Abnormality Detection from Daily Living Activities of Autistic Patients using Deep Learning Model of Video Analytics

A

thesis

submitted to

LOVELY PROFESSIONAL UNIVERSITY

For the degree of

DOCTOR OF PHILOSOPHY (Ph.D)

in

Computer Applications

By

Ankush Manocha

11616681

Supervised By

Dr. Ramandeep Singh

LOVELY FACULTY OF TECHNOLOGY AND SCIENCES

LOVELY PROFESSIONAL UNIVERSITY

Punjab

2019

Declaration of Authorship

I, Ankush Manocha, declare that this thesis titled, "Abnormality Detection from Daily Living Activities of Autistic Patients using Deep Learning Model of Video Analytics" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

Certification

This is to certify that the thesis entitled "Abnormality Detection from Daily Living Activities of Autistic Patients using Deep Learning Model of Video Analytics", which is being submitted by Mr. Ankush Manocha for the award of the degree of Doctor of Philosophy in Computer Applications from the Faculty of Technology and Sciences, Lovely Professional University, Punjab, India, is entirely based on the work carried out by him under my supervision and guidance. The work reported, embodies the original work of the candidate and has not been submitted to any other university or institution for the award of any degree or diploma, according to the best of my knowledge.

Dr. Ramandeep Singh
Associate Professor
Department of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab-144411, India
Date:

Abstract

Physical and psychological disorders are the leading cause of disability all over the world. Over 25% of the population is suffering from disability and premature mortality in the United States and Canada. The same ratio can also be expected from the other nations as well. In 2013, the United States spent approximate 201 billion dollars for public healthcare care sector focusing on physical and psychological disorders. Significant efforts have been made to develop smart technological approaches to alleviate the adverse impacts of physical and psychological disability. The physical and psychological issues are commonly predicted by analyzing the motor movements of the patients in a continuous manner. The early prediction of abnormality through continuous monitoring can be considered as the best solution for making positive change in the life of patient and their family. Usually, the observation can be done by watching exercises of the individuals either from video recordings or live, and further, the behaviors are being categorized. The process of manual monitoring is time-consuming and requires a significant amount of efforts. Therefore, automating the process of monitoring can enhance the ability of irregularity prediction and behavior observation more efficiently.

This thesis present efforts of combining the advanced monitoring principles of computer vision with modern data processing techniques to provide aid in healthcare and assistive-care domain. The core of this dissertation is to contribute to the research in smart healthcare by identifying physical or psychological issues which can be monitored remotely. For this, application-specific four distinct frameworks are proposed to monitor different health issues in real-time. Irregular activities are a class of behavior known to occur in patients. The calculated outcomes are presented to validate the ability of the proposed frameworks for behavior analysis based on the performed

physical activities. The experimental results show promising scope towards improving the quality of life and cost-effective smart monitoring solutions.

Acknowledgements

First of all, I would like to express my gratitude to my supervisor, Dr. Ramandeep Singh, for his supervision, advice, and guidance from the very first day of this research as well as giving me extraordinary experiences throughout the work. I am truly very fortunate to have the opportunity to work with him. I found this guidance to be extremely valuable.

I am grateful to the friends and fellow researchers, particularly Dr. Prabal Verma for their constructive criticism and suggestions.

I would like to show my gratitude to the entire family of Lovely Professional University for providing me a suitable research atmosphere to carry out my work in proper time. I would like to thank the Division of Research and Development and School of Computer Applications for all the support encouragement throughout the research work.

I am also very much grateful to my mother, Mrs. Mamta Manocha and my father, Mr. Satish Manocha, and all my sisters for their moral support and care that they shown me during the period of this work.

Last but not least, I thank God for sailing me through all the rough and tough times during this research work.

