model to create and the development of the model. There is a large number of data which we have used for the model. When the output from video will show us the activity percentage of people.

3.2 Data Analysis and Pre-processing

There are two types of data for our project. The first one is video data. Where there are many video labels data. Another one is accelerometer data. Accelerometers are widely used to measure sedentary time, physical activity, physical activity energy expenditure (PAEE), and sleep-related behaviors, with the ActiGraph being the most frequently used brand by researchers. [39] The accelerometer data need to label. These data are not labeled so these data are labeled. There is no null value on accelerometer data. They are a few categories of these data like:

- Downstairs
- Jogging
- Sitting
- Standing
- Upstairs
- Walking

These categories are showing the classification of this video. We have analyzed the categories of this video.

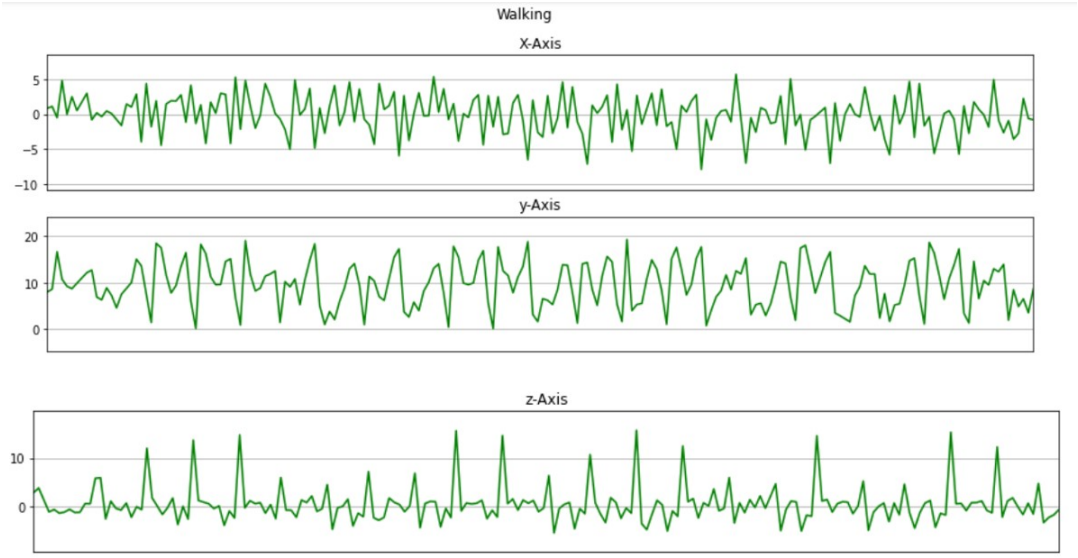


Figure 3.2: Walking category

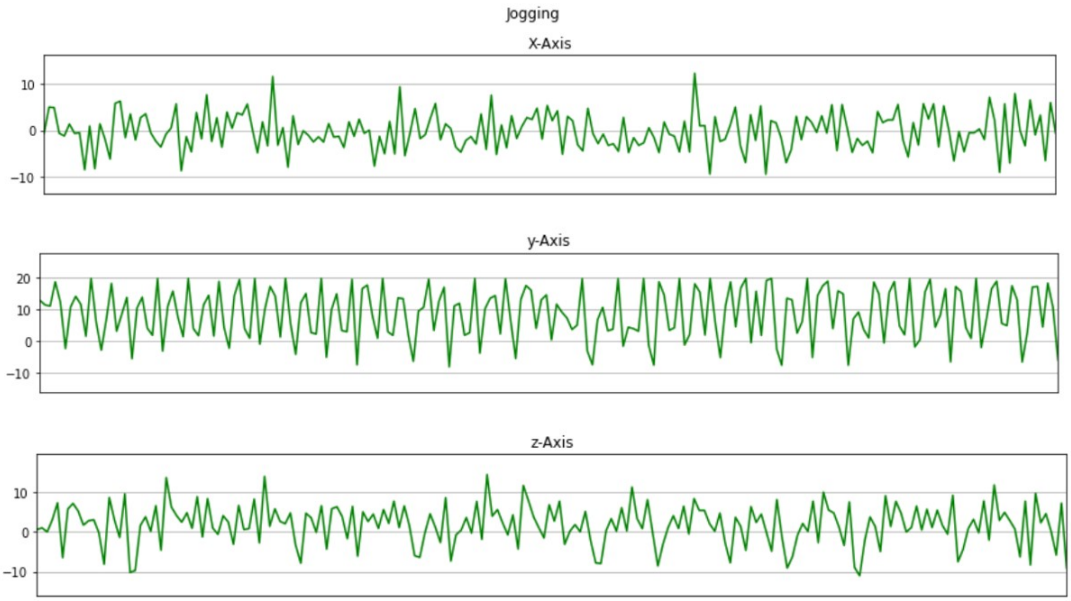


Figure 3.3: Jogging category

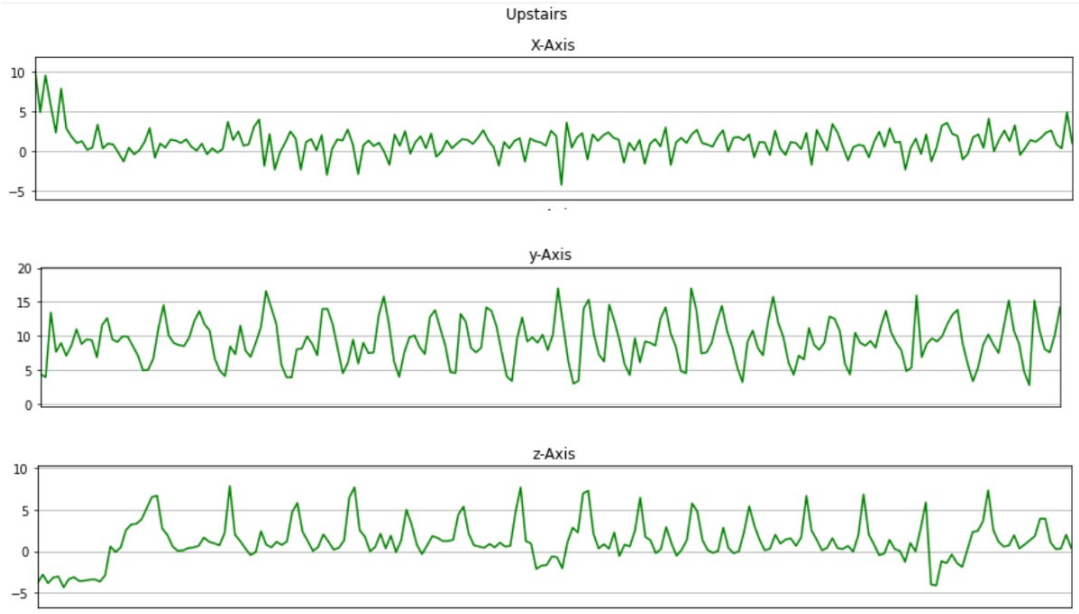


Figure 3.4: Upstairs category

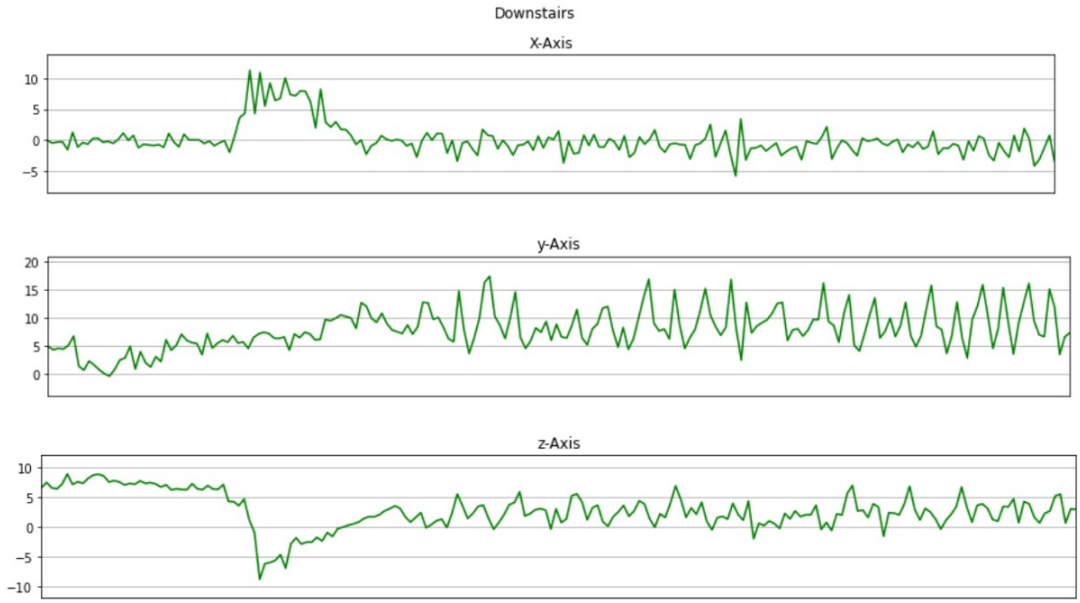


Figure 3.5: Downstairs category

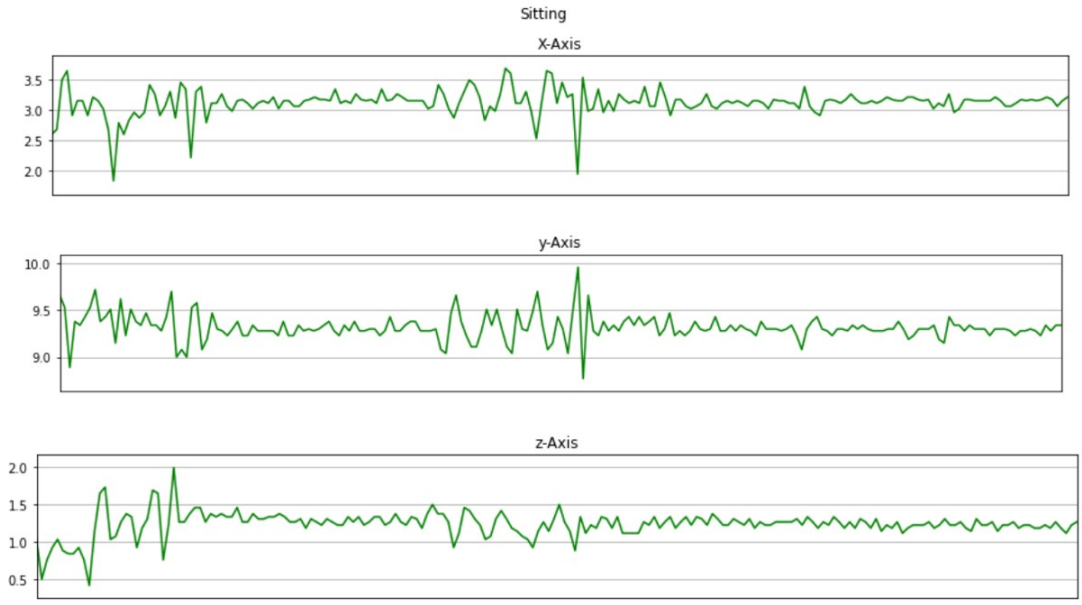


Figure 3.6: Sitting category

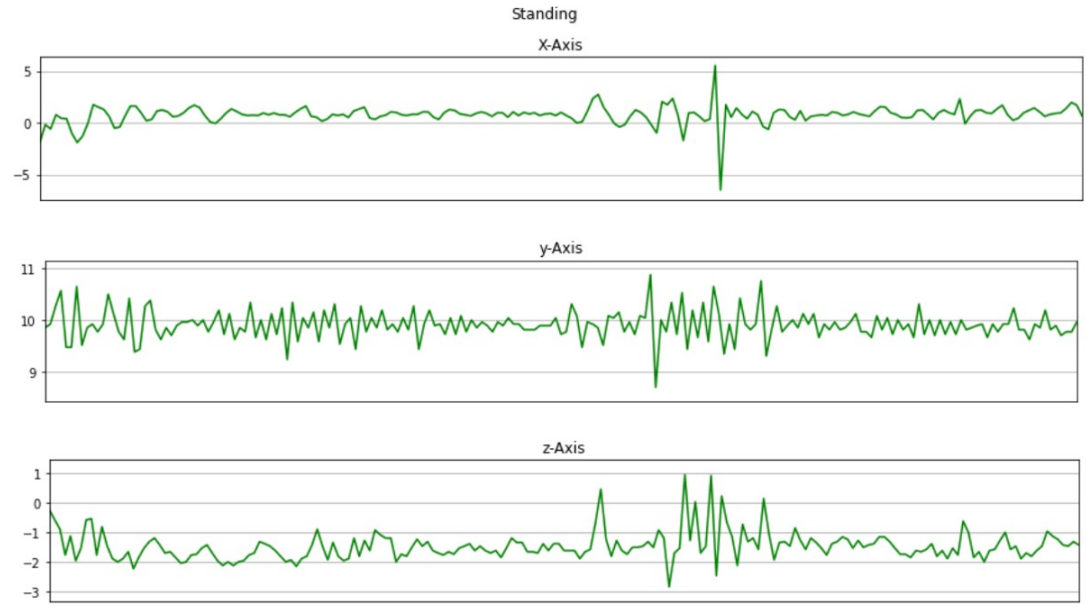


Figure 3.7: Standing category

The raw data was not balanced. There is a lot of data, the total counted data are:

- Walking (137375)
- Jogging (129392)

- Upstairs (35137)
- Downstairs (33358)
- Sitting (4599)
- Standing (3555)

As these data are not balanced so data will not work properly. All data is measured by standing data. So total data will show balanced.

