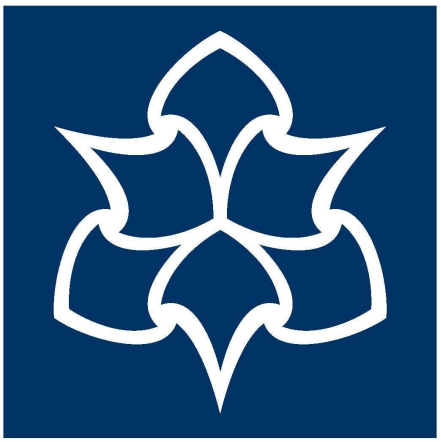
SMART DAILY HUMAN ACTIVITY REPORTING SYSTEM

A DISSERTATION SUBMITTED TO MANCHESTER METROPOLITAN UNIVERSITY

FOR THE DEGREE OF MASTER OF SCIENCE

IN THE FACULTY OF SCIENCE AND ENGINEERING



2023

By

Shreeya Gulawani

Department of Computing and Mathematics

Table of Contents

List of Tables

List of Figures

Abstract

This thesis focuses on the automatic identification of human actions in videos. The definition of human action recognition is the automatic identification of human actions in a video. Due to the numerous difficulties, including but not limited to changes in human shape and motion, occlusion, a crowded background, moving cameras, changing lighting conditions, and different viewpoints, this is a challenging problem.

The suggested action recognition framework, which is based on the literature study, is then based on machine learning and deep learning techniques. These approaches are used throughout the thesis by embedding unique action recognition algorithms in deep learning domains. Machine learning, deep learning, and real-world applications come together in the emerging topic of human activity recognition (HAR), which aims to give computer programs the ability to comprehend and interpret human activities and behaviours. It is crucial to be able to automatically identify and analyse human activities in the digital age, where wearable sensors and smart devices are pervasive. With the help of a diverse array of algorithms, including Convolutional Neural Networks (CNN), Long Short-Term Memory networks (LSTM), Support Vector Classifiers (SVC), Random Forest Classifiers, and Recurrent Neural Networks (RNN) with Gated Recurrent Units (GRU), this research project explores the complex HAR landscape. The main goal is to interpret the intricate structure of human behaviour while assuring accuracy and context awareness in recognition.

The most important parts of our adventure as we go across the algorithmic environment are algorithm selection and development. CNN+LSTM, which combines spatial and temporal analysis, has proven to be a versatile candidate that can identify subtle patterns in data. Conv2D, which is built on convolutional layers, performed particularly well in situations where spatial features were important. Although promising, the RNN+GRU design had issues, which highlighted the significance of algorithmic adaptability to particular recognition tasks. While Random Forest Classifier brought ensemble learning for increased stability, Support Vector Classifier proved robustness and accuracy. A complete set of evaluation metrics, including accuracy, precision, recall, F1-score, and confusion matrices, guided model training and validation and revealed the intricacies of recognition performance. We were able to determine the algorithmic strengths and limitations thanks to the comprehensive assessment, which also helped us choose the best recognition strategy for various tasks and environments.

Declaration

No part of this project has been submitted in support of an application for any other degree or qualification at this or any other institute of learning. Apart from those parts of the project containing citations to the work of others, this project is my own unaided work. This work has been carried out in accordance with the Manchester Metropolitan University research ethics procedures, and has received ethical approval number 57974.

Signed: Shreeya Gulawani
Date: 28/09/2023

Acknowledgement

I'd like to convey my heartfelt gratitude to the individuals and organisations who played critical roles in the successful completion of my MSc research. This endeavour was a huge academic milestone, and their steadfast support and advice were crucial in its success.

First and foremost, I want to express my heartfelt gratitude to my distinguished project supervisor, Dr.Li Guo. His knowledge, mentorship, and important insights have been the foundations of this research project. Their dedication to academic excellence and to my scholarly growth has substantially influenced the quality and direction of my work. I am eternally grateful for their counsel during this journey.

I would also like to thank Manchester Metropolitan University for offering an enriching academic environment and access to key resources, both of which were critical to the successful completion of my project. My scholastic development has been aided by the institution's dedication to academic achievement.

Finally, I am grateful to the pioneering scholars, authors, and researchers whose work has informed and inspired mine. Throughout the course of our research, the amount of knowledge within the academic community has been an excellent resource.

Abbreviations

ML	Machine Learning
DL	Deep Learning
CV	Open Computer Vision
IoT	Internet of Things
GPS	Global Positioning System
HAR	Human Activity Recognition
HCI	Human-Computer Interaction
CNNs	Convolutional Neural Networks
SVM	Support Vector Machines
UCF	University of Central Florida
SVM	Support Vector Machine
RNNs	Recurrent Neural Networks
ReLU	Rectifier Linear Unit
LSTM	LongShort-Term Memory
GRU	Gated Recurrent Unit
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
MSE	Mean Squared Error

Chapter 1 Introduction

1.1 Project Overview

Rapid advancements in the field of activity recognition and the proposal of numerous new models based on scientific and technological advancements have led to enormous progress in this area. Future research in this area now has more chances thanks to the advancement of deep learning and OpenCV with highly trained datasets. Such development could result in the genuine and beneficial implementation of such models in this technologically advanced world for the wellbeing of all living things. Numerous researchers and developers have applied these models in a variety of settings thanks to the utilisation of cutting-edge technology in this field.These highly trained models make it possible to monitor operations in real time in a very efficient and effective way. It is possible to deal with strange or suspicious behaviour using practical techniques that promote harmony and peace among living things.

The camera module that collects the raw data or collection of videos that are readily accessible as an input to the recognition system can be used to implement the human activity recognition model. After feature extraction, activity categorization is carried out by building several frames from the input data. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), or mixtures of both have been used in several publications utilising deep learning approaches for HAR problems based on or intended for portable platforms.
Human Action Recognition can be categorised as follows:

- Still-image based action recognition:

In the computer vision problem of still image-based action recognition, a computer software recognizes actions or activities in a single image. Unlike with videos, it doesn't necessitate the analysis of a series of images. The gathering of tagged picture datasets, feature extraction from images, action pattern learning using neural networks, and accuracy testing of the model are important phases. Managing different stance and viewpoint variations is a challenge. Applications include robotics, healthcare, human-computer interface, and surveillance.

- Action representation and analysis HAR:

Analysis-based action representation Identifying human behaviours or activities using data from sensors or movies is called "human activity recognition" (HAR). To do this, data must be gathered, actionable representations of actions such as skeletal positions or body movements, must be created, classification of these representations must be made using machine learning models, the models must be trained and validated, and finally the models' correctness must be assessed. Applications for HAR include security, human-computer interaction (HCI), and sports analysis. These fields require computers to comprehend and react to human behaviour.

- Abnormal activity recognition:

The process of identifying uncommon or outlier patterns in a dataset using data analysis tools, such as machine learning or statistical methods, is known as abnormal activity identification. This is frequently used in situations like security surveillance, healthcare monitoring, fraud detection, and quality control in manufacturing, when spotting anomalies or unexpected events is essential. The objective is to discern between abnormal behaviour and typical patterns in order to raise alarms or take action when anomalies are found. It entails gathering data, developing a model of typical behaviour, deciding what constitutes aberrant conduct, and then spotting departures from the model.

- Sensor-based HAR:

Utilising sensors like accelerometers, gyroscopes, or GPS to automatically recognize and categorise human actions is known as sensor-based human activity recognition (HAR). These sensors gather data, which is then used to represent human motion and train machine learning algorithms to distinguish various activities. Applications for which the system can initiate actions or warnings based on identified activities include security, fitness tracking, health monitoring, and more.

1.2 Project Problems

The intrinsic complexity and variety of human actions present a huge technical hurdle for HAR. Human conduct has many facets and differs greatly among people, cultures, and environments. Therefore, creating algorithms and models that can precisely recognize and comprehend this complex web of movements and gestures offers a significant challenge. Activities can differ slightly from one another, making it challenging to develop a solution that works for everyone. Furthermore, distinguishing between activities like walking and jogging that have comparable characteristics or movements calls for highly advanced algorithms that can capture minute differences. The requirement for reliable feature extraction and sensor fusion techniques creates another technical barrier. Data from various sensors, such as accelerometers, gyroscopes, and even cameras or microphones, is frequently used by HAR systems. A difficult issue is ensuring the seamless fusion of these various data sources and collecting pertinent aspects that capture the essence of human activity. The accuracy of activity recognition can be hampered by noise, drift, and sensor errors, which can further complicate data preparation. Additionally, striking the right balance between selecting the best combination of characteristics and preventing either over- or under-fitting is difficult.

Data imbalance problems, where some activities are noticeably more abundant in the dataset than others, are a common problem for HAR systems. This disparity may skew model training and result in subpar recognition of minority activities. To achieve fair and accurate recognition

across all activities, this issue must be addressed using strategies like oversampling, undersampling, or adding the proper weighting.

Concerns about privacy, consent, and possible technology abuse are at the centre of ethical issues in HAR research and implementation. People's privacy rights may be violated when sensor data is collected and analysed from them in public or private areas. A fundamental ethical challenge is finding a balance between the need to collect data for identification and the need to protect individual privacy. Furthermore, it is essential to seek consent that has been informed and to ensure that the use of the data will be transparent. Furthermore, there is a chance that HAR technology will be used for intrusive or surveillance activities, which raises questions about both widespread surveillance and individual privacy. For many HAR applications, notably in the fields of healthcare, safety, and human-computer interaction, real-time recognition is a crucial necessity. It is extremely difficult to achieve low-latency, real-time processing while keeping excellent accuracy. HAR systems must be effective enough to continually analyse data streams and produce accurate predictions. Real-time processing has high computing demands, particularly for deep learning-based models, which call for hardware acceleration and efficient methods.

1.3 Aims & Objectives

Human Activity Recognition (HAR) systems require diverse purposes and objectives that cover the fields of algorithm development, real-time processing, context-aware recognition, and user-centric design and evaluation. The main goal of this research project is to significantly advance our knowledge and skills in identifying and interpreting human actions and behaviours in order to make contributions to the field of HAR. Several specific goals have been created with this in mind.

The first step in conducting this study will be to conduct a complete literature review, which will include a full analysis of the body of knowledge already available on HAR systems. The review will cover a variety of topics, including data collection methods, feature extraction techniques, machine learning algorithms, and practical applications, laying the groundwork for further investigation. A crucial goal is to build and study superior machine learning algorithms for HAR based on the literature review. To improve these algorithms' effectiveness, efficiency, and resilience in identifying human activities, representative datasets will be used in their development and optimization. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs), two examples of deep learning architectures, will be examined to determine their applicability and possible benefits in the context of HAR. The created models are tested against the validation dataset after being trained on the training dataset. The models gain knowledge of human behaviours during training by studying the labelled data. Techniques like regularisation and hyperparameter adjustment can be used to avoid overfitting. The models' performance during training is monitored using the validation dataset, and any necessary corrections are made. Accuracy, precision, recall, F1-score, and confusion matrices are some of

the evaluation measures. The effectiveness of the models in accurately recognizing activities and reducing false positives and false negatives is quantified by these indicators.Determining the best-performing model based on the performance criteria and knowledge obtained from the comparison. depending on the research objectives and application requirements, taking into consideration elements like accuracy, computational efficiency, and scalability.
This dissertation intends to significantly advance the field of human activity recognition by pursuing these goals and objectives.

1.4 Tools & Timeline

Given the nature of the task, an effective combination of advanced techniques and technologies have been used to accelerate creativity and meet our research goals. It is able to easily access computing resources thanks to Google Colab, which serves as a cloud-based development environment and ensures the scalability needed for ML and DL operations. The programming language of choice is Python, which offers the flexibility and adaptability required to create complex ML and DL models thanks to its rich ecosystem of libraries and frameworks. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which are at the cutting edge of neural network technology, are used in these models,in order to decode complicated patterns in the data.

When used collectively, these technologies produce a synergy that enables users to solve the complex problems of human activity recognition with accuracy and efficiency, pushing the bounds of what is practical in this dynamic and developing field.

Task	Start Date	End Date	Duration
ToR & Ethics	09/07/2023	12/07/2023	3
Literature Review	09/07/2023	28/07/2023	19
Data Collection	09/07/2023	20/07/2023	11
Data Analysis	20/07/2023	02/08/2023	13
Model Development & Evaluation	01/08/2023	16/08/2023	15
Report Writing	15/08/2023	28/09/2023	44

Table. 1.1 Project Timeline

1.5 Report Structure

This report's content would be divided into several chapters. As noted below, each chapter will focus on a distinct stage of the project:

• Chapter 1 will address the fundamental outline of the project, as well as the potential challenges that may arise, as well as the main aims and objectives of the project.

• Chapter 2 will go over the various strategies and concepts available for the project's goal. There will also be an analysis of earlier comparable work undertaken by others.

• Chapter 3 will go through the approaches that will be employed during the project's experimentation phase.

• Chapter 4 will provide an analysis of the project's experimentation findings.

• Chapter 5 will present ideas that could not be verified in the current thesis but are important ideas to examine in the future.

• Chapter 6 will provide a conclusion to the entire report depending on whether the project's goals and objectives were met.

Chapter 2 Literature Review

2.1 Introduction

In the areas of computer vision and machine learning, human activity recognition (HAR) is a significant study topic. To automatically recognize the type of action being performed in the video is the HAR's main function. A lot of difficulties with HAR make this a particularly challenging topic. These difficulties include occlusion, varying human shape and motion, complex backgrounds, stationary or moving cameras, various lighting conditions, and viewpoint alterations. Nevertheless, the degree to which these difficulties are felt may differ depending on the type of activity being examined. Gestures, actions, interactions, and group activities can be divided into four general categories (Beddiar et al.).

- Gesture:

One of the most basic forms of non-verbal communication is gesture, in which people express ideas, intentions, or feelings through certain bodily gestures, postures, or activities. They can be both intentional and spontaneous and are very important in human contact.To recognize gestures in a HAR system, unique movement patterns connected to certain gestures are identified by collecting and analysing sensor data, such as accelerometer measurements. These patterns can be recognized and the corresponding gestures can be categorised by machine learning algorithms.
Examples: A thumbs-up gesture denotes agreement or approval. Pointing towards something draws attention to it or expresses interest in it.