Ankush Manocha **Date.....**

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviations	Description
ASD	Autism Spectrum Disorder
OCD	Obsessive-Compulsive Disorder
GAD	Generalized Anxiety Disorder
ADL	Activities of Daily Living
HAR	Human Activity Recognition
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
3D	3 Dimensional
2D	2 Dimensional
GRU	Gated Recurrent Unit
MLP	Multilayer Perceptron
ReLU	Rectified Linear Unit
MLP	Multilayer Perceptron
FS	Frame Segment
IoA	Index of Abnormality
ARS	Activity Recognition Score
WBAN	Wireless Body Area Network
QoS	Quality of Services
TAL	Temporal Activity Logs
SVM	Support Vector Machine
HMM	Hidden Markov Model
HVI	Health Vulnerability Index
DoI	Degree of Irregularity
IDT	Improved Dense Trajectory

RGB	Red Green Blue
GPU	Graphical Processing Unit
SoA	Scale of Abnormality
IoMT	Internet of Multimedia Things
EOT	Edge of Things
FPS	Frames Per Second

I would like to dedicate my thesis to my family and especially, one of my beloved sister, Mrs. Ruchi Arora.

Abstract

Physical and psychological disorders are the leading cause of disability all over the world. Over 25% of the population is suffering from disability and premature mortality in the United States and Canada. The same ratio can also be expected from the other nations as well. In 2013, the United States spent approximate 201 billion dollars for public healthcare care sector focusing on physical and psychological disorders. Significant efforts have been made to develop smart technological approaches to alleviate the adverse impacts of physical and psychological disability. The physical and psychological issues are commonly predicted by analyzing the motor movements of the patients in a continuous manner. The early prediction of abnormality through continuous monitoring can be considered as the best solution for making positive change in the life of patient and their family. Usually, the observation can be done by watching exercises of the individuals either from video recordings or live, and further, the behaviors are being categorized. The process of manual monitoring is time-consuming and requires a significant amount of efforts. Therefore, automating the process of monitoring can enhance the ability of irregularity prediction and behavior observation more efficiently.

This thesis present efforts of combining the advanced monitoring principles of computer vision with modern data processing techniques to provide aid in healthcare and assistive-care domain. The core of this dissertation is to contribute to the research in smart healthcare by identifying physical or psychological issues which can be monitored remotely. For this, application-specific four distinct frameworks are proposed to monitor different health issues in real-time. Irregular activities are a class of behavior known to occur in patients. The calculated outcomes are presented to validate the ability of the proposed frameworks for behavior analysis based on the performed physical activities. The experimental results show promising scope towards improving the quality of life and cost-effective smart monitoring solutions.

Chapter 1

Introduction

A group of physical or psychological disorder originating in childhood are categorized as developmental issues. Psychological disorders comprise of issues pertaining to learning, language, motor movement and autism spectrum disorders. Broadly, these disorders are characterized as neurodevelopmental disorders. Identifying and determining irregularities at its initial stage is the most desirable. The perception of individuals in a continuous manner is one of the most challenging tasks. Visual sensors-based continuous monitoring can become a key aspect to deal with these issues at ground level. The monitoring can either be in real-time or recorded for later annotation. This helps medical professionals to intervene earlier to get positive outcomes. Therefore, the auto-prediction of such irregularities is highly desirable both for parents and medical representatives.

Computer vision is a field of science that provides machines with the ability to visualize (Szeliski, 2010) and have a profound impact on multiple domains as shown in Figure 1.1. Tracking and following the activities of an individual from videos (Yilmaz, Javed, and Shah, 2006) and multi-model sensors (Munaro and Menegatti, 2014) is considered to be the most interesting topic of computer vision. One of the primary requirements of these applications is to learn a recognition model that can differentiate different types of activities. Several advanced computer vision assisted monitoring solutions have been proposed to analyze the type of behavior based on physical activities. In its early stage, several human tracking (Kazemi and Sullivan, 2014) and face detection (Krizhevsky, Sutskever, and Hinton, 2012; Viola

and Jones, 2004) solutions have been introduced by utilizing handcraft approaches with satisfactory detection accuracy. As the quality and the amount of the data has been increased, the fewer data processing capability of handcraft techniques became the major cause of the low satisfaction level in event understanding (Papert, 1966). Some of the challenges such as, large intra-class variance among the activities, large variability in spatiotemporal scale, the variability of human pose, periodicity of human action, scarcity of labeled data, and many other are the most common challenges in computer vision which are very hard to achieve by utilizing handcraft techniques.