3.3 Model Development

There is a total of three parts of our model development, we are using the sequential model on this part. Lastly, we will see the confusion matrix from our given data dataset. Here are three parts of our model development:

- First, we will predict the CNN-LSTM model using TensorFlow from the UFC101 data set, there will be showing the video and the activity part of the test sample.
- Second, there will be showing the percentage of the video and there will be feature extraction using CNN model.
- There will be output from accumulator data, and there will confusion matrix to understand the output of the model. Since the accommodators data has covert into normalization so we used the CNN-LSTM.

3.3.1 CNN And CNN-LSTM model

After completing the data processing from each model, there will be three models to predict the output. Simply placing the Keras layers in a sequential order is the fundamental concept behind Sequential API, hence the name. The majority of ANNs also have layers that are arranged in sequential order, and data flows from one layer to the next in the designated order until it eventually reaches the output layer. Artificial intelligence is now can capable of doing and

understanding human behavior. [40] So nowadays, the model can understand human behavior. Machine learning one of the best fields is deep learning where the model is inspired by the brain function and it is working like the brain function. [41] By the help of working like a brain function model can learn self-learning so that they can predict the behavior. A perceptron is trained by providing it with numerous training samples and computing the output for each one. The weights are modified after each sample to reduce output error, which is typically characterized as the disparity between the desired (target) and actual outputs. The perceptron's capacity to learn classification is crucial since many cognitive processes rely on classification. Similar algorithms could learn to have the same behavior as humans. So, if the model will calculate the same behaviors from the first layer it will process another second layer. The layer will pass the data consistently and finally, the output will show the result of which activity from the human.

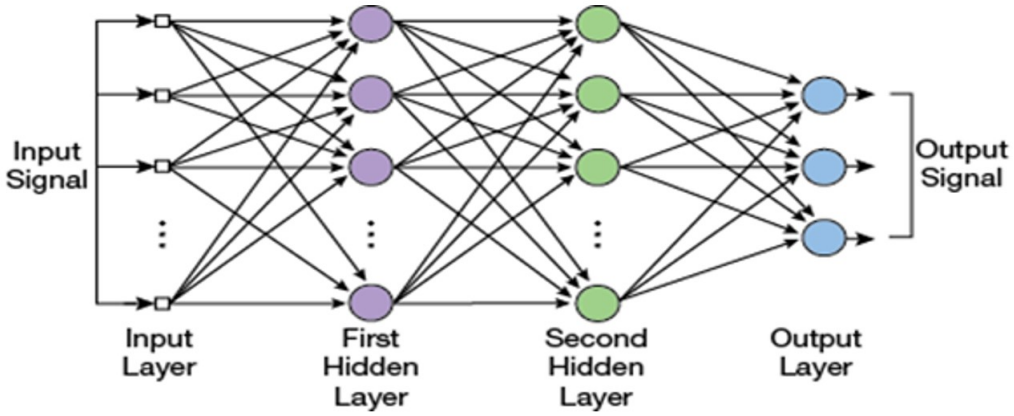


Figure 3.8: Layers from the sequential model

Sometimes there is almost similar activity from different label so in this case the layer is showing almost same but there will be different types of layer and data will go through those layers so that the input signal and output signal will be the same.

There is an arrangement sequence, so when there is a video to put into the algorithm, the frames will search for this functionality, which will be read from

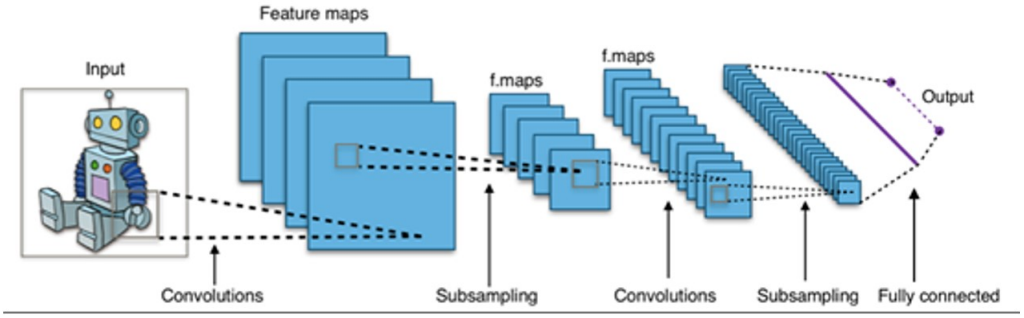


Figure 3.9: Architecture of sequential model

the videos. Then frames will be extracted from the video until the maximum frame count is reached. Most of the time there will be different frames, so when there are different areas in every video, those data will be put into the batches. For the solution of this problem, padding is used on the model. So that number of equal frames is in real shape. The image size is 224, and there are two functions on the CNN model. One will crop the center square, and another will load the video function into the CNN algorithm. Max sequence length and a number of features are important because now there will be a CNN network, which means a convolutional neural network.

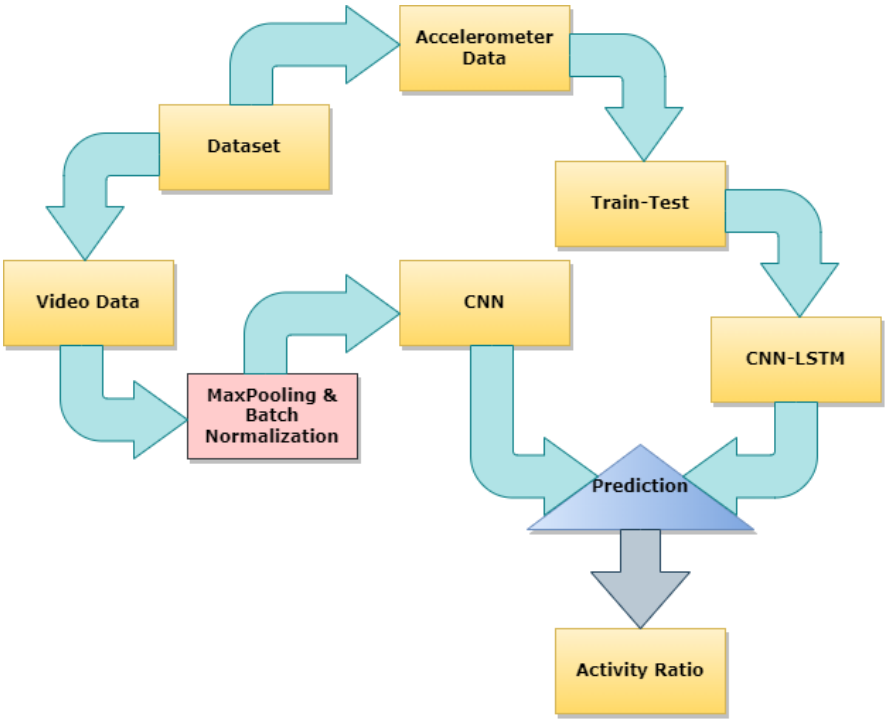


Figure 3.10: Proposed hybrid CNN-LSTM architecture to recognize human activity.