- Action:

Activities include a broader variety of acts and behaviours that people engage in over time. Depending on the situation, these activities might be straightforward or complex and frequently have a purpose or goal. By evaluating data from numerous sensors, such as accelerometers, gyroscopes, or video cameras, HAR systems seek to identify and classify actions. Machine learning models can be trained to link particular activity patterns in sensor data with those patterns, enabling the system to guess what a person is doing at any given time.
Examples: The activity of cooking a meal, which requires a series of steps, including chopping ingredients, cooking, and serving.
Turning pages, paying attention to the text, and flipping the book are all parts of reading a book.

- Interaction:

The way people interact with one another or with the things around them is called an interaction. Interactions can be verbal, in which case spoken language is used to communicate, or non-verbal, in which case physical movements, body language, or gestures are used to communicate.The way people interact with one another or with the things around them is called an interaction.

Interactions can be verbal, in which case spoken language is used to communicate, or non-verbal, in which case physical movements, body language, or gestures are used to communicate.
Examples: Verbal Conversations take place when two people talk to one other, listen to each other, and reply to what they say.
Nonverbal communication may include crossed arms may suggest defensiveness, while a nod of the head may denote agreement.
Shaking hands to express a greeting or agreement is a common example of physical engagement in business settings.

- Group Activity:

Individuals working together or taking part in activities as a group constitute group activities. To accomplish a common objective or produce a group experience, these behaviours are frequently synchronised and coordinated.By examining data from numerous sensors and participant coordination patterns, HAR systems may identify group activities. For instance, the system may examine player posture and movement in team sports to determine the current activity.
Examples: In team sports like basketball or soccer, players cooperate to score goals or points by using synchronised movements and tactics.
Players participate in cooperative gameplay, which frequently involves bargaining and competitiveness, when playing board games like chess or Monopoly.

2.2 Background & Related Work

The history of Human Activity Recognition (HAR) is a comprehensive investigation of the relevance, development, and potential for transformation of the area. With broad implications for a range of industries, from healthcare and fitness tracking to smart environments and security systems, HAR has recently emerged as a vibrant and interdisciplinary research area at the intersection of computer science, signal processing, machine learning, and sensor technology. The earliest research on HAR was done in the late 20th century, with a focus on simple activity monitoring and gesture detection. On the basis of sensor data, researchers sought to create systems that could automatically evaluate human behaviour (Gupta et al.).
The emergence of wearable sensors such as accelerometers and gyroscopes was one of the innovations that opened up new opportunities for data collecting and processing. These sensors, which were frequently integrated into smartphones and smartwatches, enabled the monitoring and recognition of a greater range of actions and gestures. As a result, academics began to investigate the potential of HAR beyond interface design, imagining applications in healthcare, fitness, and other fields.

With the introduction of machine learning techniques, particularly supervised learning, which enabled the construction of data-driven models capable of distinguishing actions from sensor data, HAR's breakthrough moment arrived. Researchers used labelled datasets to train algorithms that could recognize a wide range of movements, from simple hand gestures to

complicated physical activities such as jogging or cycling. At the intersection of technical advancement and our understanding of human behaviour, Human Activity Recognition (HAR) offers the alluring possibility of machines that can perceive and interpret human actions. Despite the impressive advancements, the HAR area continues to face complex issues that both guide its research agenda and highlight its importance. The requirement for context awareness and adaptation is one significant problem. Human actions are essentially dependent on their environment, which is determined by elements including place, time, and social interactions. As a result, it is a challenging and emerging research area to design HAR systems that can dynamically alter their recognition algorithms in response to context. The capacity to switch between tasks with ease or to recognize complex sequences of movements, such as cooking a meal or dancing, is still a topic of ongoing research.

Data collection and processing are made more difficult because this level of adaptability demands not only advanced algorithms but also in-depth contextual information. Data collection is critical in HAR research since it is the lifeblood of system development and evaluation. To ensure the reliability and generalizability of their findings, researchers always attempt to improve data gathering procedures such as sensor selection, data collection protocols, and preprocessing methodologies. However, data imbalance issues, in which some activities are overrepresented in datasets, cause difficulties in model training and evaluation. Addressing this issue with strategies like oversampling, undersampling, or introducing suitable class weights is critical for providing equitable and accurate recognition across all activities while limiting biases inherent in imbalanced data.

In the HAR landscape, ethical concerns appear prominently. Sensor data gathering and processing, particularly in public or private locations, raises serious privacy problems. Balancing the necessity for data collection in order to promote research with the protection of individual privacy is a difficult ethical quandary. Consent processes and transparent data usage regulations are required to ensure that participants understand the goal of data gathering and its ramifications. There is also the possibility that HAR technology will be co-opted for intrusive or surveillance reasons, sparking arguments about mass surveillance, human autonomy, and civil liberties. Striking an ethical balance between innovation and privacy rights is a never-ending task that necessitates interdisciplinary collaboration, stringent rules, and society discourse. Another layer of complication is the lack of standardised benchmark datasets and evaluation metrics. The lack of widely acknowledged datasets makes it difficult to effectively compare the performance of different HAR systems. Researchers and organisations have made admirable efforts to produce and exchange standardised datasets, but the availability of diverse, complete, and representative datasets remains a difficulty. In addition to datasets, defining relevant evaluation criteria that can reflect the intricacies of activity recognition is a continuous research effort. Metrics must account for characteristics such as the temporal nature of activities, context-dependent fluctuations, and an unequal distribution of activity classes. The lack of defined standards stifles progress and makes judging the generalizability of new algorithms and models difficult.

Furthermore, the inclusiveness and accessibility of HAR technology pose a significant societal concern. While HAR has the potential to help people with impairments, monitor health conditions, and improve human-computer connection, there is a risk that certain groups will be excluded or disadvantaged if accessibility issues are not addressed adequately throughout

development. A multidimensional difficulty exists in achieving universal accessibility while retaining high recognition accuracy. Not only does it demand technological innovation, but also interdisciplinary collaboration, user-centred design, and a dedication to inclusive practices. A crucial ethical responsibility is to ensure that HAR technology serves all parts of society, regardless of age, aptitude, or socioeconomic status.

The use of HAR in the healthcare industry has shown to be particularly promising. Researchers and medical experts realised that HAR technology could be used to keep tabs on patients' movements, catch falls, and monitor their recovery. By enabling remote monitoring and early intervention, HAR provided the promise to enhance quality of life and lower healthcare expenditures for the elderly and others with chronic diseases. Additionally, HAR established itself in the world of sports and fitness. Wearables that could track their activity in real time, analyse their performance, and count steps were hailed by athletes and fitness aficionados. Thanks to the insights produced by HAR systems, individualised coaching and training programs were more widely available. The IoT and smart environments have seen great progress because of HAR. The presence of people could be detected, their activities could be understood, and the lighting, heating, and security systems in homes and businesses could be adjusted as necessary. It also promised to improve security and energy efficiency in addition to convenience.

Although they raise privacy issues, HAR security and surveillance applications have evolved. Public safety could be improved by using video cameras with HAR capabilities to spot suspicious activity in public areas. Finding the ideal balance between security and privacy, nevertheless, continues to be a major problem in this situation. The background of Human Activity Recognition, in a nutshell is an enthralling voyage through the nexus of technology and human behaviour. Since its modest origins in gesture recognition, HAR has grown into a vibrant, interdisciplinary field with limitless promise. It has had a profound impact on healthcare, fitness, smart surroundings, and security.

2.3 Key Concepts & Terminologies

A collection of key concepts guides the attempt to interpret human behaviours from sensor data in the complex field of human activity recognition (HAR).

- Temporal Context: A collection of key concepts underlies the complex field of Human Activity Recognition (HAR), which governs our effort to interpret human actions from sensor data. The idea of Temporal Context emerges as a crucial one, reflecting how time affects activity recognition in revolutionary ways. Accurate recognition depends on a comprehension of these temporal patterns.Temporal context refers to taking time-related information and patterns into account while interpreting and recognizing human actions or activities in the context of Human Activity Recognition (HAR) and data analysis in general. It emphasises that comprehending the context and importance of such acts depends critically on the time and sequencing of events, activities, or data points.

- Feature Extraction: Feature extraction in Human Activity Recognition (HAR) is undoubtedly the process of turning raw sensor data typically gathered from gadgets like accelerometers and gyroscopes into a collection of useful qualities or features. These characteristics capture important details about the observed human actions. The complexity of the data is reduced and made more manageable for machine learning algorithms by the extraction of these significant properties. This procedure is necessary because it reveals in the data the patterns and subtleties that are required for correctly differentiating between various activities. Feature extraction essentially serves as a link between the unprocessed sensor data and the algorithms in charge of identifying and categorising human behaviours, allowing for more accurate and effective activity recognition systems (Oukrich).

- Transfer Learning: Using the machine learning technique of transfer learning, a model that has been trained for one task can be modified or repurposed to carry out a related but distinct activity. In this method, the learning of new tasks is built upon a pre-trained model, frequently with a rich knowledge base. Transfer learning builds upon prior knowledge and refines it for better performance on the target task rather than starting from scratch (An et al.). When labelled data for the target job is few or expensive to collect, this strategy is especially beneficial. Transfer learning is essentially the use of pre-existing knowledge to speed up and improve learning for new, related tasks, making it a potent tool in machine learning and artificial intelligence.

- Data Adaptability: Dynamic Adaptation emphasises the value of HAR systems that change along with users. Over time, activities may vary, and HAR models must adjust. Dynamic adaptation refers to the system's capacity to pick up new information and modify its recognition abilities in response to people's changing behaviour.

- Privacy-Preserving: In the age of worries over data privacy, Privacy-Preserving Techniques are of the utmost importance. These methods enable accurate activity recognition while ensuring the privacy of sensitive user information. Applications that are responsible and secure can now be developed thanks to advances in privacy-preserving HAR.

- Human-Centric Evaluation: By including user feedback and arbitrary judgments, Human-Centric Evaluation raises the evaluation procedure. Real-world applications require an understanding of how effectively HAR correlates with user happiness and perception beyond quantitative indicators.

2.4 Human Activity Recognition

With the introduction of supervised learning techniques in particular, which enabled the creation of data-driven models capable of distinguishing actions from sensor data, HAR experienced its breakthrough. To train algorithms that could recognize a variety of movements,

from simple hand gestures to complicated physical activities like running or cycling, researchers used labelled datasets. The accuracy and adaptability of HAR systems were considerably improved by the switch from rule-based to data-driven techniques.

The introduction of deep learning was a critical turning point in the development of HAR. The HAR area was addressed using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which had already proven effective in computer vision and natural language processing, respectively. The hierarchical characteristics and sequential patterns in sensor data were particularly well-learned by these deep learning architectures, which made them suitable for accurately identifying complicated human actions. The field advanced due to the convergence of HAR and wearable technologies. Numerous sensors-equipped wearable gadgets have become a crucial part of millions of people's daily lives, generating valuable data sources for HAR research. For activity recognition, health monitoring, and customisation, smartwatches, fitness trackers, and even virtual reality headsets provide constant streams of sensor data. The IoT and smart environments have seen great progress because of HAR. The presence of people could be detected, their activities could be understood, and the lighting, heating, and security systems in homes and businesses could be adjusted as necessary. It also promised to improve security and energy efficiency in addition to convenience. Although they raise privacy issues, HAR security and surveillance applications have evolved. Public safety could be improved by using video cameras with HAR capabilities to spot suspicious activity in public areas. Finding the ideal balance between security and privacy, nevertheless, continues to be a major problem in this situation.

The creation of benchmark datasets, such as the Opportunity and UCI-HAR datasets, was essential to the advancement of HAR research. These datasets allowed researchers to objectively assess their models' performance, compare results, and track advancements in the area. In HAR research, standardised evaluation measures like accuracy, precision, recall, and F1-score have become standard, allowing for unbiased evaluations of recognition systems. Researchers are increasingly investigating multi-modal and multi-sensor techniques as HAR continues to develop. The possibility for more reliable and precise recognition of complicated behaviours is provided by combining data from many sensors, including accelerometers, gyroscopes, video cameras, and microphones. The use of fusion techniques to combine data from many sources is increasingly being studied. The use of HAR in the healthcare industry has shown to be particularly promising. Researchers and medical experts realised that HAR technology could be used to keep tabs on patients' movements, catch falls, and monitor their recovery. By enabling remote monitoring and early intervention, HAR provided the promise to enhance quality of life and lower healthcare expenditures for the elderly and others with chronic diseases.Additionally, HAR established itself in the world of sports and fitness. Wearables that could track their activity in real time, analyse their performance, and count steps were hailed by athletes and fitness aficionados. Thanks to the insights produced by HAR systems, individualised coaching and training programs were more widely available.

2.5 Human Activity Recognition Approaches

- HAR Using Image processing:

Human activity recognition (HAR) relies heavily on image processing, which has a multitude of options to improve the precision, level of detail, and context of activity recognition. Although HAR has historically relied on sensor data from devices like accelerometers and gyroscopes, the incorporation of image-based sensors like cameras has broadened the application potential of activity recognition. A more comprehensive understanding of human actions and their environmental surroundings is made possible by the rich visual information that photos and video streams offer. The depth of visual information is one of the primary drivers of image processing's significance in HAR. Images and videos provide a visual context that sensor data alone frequently struggles to completely capture. Complex interactions with objects, gestures, and the environment are a part of all real-world activities. For instance, there is a strong visual component to activities like cooking, playing an instrument, and executing challenging yoga positions. Analysing the visual clues offered by picture data is necessary to correctly identify these actions. Image processing becomes essential in these situations because it provides an extra layer of knowledge that helps with comprehension of these visually oriented activities (Anitha et al.).

Another fascinating area where image processing excels is fine-grained activity recognition. Some tasks necessitate a level of resolution and detail that sensor data by itself could find challenging to meet. For instance, the subtle information that images convey is necessary to recognize particular yoga positions, complex dance motions, or hand gestures for sign language. These fine-grained characteristics can be captured via image processing techniques, such as object detection and pose estimation, allowing for the highly detailed identification of activities. Image data is essential in the context of multi-modal fusion, a core idea in activity recognition. To increase the overall recognition accuracy, multi-modal fusion combines data from various sources or modalities, such as sensor data, audio data, and image data.