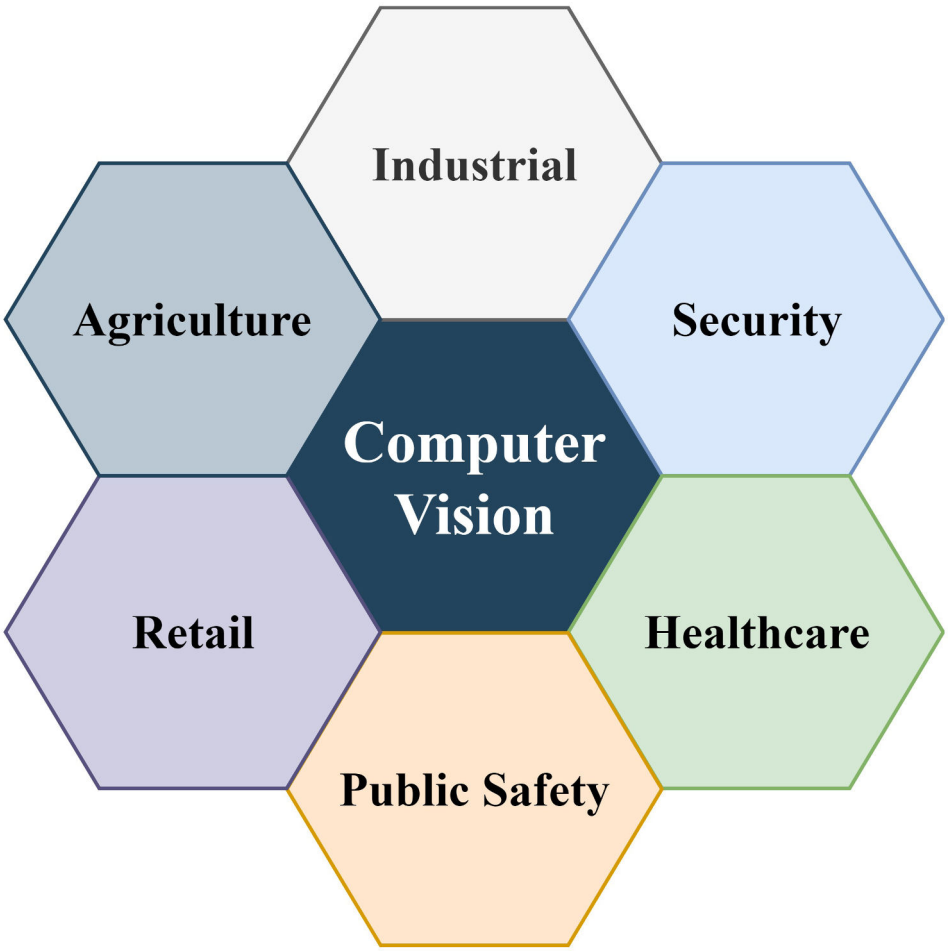

FIGURE 1.1: Computer vision and its applications.

In (Krizhevsky, Sutskever, and Hinton, 2012) and (LeCun et al., 1988),

deep learning models were introduced in computer vision for the task of content recognition from the images. Deep learning (LeCun, Bengio, and Hinton, 2015) is considered as the best option to overcome the above-mentioned challenges by utilizing the ability of data processing and achieving human-level understanding. By increasing the computational power, the data processing techniques of deep learning such as a convolutional neural network (CNN) (Krizhevsky, Sutskever, and Hinton, 2012) has prompted remarkable enhancements for image classification and robust detection of people (Ren et al., 2015). Several domains have been served by computer vision assisted deep learning enabled monitoring solutions. But the healthcare and assistive-care domain is considered as one of the most challenging application domain of computer vision. Predicting regular and irregular activities of individuals through visual sensors is considered as one of the most demanding yet challenging problems of computer vision.