As you can see, the pad is the sequence and also the text. The CNN model also shows us how much data and what kind of data there is. There are two important variables: frame feature input and mask input. These two variables are related to the CNN model. Mask input and frame feature input provides the first layer of our model, and the final layer is activation layer softnext. The activation function is using OK for this CNN model variable. So the pad frames and these kinds of steps towards sequence prediction Sequence prediction will predict the sample of the video.

The algorithm of the proposed sequential to evaluate the model's performance is shown in Algorithm 1.

Algorithm 1 Evaluation Process of sequential Model

```
loadvideo ();
dataset ();
splitData ();
loadModel ();

for each epoch in epochNumber do

        for each batch in batchSize do

            y' = model (features);
            loss = crossEntroty (y, y');
            Optimization (loss);
            Accuracy ();
             bestAccuracy = max (bestAccuracy, accuracy);
return
```

Figure 3.11: Algorithm of the CNN-LSTM for performance evaluation.

3.4 Summary

This chapter we shown models which we have used to complete the thesis project. We have shown our workflow, data ratio, input sample, our proposed architecture. And we also show our performance evaluation algorithm.

Chapter 4

4 Experimental Results

To evaluate the performance, we have used the sequential model. There are three models first model we have by using TensorFlow.Another two models have been used in the sequential model. There are two parts of the data set where there are some videos which are in train data another is test data. By using the train data, test data will predict. Another data is the accumulator data. Where there will be X, Y, and Z data which we will get from the censor. There will be many labels. Where we used our model to predict the model. There are two parts of these data there are train and test data. To get these train data and test data, we have used train test spilled to identify the train and test data.

Layer	Output Shape	Parameter
Input Layer	79*2*16	80
Inception-ResNet-V2	78*1*32	2080
MaxPoolong2D	78*1*32	0
Dense	64	159808
Dropout	2*64	0
Reshape	80*3	0
Flatten2	2496	0
Dense2	6	390
Sigmoid	1	0

Figure 4.12: Summary of the proposed model CNN-LSTM

4.1 Performance Evaluation-Metric

The effectiveness of the implemented and trained neural model is discussed here. There are many labels from our train data set where we will predict the data. There are many labels on each model. We have to predict the activity using these models. Finally, a Confusion matrix [42] is extracted to understand each label and status.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Recall = \frac{Tp}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$F1score = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{4}$$

A confusion matrix is a graphical representation of a classification process showing the degree to which the model's predictions agree with the true effects [43].A confusion matrix aids in visualizing the results of a classification task by providing a table arrangement of the various outcomes of the prediction and findings.

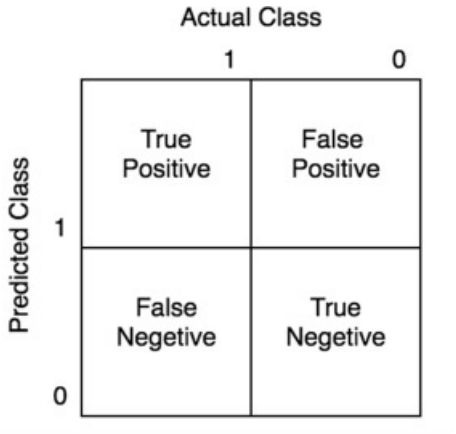
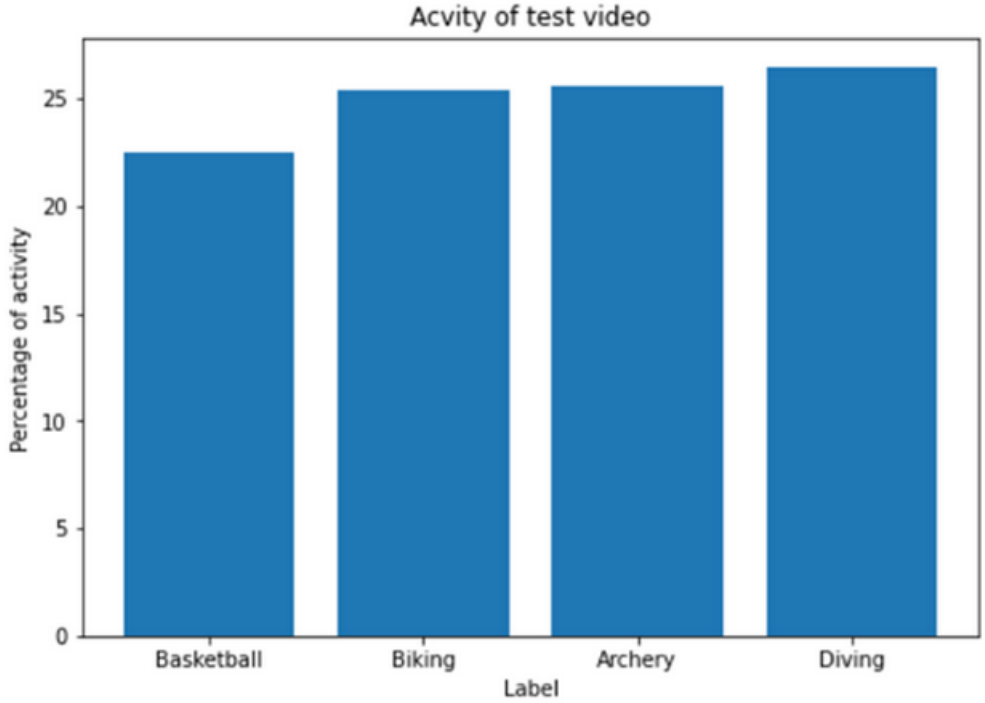


Figure 4.13: Confusion matrix of proposed model

- True positive: How frequently our real positive values match the anticipated positive. You correctly predicted a positive value, which is what it is.
- False positive: How many times our model mis predict positive values as negatives. In spite of what you had projected, the value is positive.
- True negative: How frequently our real negative values match our expected negative values. You correctly predicted a negative value, and that is what it is.
- False negative: How frequently our model predicts positive values for negative values in error. It is positive, contrary to what you had projected.