Each modality offers particular information, and the combination of various modalities can result in a more complete understanding of human behaviour. Images, with their rich visual content, are a crucial part of multi-modal fusion, enabling models to take advantage of both the strengths of sensor data and the strengths of visual data. Image processing in HAR has a huge positive impact on applications for privacy and security as well. Visual signals are used by surveillance and security systems to identify unlawful or suspect activity. Images and video streams make it possible to recognize intruders, follow their movements, and send out notifications when suspicious or potentially harmful activity is noticed. These systems can give a higher level of protection and situational awareness by incorporating image processing techniques.

Several applications, such fitness tracking and augmented reality, depend heavily on real-time feedback and monitoring. Real-time analysis of actions is made possible through image processing, providing prompt feedback and direction. For instance, while exercising, a system can measure a user's posture using camera data and offer real-time adjustments or exercise

suggestions. Similar to virtual reality, image processing in augmented reality apps can seamlessly incorporate virtual items or information into the user's surroundings, improving the overall user experience.

Image data can be a useful resource when augmenting training data. Large and varied datasets are advantageous for training deep learning models, which are becoming more and more common in HAR. When there is a lack of real sensor data, image data can be utilised to supplement training datasets. HAR's integration of image processing opens up a wide range of opportunities for developing more precise and context-aware recognition systems. By enhancing the data representation, it makes it possible to recognize fine-grained actions and activities with significant visual components. In a time when visual information is widely available, image processing enables HAR systems to achieve higher levels of precision and context awareness.

- HAR Using Machine & Deep Learning:

Human Activity Recognition (HAR) through Machine Learning is a captivating and swiftly evolving field that plays a pivotal role in various applications. It encompasses the development and implementation of algorithms and models to automatically detect and classify human activities based on data collected from sensors or other input sources. HAR has a broad range of practical applications, from healthcare and fitness tracking to security systems and beyond. In this comprehensive exploration, we will delve into the intricacies of HAR using Machine Learning, covering the entire process from data collection and preprocessing to model selection, training, validation, deployment, and ongoing improvement (Gupta et al.). The journey of HAR starts with data collection, where the groundwork for recognizing human activities is established. Typically, this involves using various sensors, such as accelerometers, gyroscopes, magnetometers, and sometimes even audio or video sources. These sensors work together to capture a wealth of information, including individuals' movements and the surrounding environmental conditions. The quality of the input data is crucial, as the reliability of the entire system depends on it.

After collecting data, it must go through a careful preprocessing phase. During this step, the raw data is subjected to a series of operations to ensure its cleanliness and relevance. Noise is eliminated, irrelevant information is filtered out, and the data is standardised. Techniques like data smoothing, feature extraction, and data scaling are applied to prepare the data for the following stages of the HAR pipeline. Feature extraction follows data preprocessing and is a critical step in the HAR process. It involves identifying and extracting relevant features from the preprocessed data. These features serve as the building blocks that enable the model to accurately distinguish and classify different human activities. These features can encompass a wide range of characteristics or patterns, including statistical measures like mean and standard deviation, frequency components, and more complex data representations. In supervised learning scenarios, where the model requires labelled data for learning, the dataset must be annotated or labelled accordingly. Human annotators play a crucial role in assigning activity labels to segments of data corresponding to specific activities. This labelling process bridges the gap between raw sensor data and the machine learning model's ability to recognize and classify human activities.

After selecting the model, the chosen model must undergo a rigorous training process. During this phase, the model is exposed to the labelled dataset, and it learns to recognize patterns and associations between the extracted features and the corresponding activity classes (Marszalek

et al.). The success of the training process depends on factors such as the size and quality of the training dataset, as well as the choice of hyperparameters and optimization techniques.
Validation and testing are integral components of the HAR development cycle. After the model has been trained, it is essential to assess its performance using separate datasets. The validation dataset is used to fine-tune hyperparameters and ensure the model's generalizability. Subsequently, the testing dataset is employed to evaluate the model's accuracy and effectiveness. Evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrices provide valuable insights into the model's performance, highlighting areas that may require further refinement.

Once the model has demonstrated satisfactory performance in testing, it is poised for deployment in real-world applications. This phase involves integrating the model into the target environment, whether it be a mobile application, wearable device, or embedded system. The model's seamless integration into these systems is paramount to ensuring its practical utility and usefulness. However, the journey of HAR does not end with deployment. For optimal and sustained performance, models often require ongoing improvement and adaptation. To achieve this, it is crucial to collect additional data from users and environments to refine and retrain the model. This continuous improvement process allows the model to adapt to changing user behaviours, evolving activity patterns, and shifting environmental conditions.

Deeper into the world of HAR, it becomes evident that it presents a plethora of challenges and complexities. One of the primary challenges is dealing with noisy data. Sensor data can be susceptible to various forms of noise, including interference and inaccuracies, which can adversely affect the model's performance. Rigorous preprocessing techniques, such as data smoothing and noise reduction, are necessary to address this issue. Another challenge in HAR is the handling of class imbalance. In many real-world scenarios, certain activities may be rare compared to others, leading to an imbalanced dataset. Models trained on imbalanced data may exhibit biases towards the majority class, resulting in reduced accuracy for minority classes. Techniques like oversampling, undersampling, and the use of appropriate evaluation metrics are employed to mitigate this challenge.

Furthermore, ensuring the model's generalizability across diverse user populations and environments is an ongoing challenge. The ability to accurately recognize activities for different users with distinct movement patterns and in various environmental conditions is a hallmark of a robust HAR system. This often requires the collection of diverse and representative data during both training and continuous improvement phases. In recent years, deep learning approaches have gained significant traction in the field of HAR. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have demonstrated remarkable capabilities in learning complex patterns in sequential data. CNNs excel in extracting spatial features from sensor data, while RNNs are adept at capturing temporal dependencies. However, deep learning approaches typically demand larger datasets and more computational resources for training and inference. Their adoption depends on the available resources and the specific requirements of the application.

In conclusion, Human Activity Recognition using Machine Learning is a dynamic and multifaceted field that continues to advance rapidly. It empowers a wide array of applications, from health monitoring and fitness tracking to security systems and beyond. The journey from data collection and preprocessing to model selection, training, validation, deployment, and

ongoing improvement involves a complex interplay of techniques, algorithms, and considerations. Addressing challenges such as noisy data, class imbalance, and generalizability is essential to building robust HAR systems. With the advent of deep learning, HAR has reached new heights, enabling the development of highly accurate and adaptable models. As technology continues to evolve, the potential applications of HAR are boundless, promising a future where machines understand human activities more deeply and accurately than ever before.

2.6 Human Activity Recognition Applications

Human Activity Recognition (HAR) is a revolutionary field with broad applications in numerous fields that uses technology to improve people's lives, increase safety, and simplify procedures.

- Human-Computer Interaction:

HAR in HCI facilitates intuitive and natural interactions between people and machines. Applications like gaming, virtual reality, and augmented reality use gesture recognition, a subset of HAR. Users can utilise hand and body motions to control video games, browse user interfaces, or interact with virtual items. Additionally, this technique has implications in the recognition of sign language, improving the usability of digital communication for those who have hearing loss.

- Healthcare:

HAR has transformed patient care and monitoring in the healthcare industry. Patients' movements and activities are tracked by wearable gadgets containing motion sensors, such as accelerometers and gyroscopes. This information assists in detecting falls in the elderly, assisting in accident prevention, and guaranteeing prompt medical care. In order for healthcare professionals to remotely monitor patients and evaluate their wellbeing, HAR is also essential for remote patient monitoring. Additionally, by monitoring patients' exercise routines and giving them immediate feedback, HAR supports individualised rehabilitation programs that aid patients in recovering from accidents or surgery.

- Sports & Fitness:

HAR is becoming a crucial component of the sports and fitness sectors. HAR is used by fitness trackers and smartwatches to track users' physical activity, tally steps taken, and calculate calories burned. Sports performance analysis uses HAR to track and analyse motions to improve technique and training, which benefits athletes. In order to improve the viewing experience for spectators, HAR is also used in sports broadcasting. It gives viewers real-time data on players' actions, such as sprinting pace, jump height, and heart rate.

- Smart Homes:

Smart homes and the Internet of Things (IoT) are developments made possible by HAR. In order to comprehend and adjust to occupants' preferences and behaviours, smart home systems can leverage HAR. Systems for lighting, heating, and cooling, for instance, can be automatically

modified based on identified activities and occupancy. This improves convenience while lowering energy use, improving the sustainability and energy efficiency of households.

- Security & Surveillance:

HAR assists in spotting and responding to abnormal or suspicious activity in security and surveillance. The use of HAR algorithms in video surveillance systems allows for the real-time detection of illicit entry, intruders, and strange behaviour patterns. The improvement of security in public areas, airports, and crucial infrastructure is dependent on this technology. Additionally, access control and identity verification using facial recognition and HAR can be utilised to ensure secure access to limited places.

- Transportation:

HAR supports safety and driver assistance systems in the automotive and transportation sectors. A driver's attention can be tracked using HAR algorithms, which can also spot indicators of distraction or tiredness. In order to increase car automation and safety, it is also employed in advanced driver assistance systems (ADAS), such as lane departure alerts and adaptive cruise control.

- Gaming:

Users' motions are tracked by motion-sensing controllers and virtual reality systems using HAR, enabling them to interact with characters and environments in virtual worlds. With the help of fitness applications and dance games, users are encouraged to stay active while having fun.

- Agriculture:

HAR helps with crop automation and monitoring in agriculture. By following the movement of plants and detecting problem areas, drones using HAR technology can examine crop health and growth patterns. HAR is used by automated harvesting equipment to distinguish between ripe and unripe fruits or vegetables, improving harvest efficiency and lowering waste.

Chapter 3 Experimental Methodology

3.1 Introduction

The Experimental Methodology, an essential component that determines the course of Human Activity Recognition (HAR). By using cutting-edge algorithms to uncover hidden patterns in data, the goal is to offer more insight into the complex world of human activities and behaviours. In order to get exact and context-aware recognition results, this methodology follows an approach that starts with data collecting and ends with thorough model evaluation.

Algorithm development and choice are at the core of experimental methodology. An eclectic group of algorithms, each with special advantages adapted to the peculiarities of HAR. Random Forest Classifier, Conv2D, RNN+GRU, SVC, and CNN+LSTM were all included in the ensemble. The precise qualities of the data and the recognition task drove the algorithm selection. To increase the accuracy of recognition, each algorithm underwent a thorough development process that included tweaking and optimization of hyperparameters.

A complete set of evaluation measures serves as the foundation for experimental methodology. These measurements, which included confusion matrices, F1-score, recall, and accuracy, presented an in-depth overview of recognition performance.

In addition to making accurate predictions, they were able to capture the intricacies of false positives and false negatives, which are critical in real-world applications where misinterpretation can have real-world repercussions. The comprehensive analysis made it possible to pinpoint algorithmic strengths and limitations, assisting in the decision of the best recognition strategy. The flowchart below depicts the flow of proposed methodology of this study:

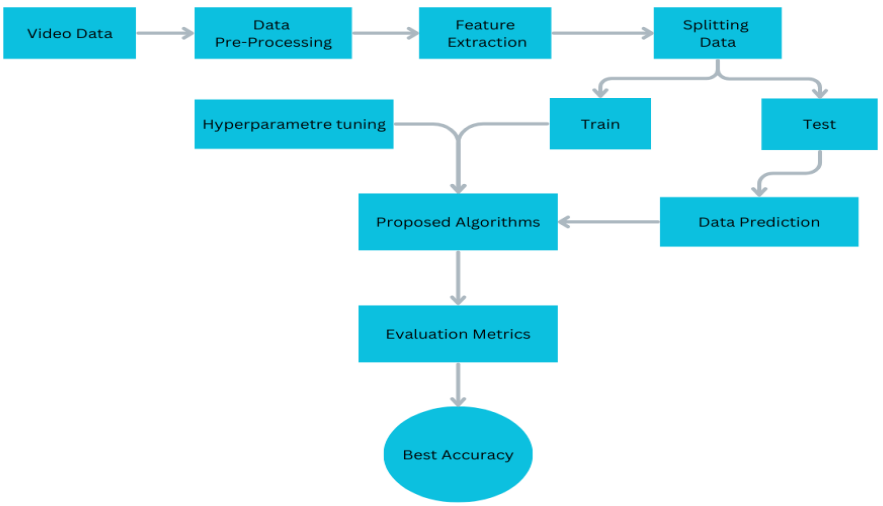


Fig. 3.1 System Workflow

3.2 Dataset Description

The following elements like data quality, relevance to the tasks to be performed and precisely labelled dataset was considered.Currently, the largest dataset of human actions is UCF101. Over 13k films, 101 action lessons, and 27 hours of video content are included. Realistic user-uploaded films with camera motion and a crowded background can be found in the database. On this new dataset, we also present baseline action recognition results using the conventional bag of words method, with an overall score of 43.9%. It is the most difficult dataset of actions at the moment because of the sheer volume of its classes, clips, and unrestricted nature of those clips. UCF101 is a dataset comprising 101 actions and 13320 videos, approximately twice the amount of actions and clips than the largest dataset that is accessible. Most action recognition datasets now in use have drawbacks like the amount of classes they offer and the videos captured unrealistically. The other datasets are put together with professional film crews' recordings of movie snippets. In recent times, unrestricted user-uploaded data has been incorporated in web videos (“Papers with Code - UCF101 Benchmark (Action Recognition)”).