1.1 Prior research

In this section, we have discussed several aspects of the proposed study, from mental health assessment to computer vision and learning techniques. These techniques can be utilized to comprehend observed symptoms as follows:

1.1.1 Observable physical and psychological irregularities

The symptoms of psychological illnesses can be observed visually. To observe psychological irregularities, continuous observation can become the most reliable solution. Visually recognizable indicators are available for several psychological illnesses including autism spectrum disorder (ASD), obsessive-compulsive disorder (OCD) and schizophrenia. Computer-vision assisted continuous monitoring can become a primary source of data collection that can further used for medical or assistive assessment.

ASD is considered as one of the most common neurological developmental disorder in the area of psychological disorder. Behavioral symptoms including social behavior deficits and communication as well as repetitive and restricted behavior patterns such as spinning, hand flapping and head banging are the most common issues of ASD (Goldman et al., 2009). These impairments may vary on the scale of severity from individual to individual. The report conducted by the Centers for Disease Control and Prevention in 2000 listed that around 1 of 150 children were experiencing ASD (Health and Services, 2017). However, the latest report of 2014 clarified that around one of 59 children by the age of 8 years are experiencing ASD in the United States (Baio et al., 2018).

As we can observe that ASD turning out to be progressively predominant consistently. The ratio of males diagnosed with ASD is always higher as compared to females. 3 males out of 4 individuals are determined to have ASD. It can be analyzed that the behavior of the child is the exact opposite of a normal child. The combination of odd behavioral effects starting at a young age can demonstrate that the child may have developmental issues at a later stage and must get tested. When a child is predicted with autism, the providence of appropriate predictive, diagnostic and treatment solutions can be a challenging task for parents and caretakers. Autism Diagnostic Observation Schedule (ADOS, Lord et al., 2000) is a procedure which is used to diagnose the state of the child and usually takes a minimum of six months. Once analyzed, the child with mental imbalance may get concentrated before atypical examples of behavior and the functioning of the brain become firmly settled.

OCD is another psychological illness that affected 2-3% of the adult population. The individual suffering from OCD usually experienced a high scale of distress and anxiety. The disorder is characterized by having intractable, intrusive obsessions (Grabill et al., 2008). As like most of the psychological disorders, this disorder is also evaluated through self-reporting, checklists

and interviews. Children's Yale-Brown Obsessive Compulsive Scale (CY-BOCS) (Scahill et al., 1997) checklist is considered as the most common solution to measure the scale of anxiety. Characteristics of this illness are easily recognizable. Usually, patients have an impulsive desire to organize things in a specific manner (Radomsky and Rachman, 2004). The most common habits in OCD include: repeatedly washing hands; following a specific order for doing tasks every time; repetitively checking doors, switches etc.; fear of touching doorknobs or shaking hands.

Individuals suffering from schizophrenic are more likely to have symptoms like delusions and hallucinations. Psychiatric analysts have detailed about recognizing the difference to determine whether the individual is suffering from schizophrenic or not from video recordings (Walker, Savoie, and Davis, 1994; Schiffman et al., 2004). The individuals were observed by analyzing motor skills and further rated to differentiate the scale of the abnormality. The proposed study found a strong correlation in a high frequency of neuro-motor irregularities and the inevitable analysis of schizophrenia. Schizophrenic subjects suffering from neuro-motor irregularities cause several abnormalities named as musculoskeletal abnormalities, postural abnormalities, and spastic movements. In article (Schiffman et al., 2004), the authors utilized recordings of Danish children which were recorded in 1972. They also utilized data of the individuals suffering from mental disabilities. The researchers found that the subjects who were determined to have schizophrenia had scored lower on their social rating and in motor skills. Perception of neuro-motor based irregularities are considered as the best possibility for programmed perception.