4.2 Results and Discussion

The proposed sequential model one is evaluated with an accuracy is 65%, loss = 0.9911, and value loss= of 1.1389. The output is showing for each test video. Here is the test video output:



Our model's two accuracies throughout training and validation are shown in Fig. 4.7.The Keras sequential model evaluation. We measured the training and validation precision while experimenting with various epoch counts. After 10 epochs, the model achieves its maximum accuracy in training and testing as well as verification. In the epoch no. 5, the model sustains in the low accuracy and it varies all the way with no. of epoch changes. The accuracy of the model is 92%, loss=0.25.41, and value loss=0.2056.

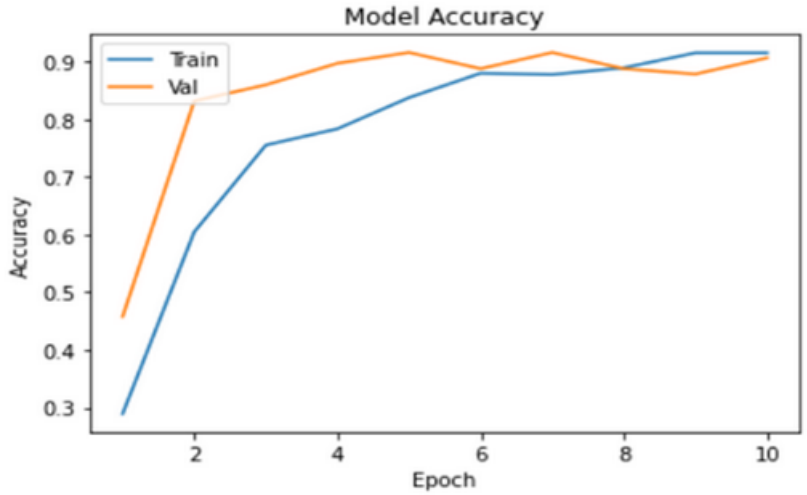


Figure 4.14: Activity of test video

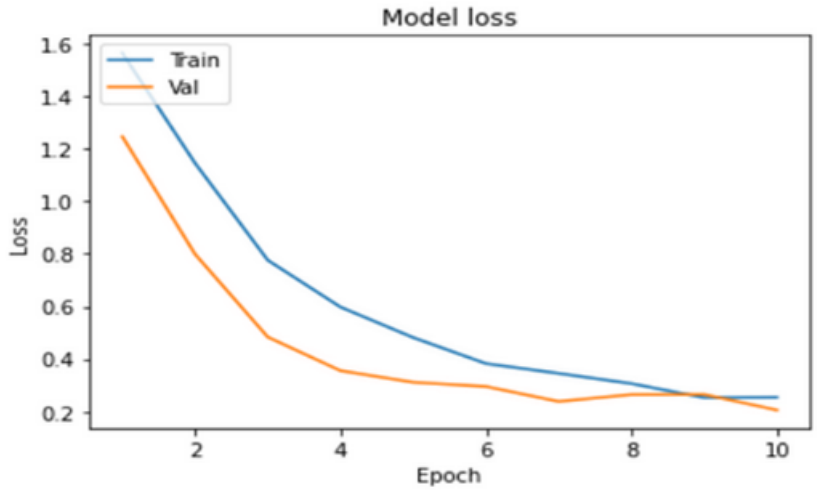


Figure 4.15: Accuracy of the Sequential model

Here is the model confusion matrix.Where we will see the output of the labeled data.

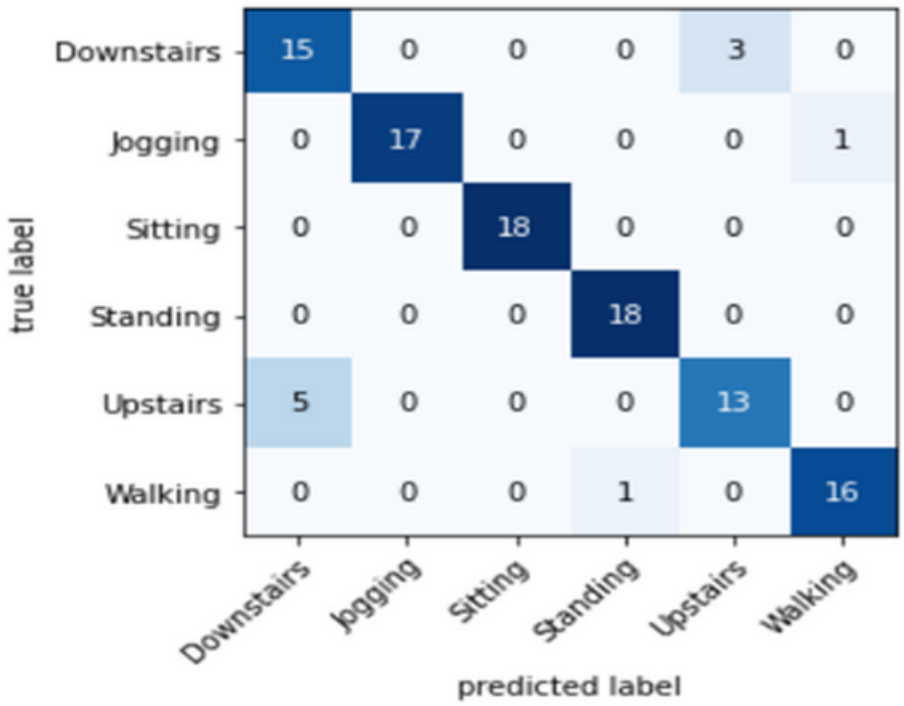


Figure 4.16: Labes of different catagory

We will now see the output of model one.Here is the showing a sample of the video:

Figure 4.17: Demo photo from video dataset

Here is the showing of the activity of the predicted label. After training the sample video then it will show the predicted output.

```
Top 5 actions:
  roller skating          : 96.85%
  playing volleyball      :  1.63%
  skateboarding           :  0.21%
  playing ice hockey      :  0.20%
  playing basketball      :  0.16%
```