UCF101 consists of 101 action classes in total, which we have classified into five categories: Human-Object Interaction, Body-Motion Only, Human-Human Interaction, Playing Musical Instruments, and Sports.UCF101 is an expansion of UCF50, which includes the following 50 action classes. The UCF50 contained the following classes: Baseball Pitch, Basketball Shooting, Bench Press, Biking, Billiards Shot, Breaststroke, Clean and Jerk, Diving, Drumming, Fencing, Golf Swing, High Jump, Horse Race, Horse Riding, Hula Hoop, Javelin Throw,, Juggling Balls, Jumping Jack, Jump Rope, Kayaking, Lunges, Military Parade, Mixing Batter, Nun chucks, Pizza Tossing, Playing Guitar, Playing Piano, Playing Tabla, Playing Violin, Pole Vault, Pommel Horse, Pull Ups, Punch, Push Ups, Rock Climbing Indoor, Rope Climbing, Rowing, Salsa Spins, Skate Boarding, Skiing, Skijet, Soccer Juggling, Swing, TaiChi, Tennis Swing, Throw Discus,Trampoline Jumping, Volleyball Spiking, Walking with a dog, Yo Yo. The following 51 new classes are introduced in UCF101: Apply Eye Makeup, Apply Lipstick, Archery, Baby Crawling, Balance Beam, Band Marching, Basketball Dunk, Blow Drying Hair, Blowing Candles, Body Weight Squats, Bowling,Boxing-Punching Bag, Boxing-Speed Bag, Brushing Teeth, Cliff Diving, Cricket Bowling, Cricket Shot, Cutting In Kitchen, Field Hockey Penalty, Floor Gymnastics, Frisbee Catch, Front Crawl, Hair cut, Hammering, Hammer Throw, Handstand Pushups, Handstand Walking, Head Massage, Ice Dancing, Knitting, Long Jump, Mopping Floor, Parallel Bars, Playing Cello, Playing Daf, Playing Dhol, Playing Flute, Playing Sitar, Rafting, Shaving Beard, Shot put, SkyDiving, Soccer Penalty, Still Rings, Sumo Wrestling, Surfing, Table Tennis Shot, Typing, Uneven Bars, Wall Pushups, Writing On Board. The bar chart below is from the official dataset paper and it depicts the distribution of videos over the action categories. The blue colour is for Human-Object Interaction, red for Body-Motion, purple for Human-Human Interaction, yellow for Playing Musical Instruments and green for sports.
Below given images are from the official paper of UCF-101 dataset which displays the total time and average clip duration over all the classes.

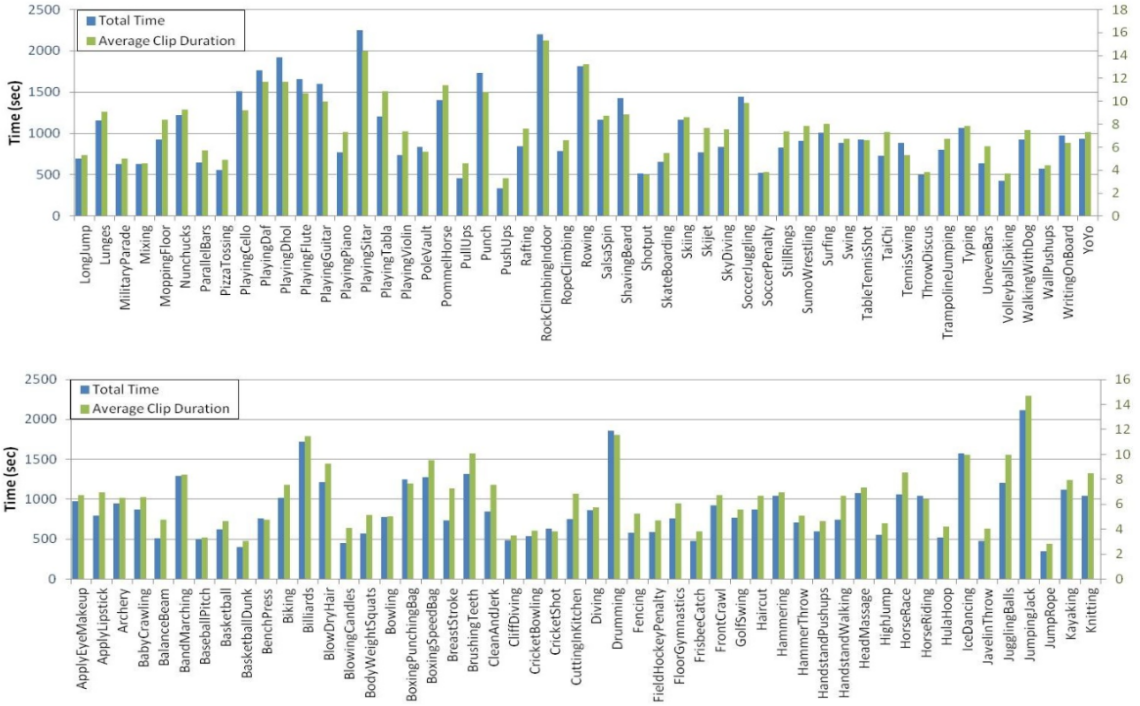


Fig. 3.2 Distribution of Clips over Classes

One action class's clips are split up into 25 groups, each of which has a different set of clips. The clips in one collection have a few features in common.

The videos come from YouTube and are downloaded. The frame rate and resolution of each clip are fixed at 25 FPS and 320 240, respectively '.avi' files are used to store the videos. The properties of the dataset are outlined in the table below:

Action	101
Clips	13320
Groups	25
Total Clip Length	1600 mins
Frame Rate	25 fps
Resolution	320*240

Table 3.1 Dataset Properties

Each clip's naming scheme corresponds to its format: v_**_g_**_ c**.The symbols * stand for the labels of the action class, the group, and the clips, respectively. v_HandstandWalking_g05_c07, for instance.Avi is the corresponding file for clip 07 of group 05 of the HandstandWalking action class.

To provide a baseline for results on UCF101, an experiment was conducted using the bag of words method, which is widely regarded as a standard action recognition technique. Actions in context, 2009. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) collected Harris 3D corners from each clip, and for each, 162 dimensional HOG/HOF descriptors were generated. Each video clip was represented by a 4000-dimensional histogram of its words, and the descriptors were allocated to their closest video words using a technique called nearest neighbour classifier. The histogram vectors from the training set were used to train an SVM utilising three train/test splits. The trained SVM was used to construct and classify a comparable histogram representation for the test video. The accuracy for the predefined action types is: Sports (49.40%), Playing Musical Instrument (42.04%), Human-Object Interaction (36.62%), Body-Motion Only (37.64%), and Human-Human Interaction (42.66%). This method produced an overall accuracy of 43.9%. Sports activities are the most accurate since they often demand unique motions, which facilitates classification. To maintain the consistency of the reported tests on UCF101, the train/test split was taken into consideration for the experimental configuration. A leaveone-group-out 25-fold cross validation configuration was also used in the aforementioned experiment, yielding an overall accuracy of 44.5%. It was ensured that the clips from a group were not split between the training and testing set by testing on one group and training on the others. UCF Sports, UCF11, UCF50, and UCF101 are other similar data sets. Several associated datasets that UCF has assembled are UCF Sports, UCF11, UCF50, and UCF101; each one includes the prior one.

With regard to HAR and computer vision, the UCF-101 dataset fulfils a number of vital functions.The UCF-101 dataset offers a standardised environment for assessing the efficacy of action recognition algorithms. This dataset can be used by researchers to evaluate how well various models and methods for identifying a variety of human behaviours can recognize various human actions.It is well suited for testing an algorithm's capability to manage real-world variability because it contains films shot in a variety of environments and scenarios. If action recognition systems are to be used in real-world settings, this is essential.

In UCF-101, algorithms will be tested to recognize a variety of diverse activities after learning from a tiny dataset. A robust model's applicability in real-world situations depends on this.The dataset is used by researchers and developers to create and test action recognition-related algorithms and tools for surveillance, HCI, and entertainment and sports video analysis.In educational settings, UCF-101 is frequently used to introduce computer vision, machine learning, and action recognition ideas to students. It offers a useful and interesting method of learning about various topics (Soomro et al.).

Despite its many benefits, the UCF-101 dataset also has several limits and issues that need to be addressed.Most of the dataset's behaviours are captured in brief video clips, which may not fully represent their duration in some cases. For actions that primarily rely on temporal information, this constraint can be problematic. There are concerns with class imbalance due to the variable number of video samples for each action category. Machine learning models can be trained and evaluated differently depending on which behaviours have more examples than others. Human behaviour is extremely variable, and some activities may show notable intra-class differences. This can make it challenging for models to correctly identify actions.Due to the variety of video sources, some movies may have distracting backdrops or clutter, which can make the work of recognizing objects more difficult. Researchers are continuously addressing

these issues and building bigger, more varied datasets to further enhance action recognition algorithms as the fields of computer vision and HAR continue to grow.

3.3 Data Pre-processing

Preparing raw data for meaningful analysis and machine learning is why data preprocessing is crucial. Algorithm performance can be impacted by the noise, consistency issues, or irrelevant information that frequently exists in raw data. Cleaning, normalisation, feature extraction, and label assignment are examples of data pre-processing procedures that assist in transforming data into a format that is appropriate for modelling. It ensures that the data is precise, consistent, and pertinent so that machine learning models may learn efficiently, make precise predictions, and derive insightful information. In essence, pre-processing the data is the crucial initial step in the data analysis pipeline, establishing the groundwork for effective model training and data-driven decision-making. Particularly in the context of video-based applications, data pre-processing is a crucial but frequently ignored part of any effective machine learning pipeline. Using machine learning algorithms for tasks like action recognition, video summarization, and anomaly detection requires a comprehensive understanding of the complexities of data pre-processing for video data, which we cover in-depth in this chapter. We investigate a Python code implementation that demonstrates how to transform unstructured video data into an instructive and structured format that supports machine learning.

The two crucial functions created in this step, feature_extraction and load_video, allow for the organising of video clips into a format that is appropriate for model training and the extraction of useful features from them. Videos frequently have different resolutions, therefore it's important to resize the frames to a standard size determined by the width and height parameters for consistency and effective processing. This process of resizing makes sure that every frame has the same proportions, removing any differences that would affect the model's capacity to recognize patterns. The preprocessed frames are collected into a list rather than being processed one at a time. It is impossible to create temporal sequences or video segments that adequately depict the dynamic nature of video data without the accumulation of frames per video clip. The sequence_length parameter, which can be changed based on the desired level of feature granularity, determines how many frames are included in each segment. Data preprocessing entails categorising the data as well as processing frames in order to support supervised learning. The videos being processed are given class labels by the load_video method. Machine learning models must be trained using class labels since they provide the baseline against which predictions can be tested and assessed.

In summary, there are a number of diligent stages that go into data preprocessing for video-based machine learning applications, each with a distinct purpose. All of these processes, from frame extraction to normalisation, organising, and labelling, prepare the raw video data for model training.

3.4 Feature Extraction

In machine learning, feature engineering is an essential component of data preprocessing. It entails choosing, modifying, and producing useful features from unprocessed data to improve the functionality of machine learning models. Effective feature engineering not only decreases dimensionality but also reveals pertinent patterns and relationships in the data, improving model accuracy and generalisation. It necessitates domain expertise and an in-depth comprehension of the data, enabling data scientists to identify the most instructive features and present them in a way that helps algorithms generate more precise predictions or classifications. The flowchart below depicts the overall flow of feature engineering followed in this experimental methodology.

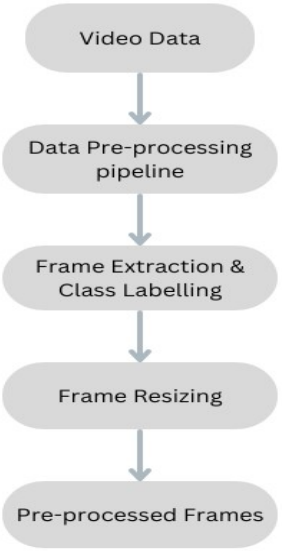


Fig. 3.3 Video Pre-processing

The pipeline for pre-processing video data relies heavily on the code component known as the feature_extraction function. Processing individual video footage and pulling out pertinent frames for further analysis is its main objective.Essential parameters like width, height, and sequence_length are first set by the feature_extraction function. These parameters specify the size of the frames and the quantity of frames that should be taken from each video clip. The performance of subsequent machine learning models may be strongly impacted by these decisions, which can also have an impact on the features that are produced. In order to store the video frames that have been extracted, the code creates an empty list called frames_list. The function reads the supplied video file and utilises the OpenCV library's cv2.VideoCapture method to access the video's frames. The function reads the video file supplied by the video_path parameter and utilises the OpenCV library's cv2.VideoCapture method to access the video's frames. The ability of this feature to handle videos of various lengths is one of its most important features. It determines a skip_interval based on the total number of frames in the video and the

desired sequence_length in order to accomplish this. How frequently video frames are sampled depends on this interval. A loop with a sequence_length iteration count extracts frames from the movie. The frame is first resized to the desired size. All frames are guaranteed to have the same size during this resizing process, which is crucial for further research. Each processed frame is then added to the frames_list.

Eventually, all of the frames required for the video clip under examination are added to this list. The video reader is released by calling the video_reader.release() function after each frame has been analysed and added to frames_list. This is an essential step in resource management because it makes system resources available for additional processing. The second function, load_video, manages the pre-processing pipeline for several movies, employs the feature_extraction function on each video, and ties the processed data to the appropriate class labels. Critical variables are initialised at the function's beginning. Images, labels, and label_index are defined. These variables act as storage spaces for processed video frames and class labels. The function calls the feature_extraction function and passes the video_path as an argument in order to extract useful features from the video. This sub-function generates a list of frames as its result after encapsulating the frame extraction, resizing, and normalisation procedures. The pictures list has a record of these frames. The designation of class labels for each video is yet another crucial component of this job.

A label_index variable is used to establish correspondence between films and their corresponding classes. Every time the outer loop iterates, it increases, making sure that each video in a class folder has the proper name. The labels list contains these labels. The function produces two essential pieces of information: photos and labels—after processing all movies in all class folders. Labels is a related numpy array holding class labels, whereas pictures are a numpy array containing processed frames from the films. These data formats are prepared for use in model training and are compatible with the majority of machine learning frameworks.

The entire data pre-processing cycle, beginning with classifying video data, feature extraction, and label assignment, is encapsulated in the load_video function.

For supervised machine learning applications, where the model learns patterns related to each class, this procedure is crucial.

3.5 Algorithm Classification

- ConVLSTM:

With the ability to mix Convolutional Layers (Conv) and Long Short-Term Memory (LSTM) Layers, the Convolutional Long Short-Term Memory (ConvLSTM) model is a potent neural network design that excels at processing spatiotemporal input. This approach is useful for jobs like video analysis, weather forecasting, autonomous navigation, and others where comprehending both spatial and temporal connections within the data is essential. To handle data arranged in grids or sequences with grid-like patterns, the ConvLSTM model fundamentally enhances the conventional LSTM by introducing convolutional processes. The primary component of ConvLSTM, convolutional layers are primarily in charge of analysing spatial data.

The learnable filters in these layers slide over the incoming data to extract regional patterns or features. These filters can simultaneously capture both spatial and temporal characteristics in the ConvLSTM context since they can operate on both a single time step and many time steps. Due to its special ability to recognize complex spatial patterns that change over time, the model is particularly well suited for tasks like video action recognition.