1.1.2 Computer vision and Neurological Development Disorder (NDD)

Due to the wide range of NDD, routine practices vary accordingly in patients experiencing Neurodevelopment issue. Thus, several studies have been developed to discover patterns of irregular behavior. A perfect solution would

perceive and classify symptoms of NDD, for example, distinguishing an individual with psychological impairment and grouping the strange behavior related to the performed exercises. Fulceri et al., 2015 demonstrated that physical movements were obviously impaired in pre-schoolers with ASD, thus, it behoved the researchers to automate the procedure of identifying irregularities to discover complex practices. In articles (Hashemi et al., 2012; Hashemi et al., 2014; Luyster et al., 2009), authors have observed and classified behavior including attention disengagement, visual tracking, and arm asymmetry. They studied ADOS-T and Mullen Scales of Early Learning (Mullen, 1995) test based clinical data. The proposed studies recognized the ability to turn head and effectively distinguished arm asymmetry from the recorded video of individuals.

Recent work in this domain deals with the extremely difficult task of distinguishing children and their practices in a classroom (Sivalingam et al., 2012). Sivalingam et al., 2012 utilized visual sensors and depth sensors to track the activities of the children in a classroom (Walczak et al., 2013) and recognized social gatherings and other risk-markers (Fasching et al., 2012). Fasching et al., 2015 effectively characterized redundant behavior of the autistic child like ear-covering, shrugging, and hand fluttering from the videos. Bernstein et al., 2017 actuated and marked irregular practices in children with obsessive-compulsive disorder in an artificial setup. They analyzed the practices of the children such as hand-washing in a dummy washroom and the organization of items on various floor covering designs. They found that arranging and moving objects and longer hand washing was related to obsessive-compulsive scale evaluations. In the proposed work (Fasching et al., 2016), authors created an environment to follow how many times participants went to the washroom and turned the water tap on and off. The study was performed in an artificial environment, therefore, children do not need to be acquainted with sensors. Rajagopalan, Dhall, and Goecke, 2013 arranged public recordings of children with ASD into normal generalizations to give researchers challenging dataset for developing effective monitoring solutions. This thesis aggregates an imperative assortment

of work that attempts to automatically and robustly distinguish, track, and classify abnormal physical movements and motions.

1.1.3 Machine learning and deep learning for physical activity-based behavior prediction

In general, activities are recognized and classified by extracting and analyzing the features from the visual data. Methodologies for activity recognition are broadly categorized based on the type of features used to analyze activities as follows:

1. Low-level feature-based methods
2. Mid-level feature-based methods
3. High-level feature-based methods

Low-level feature based methods: Low-level feature based methods processed the features in two steps – first, an interest point is detected and then local features are described to analyze the type of activity (Laptev, 2005; Wang et al., 2011). The extracted local features are further process before applying any classification methods to categorize the type of activity performed by the individual.

Mid-level feature based methods: The mid-level features directly dependent on the capacity to discover and process human and its direction in the video before action recognition. The researchers exploit trajectories (Wang and Schmid, 2013), temporal sequence of the human pose (Thurau and Hlavác, 2008) and human tracks (Minhas, Mohammed, and Wu, 2012) etc. for activity recognition.

High-level feature based methods: Exercises are outlined as a collection of semantic characteristics such as semantic model vectors (Merler et al., 2012), actoms (Gaidon, Harchaoui, and Schmid, 2013), action bank (Sadanand and Corso, 2012), and many others. Based on the above highlights, some

graphical model-based optimization algorithms are utilized to improve the execution such as conditional (Quattoni et al., 2007), petri net (Albanese et al., 2008), Markov random field (Nayak, Zhu, and Roy-Chowdhury, 2013), etc.

Recent proposed studies utilized the data processing capability of deep learning to analyze the contextual information for activity analysis. Deep learning plays an imperative role in several challenging domains such as computer vision, natural language processing, speech recognition, biology and physics problems. To provide the ability of content recognition to machines, the models need to extract dense features from video and images. Deep learning (Goodfellow, Bengio, and Courville, 2016) is considered as one of the most encouraging fields in the computer vision domain and is extremely successful for analyzing high dimensional raw data in real-time. Deep learning techniques such as CNN is considered as a most popular feature extraction approach and most commonly applied to analyze visual imagery. Convolutional Neural Network (CNN) utilized global and local features to recognize complex physical activities.