Figure 4.18: Top Actions

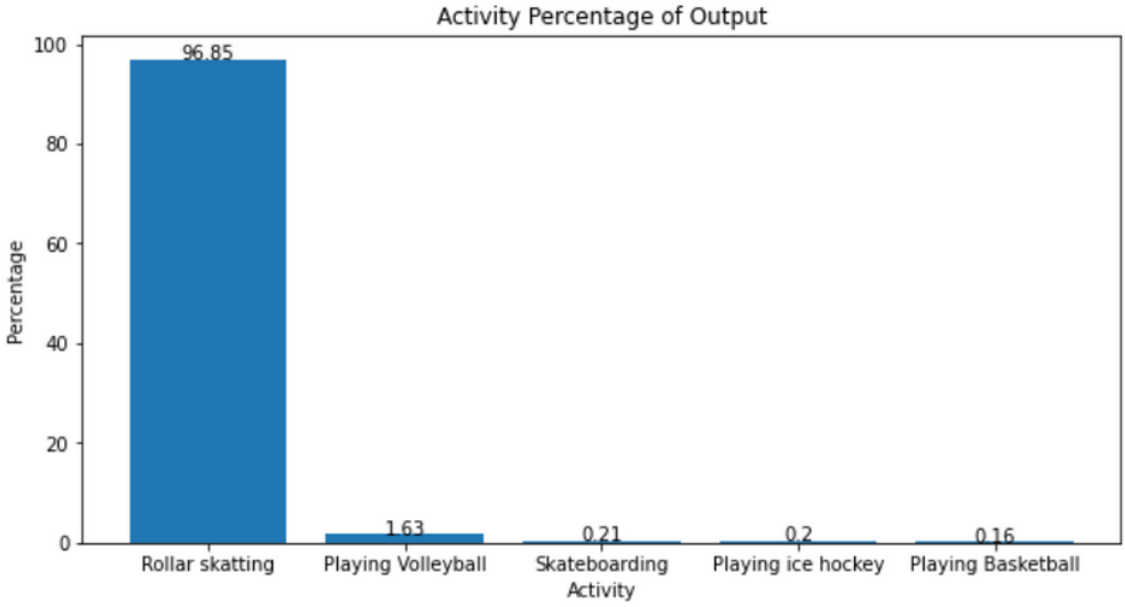


Figure 4.19: Activity Percentage of Output

Here is showing the output. Where is showing roller skating percentage is higher to predict the activity.

Reference	Focus Method	Accuracy
Andrej[35]	CNN	65.5%
ShaHuan[30]	CNN + CNN-LSTM	96.9%
Zhu[27]	CNN	96.08%
k[25]	CNN+RNN	97.5%
Sukrit[24]	CNN	96.3%
Anrin's[13]	CNN-LSTM	80.43%
Gary[16]	Decision Tree + Classifier	64%

Figure 4.20: Performance Analysis Comparisons

4.3 Results

For better accuracy, we tried different models and data ratios. We trained and evaluated 2 different machine learning models: convolutional neural network ans Long Short Term Memory. We also experimented with different data ratios. Here is a quick overview of different models and different data ratios:

No.	Models	Splitting Ration	Accuracy
1	CNN	75:25	93.18%
2	CNN-LSTM	70:30	75.34%
3	CNN-LSTM	80:20	99.67%

Figure 4.21: Different model and the dataset splitting

4.4 Summary

The experimental results demonstrate the effectiveness of the proposed sequential model to find human activity. Further research in this area can help to

improve the performance of the model and its applicability in clinical practice.

Chapter 5

5 Conclusion and Future Works

Human action recognition (HAR) has gained significant attention recently as it can be adapted for a smart surveillance system in Multimedia.[44] However, HAR is a challenging task because of the variety of human actions in daily life. Various solutions based on computer vision (CV) have been proposed in the literature which did not prove to be successful due to large video sequences which need to be processed in surveillance systems. The problem exacerbates in the presence of multi-view cameras. Recently, the development of deep learning (DL)-based systems has shown significant success for HAR even for multi-view camera systems.[45] In this research work, a DL-based design is proposed for HAR. There are many challenges involved in human action recognition in videos, such as cluttered backgrounds, occlusions, viewpoint variation, execution rate, and camera motion. To show the efficacy of the proposed design, we used two datasets, such as UCF101 and Accelerometer Data Set. This paper presents a review of various state-of-the-art deep learning-based techniques proposed for human action recognition on the two types of datasets. In light of the growing popularity and the recent developments in video-based human action recognition, this review imparts details of current trends and potential directions for future work to assist researchers.

Appendix

6 Dataset

- UFC-101 Dataset

The UFC 101 dataset is a widely-used benchmark dataset for action recognition in videos, specifically in the context of mixed martial arts (MMA) fights. The dataset consists of 101 short video clips of MMA fights, each approximately 6-8 seconds long. The videos were collected from UFC events and feature 13 different fighters performing various moves, such as punches, kicks, and takedowns. Each video clip is labeled with one of 101 action classes, such as "punch", "kick", "takedown", "submission", and so on. The videos are captured at a frame rate of 30 frames per second, and have a resolution of 320x240 pixels. They are encoded in the MPEG-4 format and are stored as individual video files. In addition to the video data, the dataset also includes precomputed optical flow vectors, which describe the motion of pixels between consecutive frames.

Figure 6.22: UFC-101 Dataset of Basketball

Figure 6.23: UFC-101 Dataset of Biking

Figure 6.24: UFC-101 Dataset of Bowling

Figure 6.25: UFC-101 Dataset of Diving

- Accelerometer Data Set

Accelerometer datasets are commonly used in research to analyze human activity and motion. These datasets typically consist of measurements taken from accelerometers, which are sensors that detect changes in velocity or acceleration. Here are some key points that could be included in a summary of accelerometer datasets for a research paper:

Data collection: Accelerometer datasets are typically collected using wearable devices that contain one or more accelerometers. These devices can be worn on different parts of the body, such as the wrist, ankle, or waist, depending on the specific research question.

Sampling frequency: The sampling frequency of an accelerometer dataset refers to the rate at which data is collected. Common sampling frequencies for accelerometer datasets range from 20-100 Hz, although higher frequencies may be used in certain applications.

Data preprocessing: Before analysis, accelerometer datasets may require preprocessing to filter out noise or correct for sensor drift. Common preprocessing steps include low-pass filtering, high-pass filtering, and calibration. Activity recognition: One common application of accelerometer datasets is to recog-

nize different types of physical activity based on the sensor readings. Machine learning algorithms can be trained on labeled datasets to classify activities such as walking, running, and cycling.

Health monitoring: Accelerometer datasets can also be used to monitor various aspects of health, such as sleep quality, heart rate variability, and energy expenditure. By analyzing the patterns of activity detected by accelerometers, researchers can gain insights into a person's overall health and well-being.

Challenges: Analyzing accelerometer datasets can present several challenges, such as dealing with missing or noisy data, selecting appropriate feature sets for machine learning models, and ensuring that the results are generalizable to different populations.