The Long Short-Term Memory (LSTM) layers are a mainstay of sequence modelling and are renowned for their capacity to identify distant dependencies in sequential data. Convolutional layers and LSTM cells combine to form a hybrid architecture in the ConvLSTM model, which is capable of learning and remembering patterns that change over both space and time. The capacity to preserve and update internal states of the LSTM makes it particularly skilled at handling sequences with complex temporal relationships, a capability essential for tasks like weather forecasting where prior data has a large impact on future forecasts. The ConvLSTM model is unique in that it can process data that has been structured into grids or sequences with grid-like patterns.

The ConvLSTM model's capacity to analyse data arranged in grids or sequences with grid-like structures is one of its distinctive characteristics. In fields where spatial linkages are important, such as image sequences or 2D spatiotemporal data, this grid-based processing is essential. The model applies convolutional and LSTM procedures to each cell individually while working on grid cells or pixels in parallel. ConvLSTM can discern spatial connections in the data thanks to this parallelism, which also enables it to record the change of surrounding cells over time. In applications like video analysis, where knowing the spatial arrangement of objects and their temporal interactions is essential for precise predictions, this grid-based technique is very beneficial (Xia et al.). Certainly, for a more full overview, let's explore further into the architecture and its components.

The model described is a complex deep learning architecture designed for the analysis of video data, particularly for tasks like action or video classification. Convolutional Long Short-Term Memory (ConvLSTM) units, which combine long short-term memory (LSTM) networks and convolutional neural networks (CNNs), are used in this model because they are excellent at detecting spatiotemporal patterns in video data (Yang et al.).

Input Layer:

-Batch Normalisation: The input data are normalised in this first layer, which improves their suitability for training. It speeds up training and enhances model convergence by keeping a running mean and variance of the data.

-ConvLSTM2D Layer: This layer serves as the model's foundation. It processes input sequences while maintaining temporal dependencies using 3x3 convolutional filters. Here, LeakyReLU is utilised as the activation function, enabling the model to recognize both positive and negative information in the data.

Pooling and Dropout Layers:

- MaxPooling3D: MaxPooling3D is used to reduce the spatial and temporal dimensions of the data after each ConvLSTM layer. In order to manage computational complexity and extract the most important characteristics from video frames, this is essential.

- Dropout layers are used to prevent overfitting in the TimeDistributed Dropout method. In order to force the model to acquire more robust and generalised characteristics, these layers randomly deactivate a portion of the network's neurons during training.

Batch Normalisation Layers:

-Each ConvLSTM layer is followed by batch normalisation. This method reduces internal covariate shift by normalising activations at each layer. This improves model convergence and stability.

Flattening Layer:

- The ConvLSTM layers produce a 3D output, which the Flatten layer converts into a 1D vector. Connecting the convolutional layers with fully linked layers requires this step..

Fully Connected Layers:

- Dense Layer (4096 units, ReLU activation): By acting as a feature extractor and building a high-dimensional representation of the data, this densely connected layer. It makes it possible for the model to recognize complex patterns in the video frames.

- Dense Layer (10 units, softmax activation): The output layer for classification tasks is the final layer. It generates class probabilities using the softmax activation function. Although the model in this instance is set up for 10 classes, it can be modified to solve unique categorization issues.

- CNN(Conv2D):

A convolutional neural network (CNN or ConvNet) is a class of deep neural networks used in deep learning that is typically used to recognize patterns in images but is also used for signal processing, computer vision, natural language processing, and other tasks. The structure of a convolutional network was influenced by the way the visual cortex is organised and mirrors the connectivity pattern of neurons in the human brain. The name of this particular form of artificial neural network comes from convolution, one of the network's most significant functions.A convolutional neural network's first layer is always a convolutional layer. Convolutional layers operate on the input using a convolution operation and send the outcome to the following layer. All the pixels in a convolution's receptive area are combined into a single value. For instance, if a convolution is applied on an image, it will both reduce the size of the image and combine all of the field data into a single pixel. A vector is the convolutional layer's final output. We can employ several types of convolutions depending on the problem type we need to solve and the features we want to learn (Lee and Ahn).

The 2D convolution layer, sometimes abbreviated as conv2D, is the most often used type of convolution. During an element wise multiplication, a filter or kernel in a conv2D layer slides over the 2D input data. Therefore, it will combine the outcomes into a single output pixel. For each point it slides over, the kernel will carry out the identical action, converting one 2D feature matrix into another 2D feature matrix.

Conv2D layers' advantages of translation invariance, automatic feature learning, and spatial hierarchies make it possible for HAR models to identify activities in various locations throughout

video frames. However, in order to create efficient HAR systems, issues including the necessity for a lot of labelled data, potential overfitting in complicated models, and architectural factors must be carefully taken into account.

Certainly, for a more full overview, let's explore further into the architecture and its components. The 3D data that is frequently used in applications like video analysis and 3D image recognition may be processed by this CNN. The breakdown of its parts is provided below:

Input Layer:

- Batch Normalisation This first layer batch normalises the input data by keeping a running mean and variance, which speeds up training and enhances model convergence.
- Conv2D Layer: This layer uses 3x3 filters to perform 2D convolution operations on the input data. In order to introduce non-linearity and collect pertinent spatial characteristics, it uses the LeakyReLU activation function.
- MaxPooling3D: MaxPooling3D is applied following each Conv2D layer. This process lowers spatial dimensions and aids in data downsampling.

Dropout Layers:

This dropout layer is used to reduce overfitting by randomly deactivating a portion of the network's neurons during training. It is applied after MaxPooling. The dropout likelihood is managed using the dropout rates (0.2 or 0.3).

Flatten Layer:

This layer prepares the 3D output from the preceding levels for fully connected layers by flattening it into a 1D vector.

Fully linked Layers:

- Dense Layer (4096 units, ReLU activation): By acting as a feature extractor and producing a high-dimensional representation of the data, this densely linked layer.
- Dense Layer (10 units, softmax activation): The output layer for classification tasks is the final layer. In order to generate class probabilities, it employs the softmax activation function.

- RNN+GRU

A deep learning model built for sequence data processing is a Recurrent Neural Network (RNN) with Gated Recurrent Unit (GRU) layers. Because traditional feedforward neural networks are built for fixed-size inputs, they are inadequate for sequential data with varying lengths. RNNs were developed to manage data sequences by inserting recurrent connections. These links allow information to be carried forward from earlier time steps. An RNN's basic principle is to keep a hidden state vector that is updated at each time step based on the current input and the prior hidden state. This hidden state encodes data about the series so far. Standard RNNs, on the other hand, have restrictions, such as the vanishing gradient problem, which can render them ineffective for extended sequences (Sun et al.).

More advanced recurrent layers, like GRUs, were introduced to solve some of the limitations of regular RNNs. GRUs are a type of gated recurrent layer that aids the model in

capturing and propagating information across longer sequences. The reset gate and the update gate are the two most important parts of a GRU. These gates govern how much information from the previous hidden state should be forgotten and how much new information should be stored. The reset gate determines which information from the previous concealed state should be reset, while the update gate determines which information should be updated with the new input. Because GRUs are more resistant to the vanishing gradient problem, they are a preferred alternative for many sequence-based applications. Multiple GRU units are stacked on top of each other to form an RNN with GRU layers. Each GRU layer processes the input sequence one time step at a time, updating its hidden state as needed. The final GRU layer's output can be used for a variety of tasks, including sequence prediction, classification, and sequence production.

Certainly, for a more full overview, let's explore further into the architecture and its components. The code creates a neural network model by utilising the Keras library, which is a high-level deep learning API built on top of TensorFlow.

A sequential model is developed, allowing layers to be stacked one on top of the other.

RNN Layer: The first layer consists of 64 SimpleRNN units. It accepts an input shape, implying that this model is intended to handle two-dimensional sequential data. This layer returns sequences, as indicated by the return_sequences option, making it suitable for stacking RNN layers. To avoid overfitting, a dropout layer with a dropout rate of 0.3 is inserted after the SimpleRNN layer.

GRU Layer : The following layer is a GRU (Gated Recurrent Unit) layer with 64 units and the'return_sequences=True' option. GRU is another form of recurrent layer that is well-known for its ability to capture sequential patterns. Following the GRU layer is another dropout layer with a dropout rate of 0.3.

Flatten Layer: Add a flatten layer after the recurrent layers. This layer is used to convert the previous layers' 3D output into a 1D vector that can be fed into a fully linked (dense) layer.

Dense Layers: Add two dense layers of 128 units each, as well as ‘relu' activation functions. In neural networks, these layers are frequently employed for feature extraction and non-linearity. A dropout layer is included to prevent overfitting.
Output Layer: The final layer is a dense layer with 10 units, assuming there are 10 classes in the classification issue and a 'softmax' activation function. Softmax is often used to generate probability distributions across classes in multi-class classification problems.

Overall, this model is intended for sequential data processing by combining SimpleRNN and GRU layers, followed by several fully connected layers for feature extraction and classification. By randomly discarding a fraction of the connections during training, the dropout layers serve to prevent overfitting.

- SVM Classifier

A support vector machine (SVM) is a sort of supervised learning algorithm used in machine learning to tackle classification and regression tasks. SVMs are particularly good at tackling binary classification issues, which require categorising data set members into two groups. A support vector machine algorithm's goal is to determine the best line, or decision boundary, that divides data points into different data classes. SVMs work by transforming input data into a higher-dimensional feature space. This modification makes it easy to discover a linear separation or classify the data set more effectively. SVMs can deal with both linearly and nonlinearly separable data. SVMs are less prone to overfitting than other methods, such as decision trees; overfitting occurs when a model performs exceptionally well on training data but becomes too particular to that data and is unable to generalise to new data. The implementation of the margin maximisation principle by SVMs aids in generalisation to previously unknown data.

Following training, the object containing the learnt model is used to generate predictions on new, previously unknown data. SVMs are particularly effective for binary classification tasks, but they can also be applied to multi-class situations.

Once the model has been trained, its performance can be evaluated by making predictions and comparing them to the ground truth labels using the testing data and related labels. This evaluation can be carried out using measures such as accuracy, precision, recall, and F1-score, which are provided via scikit-learn functions such as classification_report and accuracy_score.

- Random Forest Classifier

A random forest is a supervised machine learning technique that combines the calculations of multiple decision trees to obtain a single final result. It is well-liked since it is simple yet efficient. Classification and regression trees are two of the most prevalent types of this approach. Classification trees: This technique is used to discover which "class" a given variable belongs to. Regression Trees: A regression tree is a method in which the target variable is not fixed. A regression problem is the prediction of housing prices, which is a continuous variable.

A Random Forest classifier is a subset of the Random Forest method that is used to solve classification problems. It refers to the employment of a Random Forest ensemble to perform classification tasks, such as categorising data into multiple classes or categories, in this context. When a Random Forest is used to classify data, each decision tree inside the forest offers class predictions, and the final prediction is established by majority vote among the trees. The Random Forest classifier is designed primarily for classification issues, and it is one of the most common applications for Random Forests.

Random Forest calculates the relevance of features, which might be useful in HAR. Understanding which features (for example, sensor data) contribute the most to classification can assist researchers and practitioners in gaining insights into the underlying patterns of human activity. HAR datasets frequently feature uneven classes, with some actions occurring less frequently than others. Random Forest can handle class imbalance successfully by prioritising minority classes throughout training.

In the code, a Random Forest classifier is being created and trained. The Random Forest algorithm is an ensemble learning method that combines the predictions of multiple decision trees to improve accuracy and reduce overfitting.
The Random Forest classifier is trained using the training data and their labels. The Random Forest algorithm constructs a series of decision trees during training, each trained on a distinct subset of the training data with replacement. It also chooses a subset of features at random for each tree, which improves tree diversity. The goal is to generate a set of decision trees that are diverse and accurate. It uses ensemble learning to create a final prediction by combining the predictions of numerous separate decision trees. When predicting a new data point, each tree in the forest guesses the class independently, and the class with the most votes among the trees is chosen as the final forecast. When compared to a single decision tree, this ensemble technique often results in higher generalisation and lowers the danger of overfitting. Each tree in the forest predicts a class for each new data point, and the class with the highest votes becomes the predicted class for that data point.

3.6 Evaluation Metrics

The machine learning pipeline is not complete without evaluation metrics. They act as an indicator to assess how well a machine learning model performs and completes the task for which it was designed. It is crucial to choose the right evaluation metrics because they not only comprehend a model's capabilities but also assist to choose models, tune hyperparameters, and make overall decisions for machine learning projects (Lalwani).

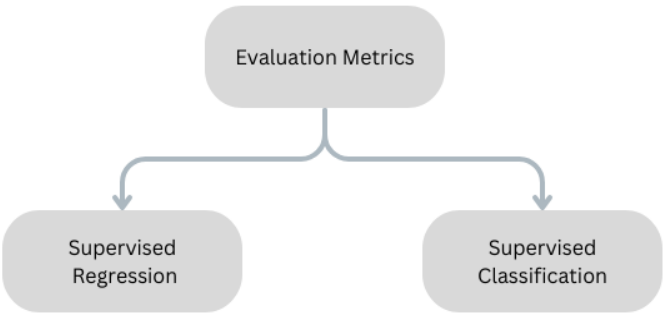


Fig. 3.4 Evaluation Metrics

- Evaluation Metrics for Classification:

Confusion matrix:

A confusion matrix, also known as an error matrix, is a specific table layout that allows visualisation of the performance of an algorithm in the field of machine learning, specifically the problem of statistical classification. In unsupervised learning, it is typically referred to as a matching matrix.

Both variations of the matrix, where each row represents examples in an actual class and each column represents instances in a predicted class, are documented in the literature.The name refers to the fact that it is simple to determine whether the system is mislabeling one class as another, or confusing two classes.This particular contingency table has two dimensions "actual" and "predicted," with identical sets of "classes" in each dimension. Each such combination of dimension and class is a contingency table variable(Lalwani). Key elements of confusion metrics are as follows:
True Positives (TP): Instances that are correctly classified as positive.
True Negatives (TN): Instances that are correctly classified as negative.
False Positives (FP): Instances that are incorrectly classified as positive (Type I error).
False Negatives (FN): Instances that are incorrectly classified as negative (Type II error).
It serves as the foundation for calculating various evaluation metrics, including accuracy, precision, recall, and F1-score.
Metrics Derived from the Confusion Matrix:
Accuracy: The proportion of correct predictions (TP and TN) out of the total.
Precision: The ratio of TP to the total positive predictions (TP + FP).
Recall (Sensitivity): The ratio of TP to the total actual positives (TP + FN).
Specificity (True Negative Rate): The ratio of TN to the total actual negatives (TN + FP).
F1-Score: The harmonic mean of precision and recall, providing a balance between the two metrics.