In conclusion, accelerometer datasets are a valuable resource for researchers studying human activity and motion. By using machine learning algorithms to analyze these datasets, researchers can gain insights into a wide range of health-related questions, from activity recognition to sleep monitoring. However, analyzing accelerometer datasets can also present several challenges, which must be carefully addressed to ensure accurate and reliable results.

```
33,Jogging,49105962326000,-0.6946377,12.680544,0.50395286;
33,Jogging,49106062271000,5.012288,11.264028,0.95342433;
33,Jogging,49106112167000,4.903325,10.882658,-0.08172209;
33,Jogging,49106222305000,-0.61291564,18.496431,3.0237172;
33,Jogging,49106332290000,-1.1849703,12.108489,7.205164;
33,Jogging,49106442306000,1.3756552,-2.4925237,-6.510526;
33,Jogging,49106542312000,-0.61291564,10.56939,5.706926;
33,Jogging,49106652389000,-0.50395286,13.947236,7.0553403;
33,Jogging,49106762313000,-8.430995,11.413852,5.134871;
33,Jogging,49106872299000,0.95342433,1.3756552,1.6480621;
33,Jogging,49106982315000,-8.19945,19.57244,2.7240696;
33,Jogging,49107092330000,1.4165162,5.7886477,2.982856;
33,Jogging,49107202316000,-1.879608,-2.982856,-0.29964766;
33,Jogging,49107312332000,-6.1291566,6.851035,-8.158588;
33,Jogging,49107422348000,5.829509,18.0061,8.539958;
33,Jogging,49107522293000,6.2789803,2.982856,2.9147544;
33,Jogging,49107632339000,-1.56634,8.308413,-1.4573772;
33,Jogging,49107742355000,3.5276701,13.593107,9.425281;
33,Jogging,49107852340000,-2.0294318,-5.706926,-10.18802;
33,Jogging,49107962326000,2.7649305,10.337844,-9.724928;
33,Jogging,49108062271000,3.568531,13.6748295,1.5390993;
```

Figure 6.26: Accelerometer Data Set

```
27,Walking,10704872296000,4.02,10.5,1.607201;
27,Walking,10704922223000,3.6,0.61,-2.0294318;
27,Walking,10704972272000,8.73,8.08,-4.099725;
27,Walking,10705022229000,5.37,8.96,-3.336985;
27,Walking,10705072309000,3.53,7.25,-0.9942854;
27,Walking,10705122205000,2.41,10.46,0.14982383;
27,Walking,10705172284000,1.04,9.7,0.6946377;
27,Walking,10705222455000,0.31,12.41,4.630918;
27,Walking,10705272260000,1.31,13.14,-2.070293;
27,Walking,10705322217000,-2.34,14.33,-2.7240696;
27,Walking,10705372266000,-3.11,13.63,2.3699405;
27,Walking,10705422284000,1.31,12.83,1.56634;
27,Walking,10705472303000,9.43,18.92,-2.982856;
27,Walking,10705522229000,12.49,13.29,1.4573772;
27,Walking,10705572278000,1.95,5.05,-3.255263;
27,Walking,10705622205000,0.5,4.67,-1.6480621;
27,Walking,10705672284000,3.34,2.18,-3.1463003;
27,Walking,10705722241000,5.13,1.27,-3.7864566;
27,Walking,10705772229000,3.76,3.49,-1.334794;
27,Walking,10705822248000,0.53,19.57,5.8567495;
27,Walking,10705872296000,1.61,12.57,3.405087;
```

Figure 6.27: Accelerometer Data Set

```
18,Upstairs,33635082263000,1.38,12.79,-1.3756552;
18,Upstairs,33635132281000,0.95,10.95,-1.0760075;
18,Upstairs,33635182238000,1.18,9.3,-0.0;
18,Upstairs,33635232287000,3.6,9.38,-1.0760075;
18,Upstairs,33635283160000,4.4,7.86,-0.3405087;
18,Upstairs,33635332263000,4.18,14.52,-0.95342433;
18,Upstairs,33635382281000,2.56,16.82,-1.607201;
18,Upstairs,33635432269000,2.79,9.38,0.6537767;
18,Upstairs,33635482318000,3.99,5.6,7.205164;
18,Upstairs,33635532305000,-1.18,9.15,-1.7978859;
18,Upstairs,33635582293000,-1.14,7.82,-3.9771416;
18,Upstairs,33635632281000,1.76,7.01,-2.5606253;
18,Upstairs,33635682269000,9.28,9.92,-2.4925237;
18,Upstairs,33635732287000,10.95,12.07,-1.7978859;
18,Upstairs,33635782244000,3.38,10.61,0.9942854;
18,Upstairs,33635832293000,0,7.93,0.46309182;
18,Upstairs,33635882281000,1.18,14.67,-0.10896278;
18,Upstairs,33635932269000,-2.64,13.33,-2.982856;
18,Upstairs,33635982257000,-1.53,7.82,-2.6014864;
18,Upstairs,33636032275000,-0.76,11.11,-6.851035;
18,Upstairs,33636082263000,2.96,9.43,-3.336985;
18,Upstairs,33636132312000,8.08,13.14,-3.3642259;
18,Upstairs,33636182299000,8.24,9.89,6.238119;
```

Figure 6.28: Accelerometer Data Set

This dataset contains time-series data collected from a smartphone accelerometer while the user performs a variety of physical activities, such as walking, running, and sitting. The dataset includes data from 30 participants, with each participant performing six different activities (walking, walking upstairs, walking downstairs, sitting, standing, and laying) while wearing a smartphone on their waist. The accelerometer sensor on the smartphone captures three-dimensional acceleration data at a sampling rate of 50 Hz, resulting in a time series of acceleration values for each axis (X, Y, and Z) for each activity. The data is split into two sets: a training set and a test set. The training set contains 7,352 samples, and the test set contains 2,947 samples. The data is labeled with the corresponding activity for each sample, allowing researchers to train and evaluate algorithms for activity recognition and classification.

References

[1] Blay Whitby. *Artificial Intelligence: A Beginner's Guide*. Oneworld Beginners' Guides Ser. Oct 2003.

[2] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. Adaptive Computation and Machine Learning Ser. Nov 2016.

[3] Roshan Singh, Alok Kumar Singh Kushwaha, and Rajeev Srivastava. Multi-view recognition system for human activity based on multiple features for video surveillance system. *Multimedia Tools and Applications*, 78(12):17165–17196, Jan 2019.

[4] Torki Altameem and Ayman Altameem. Facial expression recognition using human machine interaction and multi-modal visualization analysis for healthcare applications. *Image and Vision Computing*, 103:104044, Nov 2020.

[5] Ding-Xuan Zhou. Theory of deep convolutional neural networks: Downsampling. *Neural Networks*, 124:319–327, Apr 2020.

[6] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D Abowd, Nicholas D Lane, and Thomas Ploetz. Imutube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3):1–29, 2020.

[7] Cheng Xu, Duo Chai, Jie He, Xiaotong Zhang, and Shihong Duan. Innohar: A deep neural network for complex human activity recognition. *Ieee Access*, 7:9893–9902, 2019.

[8] Adrián Sánchez-Caballero, David Fuentes-Jiménez, and Cristina Losada-Gutiérrez. Real-time human action recognition using raw depth video-based recurrent neural networks. *Multimedia Tools and Applications*, 82(11):16213–16235, 2023.

[9] Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. Vision-based human action recognition: An overview and real world challenges. *Forensic Science International: Digital Investigation*, 32:200901, 2020.

[10] Susana Benavidez and Derek McCreight. A deep learning approach for human activity recognition project category: Other (time-series classification), 2019.

[11] Gary M Weiss. Wisdm smartphone and smartwatch activity and biometrics dataset. *UCI Machine Learning Repository: WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set*, 7:133190–133202, 2019.

[12] Christopher Reining, Friedrich Niemann, Fernando Moya Rueda, Gernot A Fink, and Michael ten Hompel. Human activity recognition for production and logistics—a systematic literature review. *Information*, 10(8):245, 2019.

[13] Djamila Romaissa Beddiar, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. Vision-based human activity recognition: a survey. *Multimedia Tools and Applications*, 79:30509–30555, 2020.

[14] Fan Zhu, Ling Shao, Jin Xie, and Yi Fang. From handcrafted to learned representations for human action recognition: A survey. *Image and Vision Computing*, 55:42–52, 2016.

[15] Tushar Dobhal, Vivswan Shitole, Gabriel Thomas, and Girisha Navada. Human activity recognition using binary motion image and deep learning. *Procedia computer science*, 58:178–185, 2015.

[16] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Sequential deep learning for human action recognition. In *Human Behavior Understanding: Second International Workshop,*

HBU 2011, Amsterdam, The Netherlands, November 16, 2011. Proceedings 2, pages 29–39. Springer, 2011.

[17] Pichao Wang, Wanqing Li, Zhimin Gao, Jing Zhang, Chang Tang, and Philip O Ogunbona. Action recognition from depth maps using deep convolutional neural networks. *IEEE Transactions on Human-Machine Systems*, 46(4):498–509, 2015.

[18] Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 461–470, 2015.

[19] Pasquale Foggia, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. Exploiting the deep learning paradigm for recognizing human actions. In *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 93–98. IEEE, 2014.

[20] Junyong You and Jari Korhonen. Attention boosted deep networks for video classification. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1761–1765, 2020.

[21] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[22] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[23] Sukrit Bhattacharya, Vaibhav Shaw, Pawan Kumar Singh, Ram Sarkar, and Debotosh Bhattacharjee. Sv-net: A deep learning approach to video based human activity recognition. In *Proceedings of the 11th International*

Conference on Soft Computing and Pattern Recognition (SoCPaR 2019) 11, pages 10–20. Springer, 2021.

[24] K Vijayaprabakaran, K Sathiyamurthy, and M Ponniamma. Video-based human activity recognition for elderly using convolutional neural network. *International Journal of Security and Privacy in Pervasive Computing (IJSPPC)*, 12(1):36–48, 2020.

[25] Mehrez Abdellaoui and Ali Douik. Human action recognition in video sequences using deep belief networks. *Traitement du Signal*, 37(1), 2020.

[26] Roberta Vrskova, Patrik Kamencay, Robert Hudec, and Peter Sykora. A new deep-learning method for human activity recognition. *Sensors*, 23(5):2816, 2023.

[27] Shanying Zhu, Wei Chen, Fulong Liu, Xiaotao Zhang, Xiupeng Han, et al. Human activity recognition based on a modified capsule network. *Mobile Information Systems*, 2023, 2023.

[28] Yi Fei Tan, Soon Chang Poh, Chee Pun Ooi, and Wooi Haw Tan. Human activity recognition with self-attention. *International Journal of Electrical and Computer Engineering (IJECE)*, 13(2), 2023.

[29] Diego Teran-Pineda, Karl Thurnhofer-Hemsi, and Enrique Dominguez. Analysis and recognition of human gait activity based on multimodal sensors. *Mathematics*, 11(6):1538, 2023.

[30] Sha Huan, Limei Wu, Man Zhang, Zhaoyue Wang, and Chao Yang. Radar human activity recognition with an attention-based deep learning network. *Sensors*, 23(6):3185, 2023.

[31] Abhisek Ray, Maheshkumar H Kolekar, R Balasubramanian, and Adel Hafiane. Transfer learning enhanced vision-based human activity recognition: A decade-long analysis. *International Journal of Information Management Data Insights*, 3(1):100142, 2023.

[32] Qiang Shen, Haotian Feng, Rui Song, Donglei Song, and Hao Xu. Federated meta-learning with attention for diversity-aware human activity recognition. *Sensors*, 23(3):1083, 2023.

[33] Salith Ziyan, MR Manu, et al. Human activity recognition using machine learning. *International Journal of Research in Engineering, Science and Management*, 4(7):253–255, 2021.

[34] Chunhui Ding, Shouke Fan, Ming Zhu, Weiguo Feng, and Baozhi Jia. Violence detection in video by using 3d convolutional neural networks. In *Advances in Visual Computing: 10th International Symposium, ISVC 2014, Las Vegas, NV, USA, December 8-10, 2014, Proceedings, Part II 10*, pages 551–558. Springer, 2014.

[35] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[36] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[37] Dominic M Sivitilli, Joshua R Smith, and David H Gire. Lessons for robotics from the control architecture of the octopus. *Frontiers in Robotics and AI*, 9, 2022.

[38] Liliana Lo Presti and Marco La Cascia. 3d skeleton-based human action classification: A survey. *Pattern Recognition*, 53:130–147, 2016.

[39] Leila Hedayatrad, Tom Stewart, and Scott Duncan. Concurrent validity of actigraph gt3x+ and axivity ax3 accelerometers for estimating physical activity and sedentary behavior. *Journal for the Measurement of Physical Behaviour*, 4(1):1–8, Mar 2021.

[40] Joohee Kim and Il Im. Anthropomorphic response: Understanding interactions between humans and artificial intelligence agents. *Computers in Human Behavior*, 139:107512, Feb 2023.

[41] Josh Neudorf, Shaylyn Kress, and Ron Borowsky. Structure can predict function in the human brain: a graph neural network deep learning model of functional connectivity and centrality based on structural connectivity. *Brain Structure and Function*, 227(1):331–343, Oct 2021.

[42] Chong Sun Hong. Confusion plot for the confusion matrix. *Journal of the Korean Data And Information Science Society*, 32(2):427–437, Mar 2021.

[43] Khan Md. Hasib, Farhana Rahman, Rashik Hasnat, and Md. Golam Rabiul Alam. A machine learning and explainable ai approach for predicting secondary school student performance. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0399–0405, 2022.

[44] Md Moniruzzaman, Zhaozheng Yin, Zhihai He, Ruwen Qin, and Ming C Leu. Human action recognition by discriminative feature pooling and video segment attention model. *IEEE Transactions on Multimedia*, 24:689–701, 2022.

[45] Yue Guan, Qiang Wei, and Guoqing Chen. Deep learning based personalized recommendation with multi-view information integration. *Decision Support Systems*, 118:58–69, Mar 2019.