Accuracy:
One of the easiest criteria for categorization issues is accuracy.
It calculates what percentage of all instances were correctly categorised.
Accuracy may not be appropriate for unbalanced datasets when one class predominates, despite being simple to read.The percentage of correctly classified data instances over all data instances is known as accuracy. A worse assessment metric will result from unbalanced data because Accuracy will favour classes with more counts. We have two options: recall or precision.
Accuracy = (TP + TN) / (TP + FP + FN + TN)

Precision:
Precision measures the proportion of accurate positive predictions to all of the model's positive predictions. When the cost of false positives is significant, this statistic is crucial.
When a False Positive is more concerning than a False Negative, it is helpful.
As much precision as possible should be used.
Precision = TP/(TP + FP).

Recall:
Recall measures how well the model can detect each positive case.
It calculates the proportion of actual positives to all positives to be true. Additional names include sensitivity and true positive rate. When a False Negative versus a False Positive is a significant worry, it is more helpful. Recall ought to be as high as is feasible.
Recall= TP / (TP + FN)

F1 Score:

The arithmetic mean of precision and recall is known as the F1 score. It strikes a balance between the trade-offs of memory and precision. F1-score is especially helpful when there is an imbalance in the distribution of the classes or when it is important to reduce both false positives and false negatives.Its harmonic mean for recall and precision. The Precision and Recall components of the F1 Score are balanced; if either is low, the F1 Score will likewise be low.

F1 Score = (2 * Precision * Recall) / (Precision + Recall)

- Evaluation Metrics for Regression:

Regression tasks in machine learning entail predicting a continuous numeric value, therefore selecting the right evaluation metrics is crucial for effectively gauging model performance. The dataset's unique properties and the objectives of the regression task determine the evaluation metrics to be used(Lalwani).

Mean Absolute Error(MAE):

The average absolute difference between expected and actual values is measured by MAE.

Compared to other metrics like MSE, it is simple to grasp and less sensitive to outliers.

When errors must be expressed in the same units as the target variable, MAE is appropriate.

$MAE = (1/n) * \Sigma |y_i - \hat{y}_i|$

where n is the number of data points, y_i is the actual value, and $\hat{y}_i$ is the predicted value for the i-th data point.

Mean Squared Error(MSE):

The average squared difference between expected and actual values is determined by MSE. It is more sensitive to outliers since it penalises bigger mistakes more severely. MSE is frequently employed in regression issues.

$MSE = (1/n) * \Sigma (y_i - \hat{y}_i)^2$

where n is the number of data points, y_i is the actual value, and $\hat{y}_i$ is the predicted value for the i-th data point.

Root Mean Squared Error(RMSE):

The advantage of returning mistakes on the same scale as the target variable is provided by RMSE, which is the square root of MSE.Since it is in the data's original units, it is easier to read.

$RMSE = \sqrt{MSE}$

RMSE is simply the square root of the Mean Squared Error.

R-squared error:

R-squared, also known as the coefficient of determination, quantifies the percentage of the target variable's variance that can be predicted from the independent variables.

Higher numbers suggest a better fit, and the value ranges from 0 to 1.

R-squared is a tool for evaluating how effectively a model accounts for data variance.

$R^2 = 1 - [(\Sigma (y_i - \hat{y}_i)^2) / (\Sigma (y_i - \bar{y})^2)]$

where n is the number of data points, y_i is the actual value, ŷ_i is the predicted value for the i-th data point, and ȳ is the mean of the actual values.

Chapter 4 Results & Discussions

This chapter summarises efforts in the Human Activity Recognition (HAR) that have used a combination of Machine Learning (ML) and Deep Learning (DL) models. The results of studies have been revealed as a result of the process through data collecting, preprocessing, feature engineering, and model construction. Showcasing the performance measures achieved by models, such as accuracy, precision, recall, and F1-score. These metrics serve as quantifiable indicators of our models' ability to recognize and classify human behaviours. The table below displays evaluation metrics of the three regression algorithms.

	ConVLSTM	**CNN(ConV2D)**	**RNN+GRU**
Accuracy	0.93	0.94	0.69
Precision	0.92	0.93	0.67
Recall	0.87	0.92	0.66
F1-Score	0.88	0.92	0.65
MSE	1.19	0.34	6.65
RMSE	1.09	0.58	2.58
MAE	0.25	0.12	1.23

Table 4.1 Evaluation Metrics of Regression Models

- ConVLSTM

Over 15 epochs, the performance of the ConVLSTM model is measured. Each epoch corresponds to one full trip through the training dataset. The training loss begins at 2.5925 and declines throughout epochs, suggesting that the model's ability to anticipate human activities is improving. The accuracy measure displays the percentage of successfully identified samples. The training accuracy is 35.79% at the start , but it steadily improves with each epoch, reaching 95.04% by Epoch 15. This training process is also being evaluated on a validation set. These metrics aid in ensuring that the model does not overfit. Overall, the model appears to be performing effectively, with training and validation accuracy increasing and loss dropping.The graphs below represent Model Loss and Model Accuracy respectively

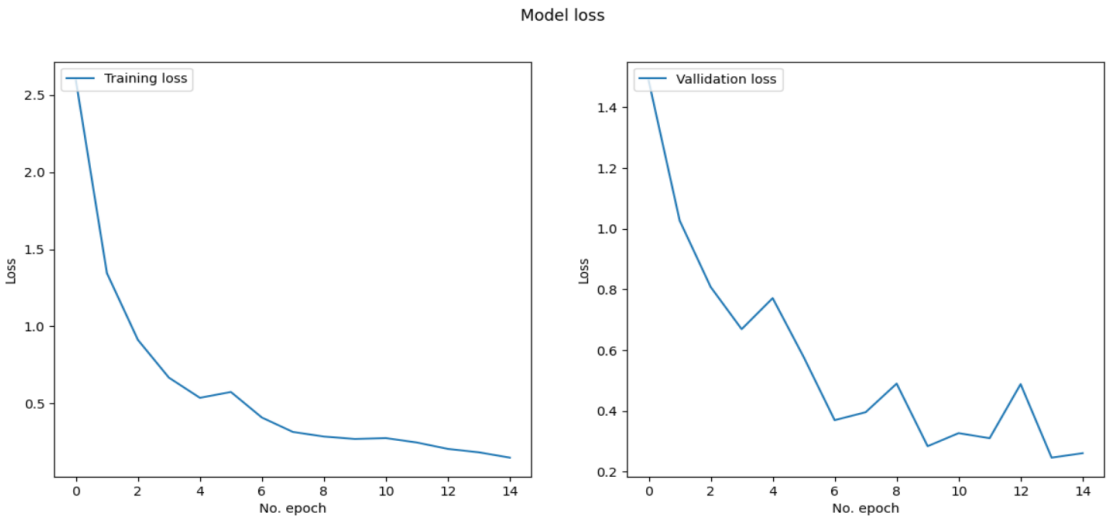


Fig. 4.1 ConVLSTM model loss

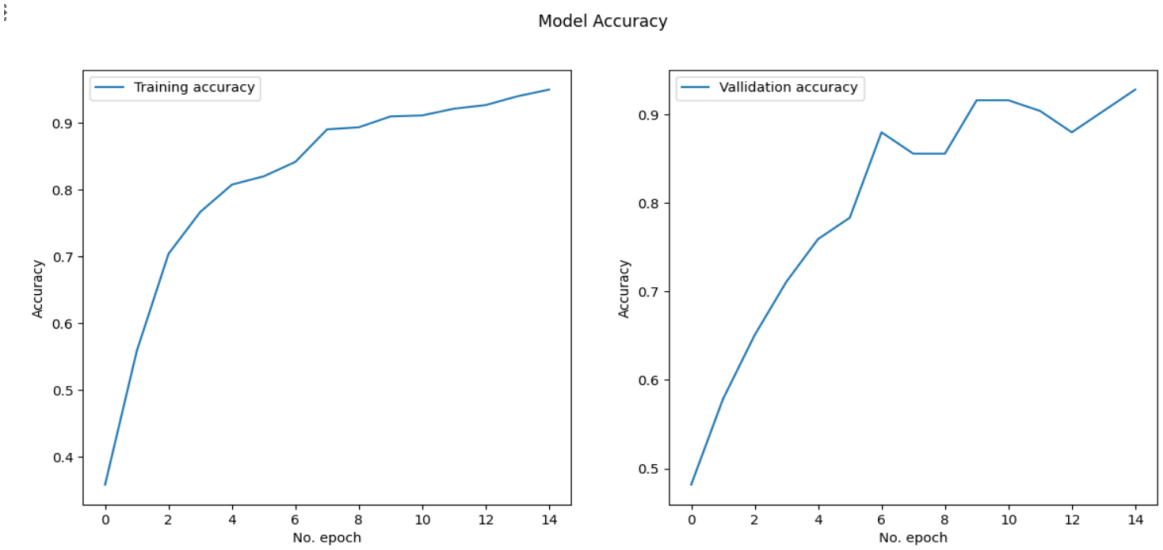


Fig. 4.2 ConVLSTM model accuracy

The evaluation results show how well the categorization model used in the Human Activity Recognition (HAR) system performed. The model has a high level of overall correctness in its predictions, with an accuracy of 93%. It has a high precision of 92%, indicating that when it predicts a specific human activity, it is correct 92% of the time. Furthermore, the model obtains an exceptional recall rate of 87%, suggesting that it correctly identifies 87% of all actual positive events. This mix of precision and recall is represented in the F1-Score, which is 0.88, demonstrating the model's ability to handle imbalanced datasets. The Mean Squared Error (MSE) of 1.19 quantifies the average squared difference between the model's predictions and actual data. Lower MSE values indicate more accuracy. At 1.09, the Root Mean Squared Error (RMSE) is the square root of MSE and gives a more interpretable measure of error in the target variable's original units. An RMSE of 1.09 indicates that the model's predictions differ from the actual values by around 1.09 units on average. Meanwhile, the Mean Absolute Error (MAE) calculates

the average absolute difference between predicted and actual values, with a value of 0.25 indicating a 0.25 unit average absolute prediction error.

- CNN(ConV2D)

The evolution of a ConV2D model over a period of 20 epochs. The model starts with a comparatively high loss of 2.4061 and a low accuracy of 0.2215 in Epoch 1, indicating poor performance. However, as training continues, the loss lessens and accuracy improves. By Epoch 20, the model had achieved a far lower loss of 0.1197 and an accuracy of 0.9620, demonstrating significant improvement in its ability to predict accurately. The validation loss and accuracy, are also tracked to ensure that the model generalises effectively to previously unseen data. The graphs below represent Model Loss and Model Accuracy respectively

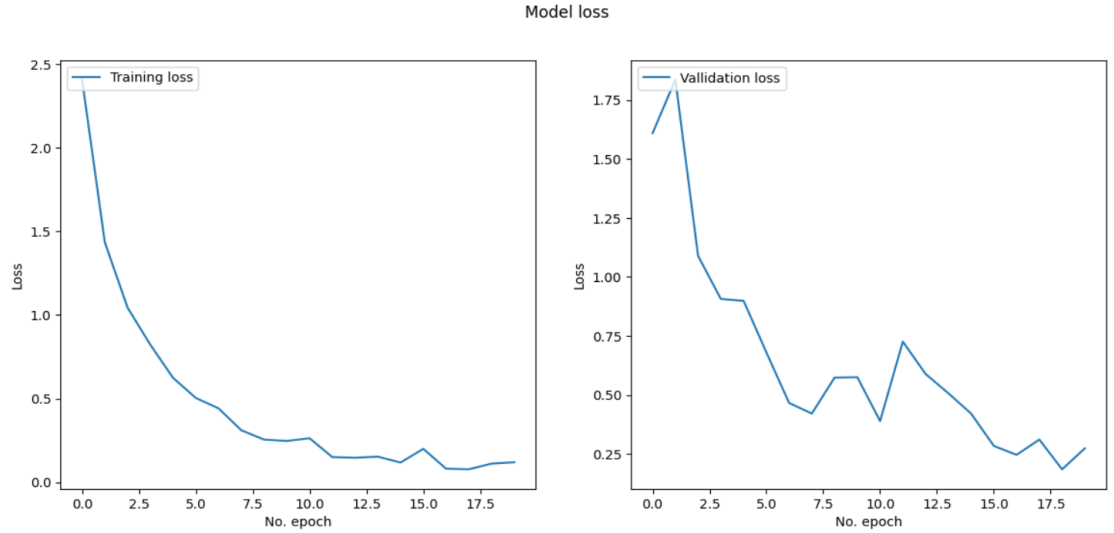


Fig. 4.3 ConV2D model loss

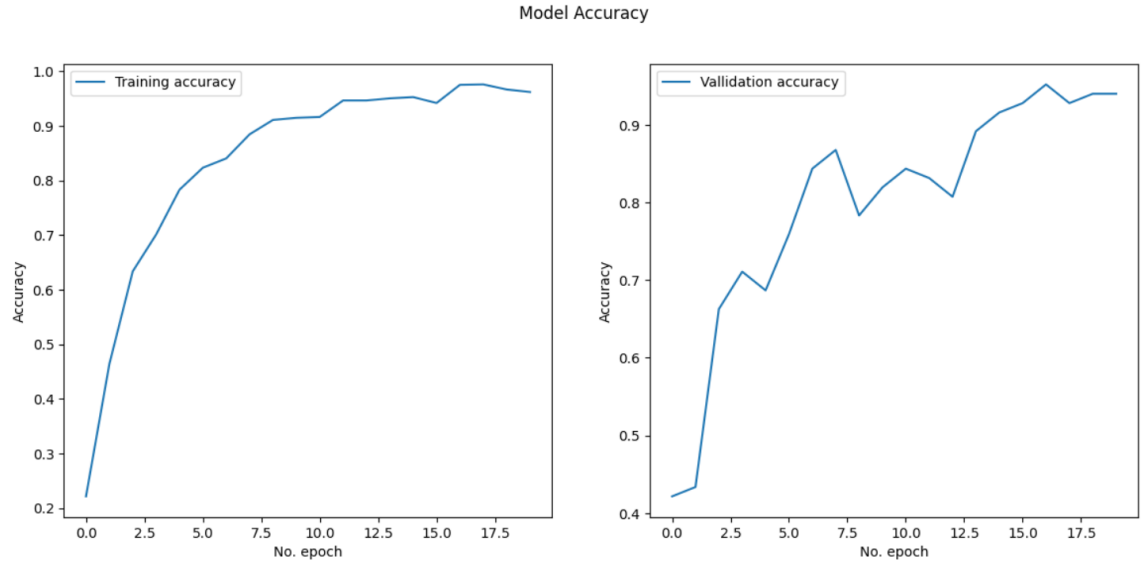


Fig. 4.4 ConV2D model loss

The model's accuracy of 0.94 indicates that it properly classified 94% of the occurrences. Precision of 0.93 means that the model is 93% accurate when predicting a positive class. Furthermore, the recall, or capacity to accurately identify positive occurrences, is 0.92,

confirming the model's ability to capture 92% of true positive situations. The F1-Score, which is similarly 0.92, reflects this balance between precision and recall, indicating the model's solid performance. The matrix's diagonal members reflect accurate predictions for each class, whereas the off-diagonal elements represent errors.

The Mean Squared Error (MSE) is a calculation that calculates the average squared difference between the model's predictions and the actual data. The MSE in this situation is quite low at 0.34, indicating that the squared errors between predicted and actual values are rather minimal on average. This is translated into a more understandable scale by the Root Mean Squared Error (RMSE), which is the square root of the MSE. With an RMSE of 0.58, it implies that the model's predictions have an average error of 0.58 units on the same scale as the target variable. Finally, the Mean Absolute Error (MAE) calculates the average absolute difference between predicted and actual values, revealing the model's precision.

With an MAE of 0.12, it shows that the model's predictions differ from the actual values by only 0.12 units on average.

- RNN+GRU

This training log shows the evolution of the RNN+GRU model over 50 epochs. Loss and accuracy are the two major parameters being tracked. The model starts with a high loss of 2.3698 and a low accuracy of 0.1115 in Epoch 1, indicating poor performance. However, as training goes, the loss diminishes consistently and the accuracy eventually improves. By Epoch 50, the model had a substantially lower loss of 0.8569 and a higher accuracy of 0.6878, indicating that it had learned to produce more accurate predictions over time. It's worth noting that this training history shows the model's learning process and convergence toward a higher-performing state. The graphs below display the Model Loss and Model Accuracy

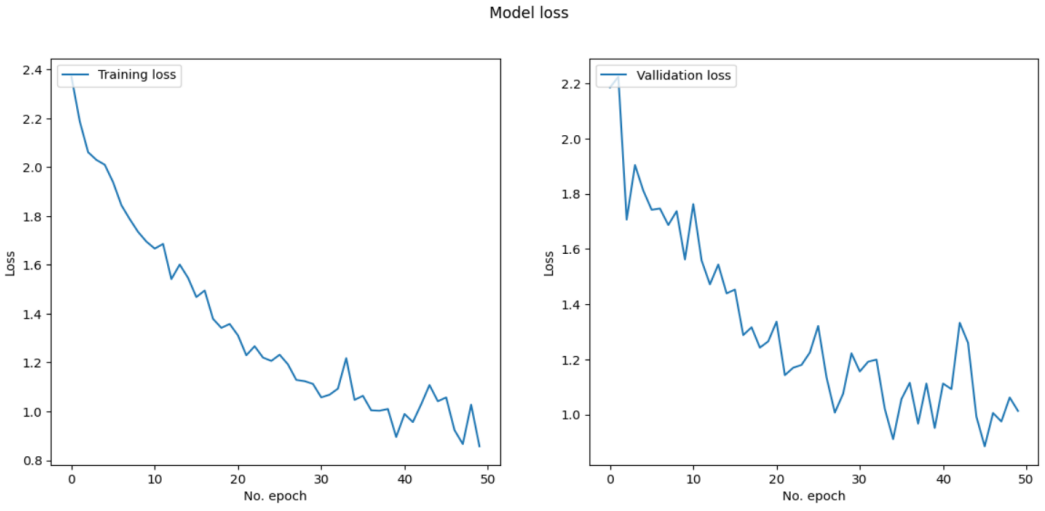


Fig. 4.5 RNN+GRU model loss

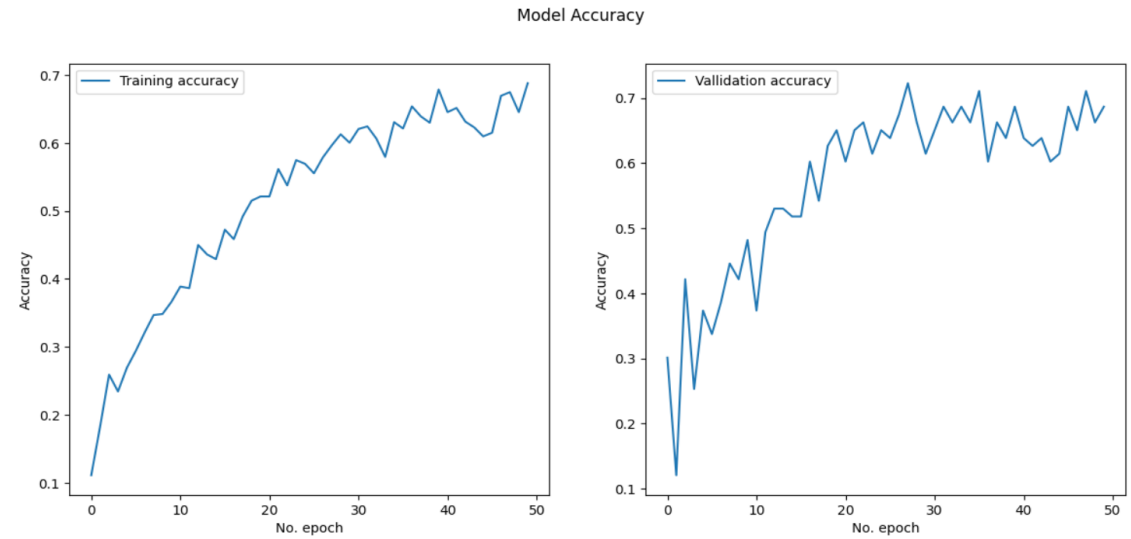


Fig. 4.6 RNN+GRU model accuracy

The classification metrics (accuracy, precision, recall, and F1-score) assess how effectively a model performs in classifying data. The model obtains roughly 69% accuracy in this scenario, which means it properly identifies 69% of the data. The confusion matrix gives a detailed breakdown of these classifications for each class, aiding in the identification of regions where the model may struggle to differentiate across classes. The Mean Squared Error (MSE) in regression evaluates the average squared difference between the model's predicted and actual values, with smaller values indicating better performance. The Mean Absolute Error (MAE) evaluates the average absolute difference between predictions and actual values, while the Root Mean Squared Error (RMSE) is a more interpretable variant of MSE. These regression measures assist in determining how well the model guesses continuous data.

- SVM

The model's accuracy is roughly 90.4%, implying that it properly identifies 90.4% of the data samples. The classification report contains a more extensive assessment of the model's performance for each class.The model achieves 100% precision and 93% recall for class 0, with an F1-score of 96%. This implies that the model is quite accurate at detecting class 0, and it also recovers the majority of true class 0 instances. Class 1, 2, 6, 7, 8, and 9 had similar high precision, recall, and F1-scores, showing strong achievement in these classes. The model's precision is poorer for classes 3 and 4, implying that it occasionally misclassifies other classes as 3 or 4. The recall is also moderate, implying that not all occurrences of these classes are captured. As a result, F1-scores are lower. While the precision for class 5 is 100%, the recall is 67%, indicating that it is good at avoiding false positives but misses some actual class 5 cases.

To summarise, the model outperforms most classes, with excellent precision, recall, and F1-scores. However, there is potential for development in classes 3, 4, and 5, particularly in terms of recall. Precision, recall, and F1-score macro and weighted averages provide overall performance indicators, with macro averaging giving equal weight to each class and weighted averaging taking class imbalance into account. For example, the weighted average F1-score is roughly 91%, suggesting high overall model performance.

Below is the classification report of the SVM Classifier.

```
Accuracy: 0.9036144578313253
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.93      0.96        14
           1       1.00      1.00      1.00        10
           2       0.89      0.80      0.84        10
           3       0.25      0.33      0.29         3
           4       0.50      0.67      0.57         3
           5       1.00      0.67      0.80         6
           6       0.89      1.00      0.94         8
           7       0.90      1.00      0.95         9
           8       1.00      1.00      1.00        13
           9       1.00      1.00      1.00         7

    accuracy                           0.90        83
   macro avg       0.84      0.84      0.84        83
weighted avg       0.92      0.90      0.91        83
```

Fig. 4.7 Classification Report of SVM Classifier

- Random Forest Classifier

The model has an accuracy of about 86.7%, which means it accurately identifies 86.7% of the data samples. The classification report contains a more extensive assessment of the model's performance for each class. The model has a precision of 100% and a recall of 93% for class 0, yielding an F1-score of 96%. This suggests that the model is quite good at detecting class 0 and effectively retrieving the majority of true class 0 occurrences.The precision for class 1 is 100%, suggesting high accuracy in classifying this category. However, the recall is only 80%, implying that it misses a few genuine class 1 instances. This class's F1-score is 89%, which is still good. Class 2 has 80% precision and 80% recall, yielding an F1-score of 80%, suggesting balanced performance. Classes 3 and 4 have precision and recall ratings of 50% and 67%, respectively, indicating that the model has difficulty correctly classifying these classes. This is evident in the 57% F1-scores.Class 5 has an F1-score of 73% with a precision of 80% and recall of 67%, suggesting good performance. Classes 6–8 perform well, with high precision, recall, and F1-scores. Class 9 performs admirably, with a precision, recall, and F1-score of 100%.
Below is the classification report of the classifier

```
Accuracy: 0.8674698795180723
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.93      0.96        14
           1       1.00      0.80      0.89        10
           2       0.80      0.80      0.80        10
           3       0.50      0.67      0.57         3
           4       0.50      0.67      0.57         3
           5       0.80      0.67      0.73         6
           6       0.89      1.00      0.94         8
           7       0.80      0.89      0.84         9
           8       0.92      0.92      0.92        13
           9       1.00      1.00      1.00         7

    accuracy                           0.87        83
   macro avg       0.82      0.83      0.82        83
weighted avg       0.88      0.87      0.87        83
```

Fig. 4.8 Classification Report of Random Forest Classifier

The bar plot below visualises the accuracy of the three regression models and two classifiers. CNN has the highest accuracy while RNN+GRU shows low accuracy among the regression models. SVM Classifier shows higher accuracy than the other, but random forest classifier has good accuracy.

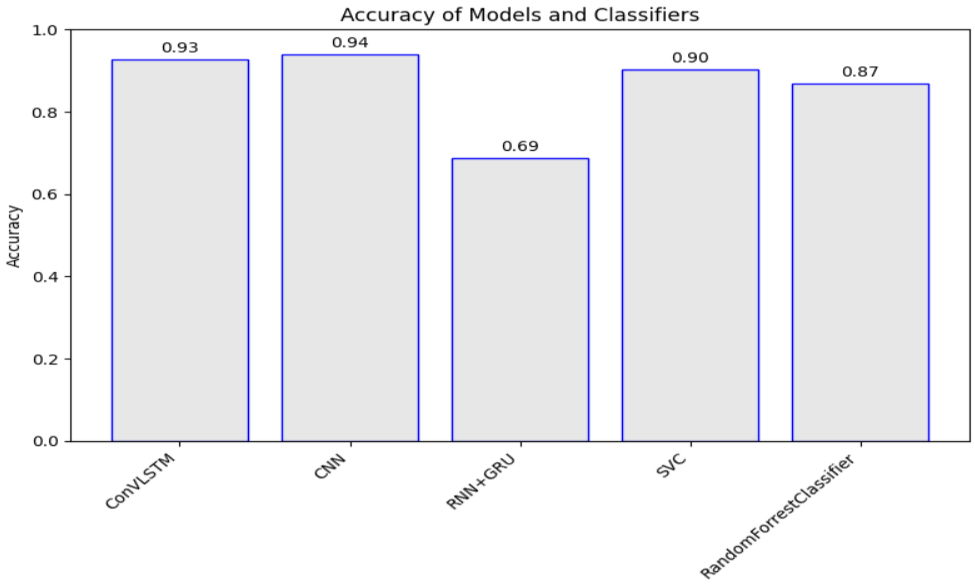


Fig. 4.9 Bar plot of Accuracy of Models and Classifiers

Using the OpenCV and Matplotlib packages, this grid of random video frames is generated. It accomplishes this by selecting a class at random from a list of class names and then selecting a video file from the relevant class directory. It uses OpenCV to read the first frame of the selected video, converts it from BGR to RGB format, then displays it as an image using Matplotlib. Each image now has a title that includes the class name, and the axis labels have been turned off.

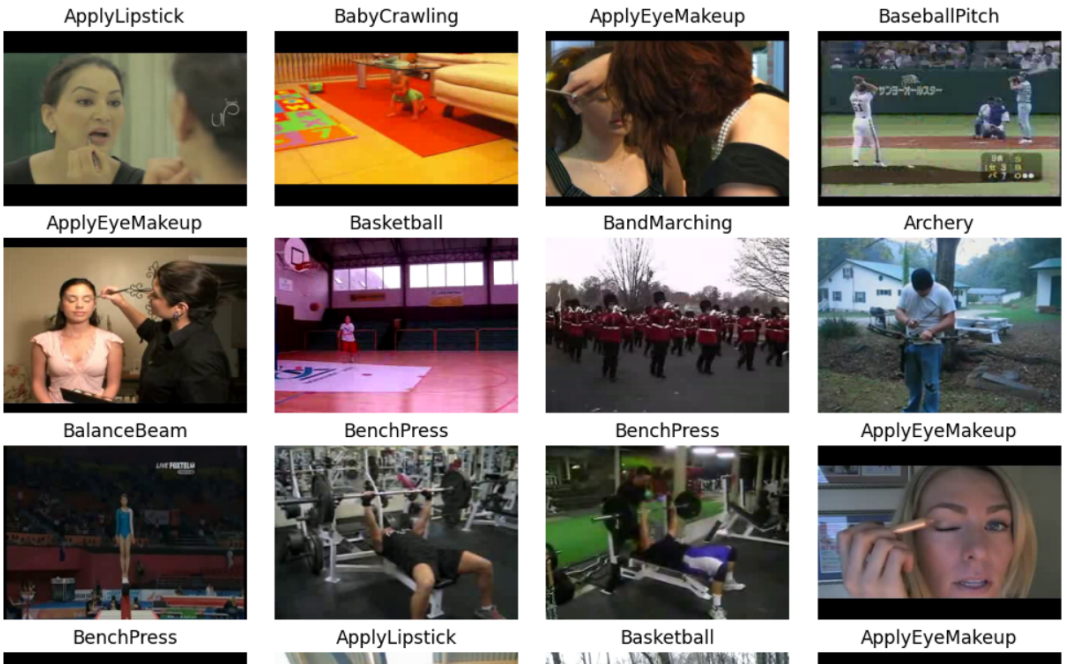


Fig. 4.10 HAR using Image Processing

Chapter 5 Conclusions

The research has been a precise exploration of algorithms' potential to unravel the complex web of human behaviours in the field of human activity recognition (HAR). We explored a variety of algorithms, such as CNN+LSTM, Conv2D, RNN+GRU, SVC, and Random Forest Classifier, each of which revealed its own particular advantages and adaptability. The experimental methodology, which included data gathering, preprocessing, choosing the best algorithm, training the model, and thorough evaluation, acted as the compass that led us on our journey. An intelligent approach to recognition was ensured through algorithmic selection that was in line with data characteristics. Comprehensive evaluation measures and robust model training revealed algorithmic performance above and beyond accuracy.

The acknowledged existence of the impact of contextual elements on human behaviours has become a focal topic, and this is known as context-awareness. Algorithmic models that can change with their environments show potential for real-world HAR applications. Each of these distinct machine learning and deep learning models revealed its own set of advantages and limitations during the evaluation process. The CNN (Conv2D) model excelled in image-based classification tasks, with an amazing accuracy of 94%. The ConVLSTM model came in second with a noteworthy accuracy of 93%, demonstrating its ability to handle sequential data. The SVM classifier also performed well, with an accuracy of 90%, making it a solid alternative for a variety of classification difficulties. The RNN with GRU, on the other hand, fell behind with an accuracy of 69%, indicating room for improvement or alternate model selection for sequential data tasks. While not the best performer in this group, the Random Forest Classifier delivered competitive results with an accuracy of 86%. Finally, the best model can be selected based on the specific requirements and characteristics of the given task, as each model excels in different scenarios.

In this study, ethical concerns were fundamentally necessary, and they emphasised the appropriate development and application of HAR technology. The tenets of informed consent, data protection, and transparency served as the unwavering foundation of our approach. In closing, we acknowledge the difficulties that lie ahead. The cutting edge of HAR research is in the reduction of data imbalances, real-time processing optimization, and dynamic ethical flexibility. Our journey serves as a testament to HAR's dynamic nature, where algorithms, ethics, and data come together to deepen our understanding of human motivations and encourage innovation in a range of fields.

Chapter 6 Future Work

A wide range of exciting new directions in the dynamic field of human activity recognition research await, each of which has the potential to increase knowledge, talents, and applications. The models, which use Long Short-Term Memory (LSTM) networks to manage temporal dependencies and Convolutional Neural Networks (CNNs) to collect spatial data, have shown remarkable promise. To maximise the potential of these systems, further effort needs to be done on tuning and optimization. Even more precise and reliable HAR systems might be developed by considering innovative hybrid combinations and implementing architectural advances. The accuracy and adaptability of HAR models can be improved by creating efficient methods to manage unbalanced datasets. To balance activity classes, look at cutting-edge data augmentation methodologies, synthetic data generation techniques, and adaptive sampling approaches. To balance activity classes, look at cutting-edge data augmentation methodologies, synthetic data generation techniques, and adaptive sampling approaches. Additionally, a more equitable evaluation of HAR systems may be facilitated by the development of uniform benchmark datasets that represent various populations and environments. Real-time HAR capabilities are necessary for many practical applications, including geriatric fall detection systems and gesture-based human-computer interfaces. It is possible to optimise algorithms and models to function in low-latency real-time scenarios. Real-time HAR systems must also have effective hardware implementations, edge computing, and low-power sensors in order to be easily incorporated into a variety of applications.

As HAR technology spreads, protecting individual privacy becomes more and more important. The study should investigate privacy-preserving methods that permit activity recognition while maintaining the privacy of user information. For creating privacy-conscious HAR systems, federated learning, secure multi-party computation, and differential privacy are attractive directions. Finding the ideal balance between data use and privacy protection is a difficult but crucial task.

In a broader sense there is a ton of potential for innovation and discovery in the field of human activity recognition. In order to bridge the gap between human actions and intelligent technology and to enable a future in which digital interactions are more natural and seamless, HAR is prepared to continue its transformative journey.

References

- Aktas, Eren. *A MACHINE LEARNING APPLICATION for CLASSIFICATION of HUMAN ACTIVITY RECOGNITION Human Activity Recognition in Machine Learning View Project*. 2018, https://doi.org/10.13140/RG.2.2.23917.95208. Accessed 27 Sept. 2023.
- An, Shuhua, et al. "Transfer Learning for Human Activity Recognition Using Representational Analysis of Neural Networks." *ACM Transactions on Computing for Healthcare*, vol. 4, no. 1, 31 Jan. 2023, pp. 1–21, https://doi.org/10.1145/3563948. Accessed 27 Sept. 2023.
- Anitha, U, et al. "ScienceDirect ScienceDirect-NC-ND License (Http://Creativecommons.org/Licenses/By-Nc-Nd/4.0/) Peer-Review under Responsibility of the Scientific Committee of the International Conference on Computational Intelligence and Data Science Robust Human Action Recognition System via Image Processing ScienceDirect-NC-ND License (Http://Creativecommons.org/Licenses/By-Nc-Nd/4.0/) Peer-Review under Responsibility of the Scientific Committee of the International Conference on Computational Intelligence and Data Science Robust Human Action Recognition System via Image Processing." *Procedia Computer Science*, vol. 167, 2020, pp. 870–877, https://doi.org/10.1016/j.procs.2020.03.426. Accessed 27 Sept. 2023.
- Au, Nathalie. *AUTOMATIC HUMAN ACTION RECOGNITION*.
- Beddiar, Djamila Romaissa, et al. "Vision-Based Human Activity Recognition: A Survey." *Multimedia Tools and Applications*, vol. 79, no. 41-42, 15 Aug. 2020, pp. 30509–30555, https://doi.org/10.1007/s11042-020-09004-3.
- Gupta, Neha, et al. "Human Activity Recognition in Artificial Intelligence Framework: A Narrative Review." *Artificial Intelligence Review*, 18 Jan. 2022, https://doi.org/10.1007/s10462-021-10116-x.
- "Human Activity Recognition (HAR): Fundamentals, Models, Datasets." *Www.v7labs.com*, www.v7labs.com/blog/human-activity-recognition. Accessed 21 May 2023.
- Lee, Junwoo, and Bummo Ahn. "Real-Time Human Action Recognition with a Low-Cost RGB Camera and Mobile Robot Platform." *Sensors*, vol. 20, no. 10, 19 May 2020, p. 2886, https://doi.org/10.3390/s20102886.
- Marszalek, Marcin, et al. "Actions in Context." *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, https://doi.org/10.1109/cvpr.2009.5206557. Accessed 27 Sept. 2023.
- Oukrich, Nadia. *Daily Human Activity Recognition in Smart Home Based on Feature Selection, Neural Network and Load Signature of Appliances*.
- "Papers with Code - UCF101 Benchmark (Action Recognition)." *Paperswithcode.com*, paperswithcode.com/sota/action-recognition-in-videos-on-ucf101?p=i3d-lstm-a-new-model-for-human-action. Accessed 27 Sept. 2023.
- Lalwani, Sanjay. "Get to Know All about Evaluation Metrics." *Analytics Vidhya*, 30 Sept.2022, www.analyticsvidhya.com/blog/2022/09/get-to-know-all-about-evaluation-metrics/.
-

- Sharma, Mr. Siddharth. “Human Activity Recognition Using OpenCv & Python.” *International Journal for Research in Applied Science and Engineering Technology*, vol. 8, no. 5, 31 May 2020, pp. 677–681, https://doi.org/10.22214/ijraset.2020.5106. Accessed 27 Sept. 2023.
- Soomro, Khurram, et al. *UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild*. 2012.
- Sun, Wei, et al. *Human Activity Recognition: Suitability of a Neuromorphic Approach for On-Edge AIoT Applications You May Also like Deep Learning Model for 3D Profiling of High-Aspect-Ratio Features Using High- Voltage CD-SEM - Multi-Branch CNN GRU with Attention Mechanism for Human Action Recognition Progress Report on High Aspect Ratio Patterning for Memory Devices OPEN ACCESS RECEIVED Human Activity Recognition: Suitability of a Neuromorphic Approach for On-Edge AIoT Applications*. 2022, https://doi.org/10.1088/2634-4386/ac4c38. Accessed 27 Sept. 2023.
- Umakanthan, Sabanadesan. “Human Action Recognition from Video Sequences.” *Eprints.qut.edu.au*, 2016, eprints.qut.edu.au/93749/. Accessed 27 Sept. 2023.
- Vrigkas, Michalis, et al. “A Review of Human Activity Recognition Methods.” *Frontiers in Robotics and AI*, vol. 2, 16 Nov. 2015, https://doi.org/10.3389/frobt.2015.00028.
- Wasim, Muhammad, et al. “A Novel Deep Learning Based Automated Academic Activities Recognition in Cyber-Physical Systems.” *IEEE Access*, 2021, pp. 1–1, https://doi.org/10.1109/access.2021.3073890. Accessed 2 May 2021.
- Xia, Kun, et al. “LSTM-CNN Architecture for Human Activity Recognition.” *IEEE Access*, vol. 8, 2020, pp. 56855–56866, https://doi.org/10.1109/access.2020.2982225.
- Yang, Ruoyu, et al. “CNN-LSTM Deep Learning Architecture for Computer Vision-Based Modal Frequency Detection.” *Mechanical Systems and Signal Processing*, vol. 144, 1 Oct. 2020, p. 106885, www.sciencedirect.com/science/article/pii/S0888327020302715, https://doi.org/10.1016/j.ymssp.2020.106885.
- Ye, Juan, et al. “USMART.” *ACM Transactions on Interactive Intelligent Systems*, vol. 4, no. 4, 28 Jan. 2015, pp. 1–27, https://doi.org/10.1145/2662870. Accessed 20 May 2022.
- Zahin, Abrar, et al. “Sensor-Based Human Activity Recognition for Smart Healthcare: A Semi-Supervised Machine Learning.” *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, 2019, pp. 450–472, https://doi.org/10.1007/978-3-030-22971-9_39. Accessed 27 Sept. 2023.

Appendix

Appendix A Terms Of Reference

Department of Computing and Mathematics Computing and Digital Technology Postgraduate Programmes Terms of Reference Coversheet	
Student name:	Shreeya Gulawani
University I.D.:	22515406
Academic supervisor:	Dr. Li Guo
External collaborator (optional):	
Project title:	Smart Daily Human Activity Reporting System
Degree title:	MSc Data Science
Project unit code:	6G7V007
Credit rating:	60
Start date:	15/06/2023
ToR date:	09/07/2023
Intended submission date:	22/09/2023
Signature and date student:	11/07/2023
Signature and date external collaborator (if involved):	

Aims & Objectives

This project aims to create a smart system for reporting daily human activity using machine learning models. The objective is to develop a reliable and effective system that uses sensor data to identify and categorise different human activities. The system will acquire data, separate useful features, develop machine learning models, and provide activity tracking and reporting in real-time. In addition, the project addresses user customization, scalability, privacy, and security issues intending to develop a workable and user-friendly solution that can be used in actual environments.

Learning Outcomes

The learning outcomes for this project are as mentioned below:

- Collecting, and analysing the data and carrying out feature engineering, data cleansing, and visualisation methods.
- Learning hyperparameter tuning, model selection, and optimization strategies to enhance the performance and accuracy of the models.
- Learning about the various machine learning algorithms, including how they might be used to solve activity identification issues.
- By tackling issues with data quality, model performance, scalability, and real-time processing, improving problem-solving abilities.

Project Description

The necessity for automated technologies that can precisely track and record human activity is what spurred the initiative. Traditional approaches are frequently arbitrary and time-consuming. The system intends to give accurate information by utilising these technologies, encouraging a healthier and more informed way of living.

Using sensor data gathered from sources, the system will automatically track and report various human actions. The system will be able to identify and categorise actions including walking, jogging, sitting, working, and sleeping by using machine learning techniques. Human activities will be recognized and classified by machine learning models. The smart daily human activity reporting system can be used in several industries, including healthcare, fitness tracking, etc. Individuals will be able to get perspective on their daily activities, make wise decisions, and enhance their general well-being.

References

- Oukrich, N. (2019). Daily Human Activity Recognition in Smart Home based on Feature Selection, Neural Network and Load Signature of Appliances. [online] hal.science. Available at: https://hal.science/tel-02193228.
- Krasimir Tonchev, Sokolov, S., Yuliyan Velchev, Georgy Balabanov and Vladimir Poulkov (2015). Recognition of Human daily activities. doi:h ttps://doi.org/10.1109/iccw.2015.7247193.
- Salguero, A.G., Medina, J., Delatorre, P. and Espinilla, M. (2018). Methodology for improving classification accuracy using ontologies: application in the recognition of activities of daily living. Journal of Ambient Intelligence and Humanized Computing, 10(6), pp.2125–2142. doi:https://doi.org/10.1007/s12652-018-0769-4.

Evaluation Plan

Testing and performance measurements are part of the evaluation plan for the smart daily human activity reporting system. Evaluation of the effectiveness of the system's activity recognition and real-time tracking using a variety of measures will be done. It will be determined whether the system's scalability, security, and privacy controls meet requirements. Comparing and recording results along with suggestions for advancement in the future.

Appendix B Experimental Code

Available on request from Dr. Li Guo.