To access the temporal features from the sequence of frames for sequential activity representation, 3D CNN (Baccouche et al., 2011; Ji et al., 2013; Tran et al., 2015) extract the graphical characteristics from a set of frames in a video. A most common challenge of 3D CNN based activity prediction methodology is discovering a large amount of labeled data for training and getting robust prediction accuracy. The recent studies examined the performance of CNNs in the detection of more complex activities. In article (Karpathy et al., 2014), authors utilized convolutional networks to find the relationship between temporal information for activity recognition. They also observed that the prediction performance can be improved by incorporating temporal features properly. In recent studies, (Ji et al., 2013; Simonyan and Zisserman, 2014a; Tran et al., 2015; Gkioxari and Malik, 2015), the activity prediction performance has also been improved by augmenting the proposed approaches using optical flow-based features. The utilization of features varies network

by network.

1.2 Motivation of the study

Human eyes can capture ten gigabits data for each second from general surroundings and the human brain can process more than three million bits of the data for every second (Anderson, Van Essen, and Olshausen, 2005). Our brain utilizes the data to gain proficiency with an amazing representation of the world (Barlow, 1989). Similar to human vision, smart monitoring is a process toward finding relative meanings from the visual data (Man and Vision, 1982). However, developing an intelligent monitoring system as vigilant as a human vision is a challenge for researchers of several fields such as philosophy, physiology, engineering, psychology, artificial intelligence and computer science.

Computer vision has begun to serve several domains in which different methodologies can be connected together to improve the efficiency of monitoring (Hashemi et al., 2014; Rehg et al., 2013; Rajagopalan, Dhall, and Goecke, 2013; Ciptadi, Goodwin, and Rehg, 2014). In recent years, it has become an essential solution to observe the regular and irregular activities of an individual or a group of individuals. Healthcare domain is considered as one of the most challenging domain of computer vision. However, computer vision prompt issues related to activity misclassification. The occurrence of misclassifications can be handled by training the model with more complex activity samples. A typical computer vision assisted activity monitoring system is represented in figure 1.2.

With a continuous progression in the amount of high quality visual data and computational power, the technology of computer vision promises a significant advancement in smart monitoring. A noteworthy advantage of computer vision has the potential of combining the data processing capability of deep learning for recognizing more complex physical activities with satisfactory scale of precision. Another favorable advantage of this technology

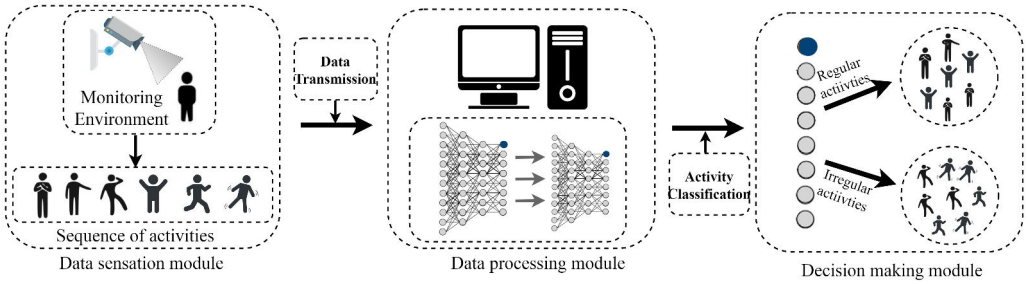

FIGURE 1.2: A typical computer-vision assisted monitoring framework.

is, an individual can be monitored without installing any wearable sensors on their body. The above mentioned advantages of computer vision and deep learning become a primary motivation of developing smart monitoring frameworks for healthcare and assistive-care domain. These frameworks can help caretakers and doctors to deal with patient's health irregularities in real-time. In addition, having the ability of designing smart monitoring solutions for the patients suffering from psychological disability may add our comprehension in the study of neuroscience (Sterling and Laughlin, 2015).

1.3 Objectives of the study

In order to investigate the benefits of utilizing and the advanced monitoring principles of computer vision with the data processing benefits of deep learning, the following objectives were proposed:

- To develop physical irregularity recognition and behavior analysis framework for patients suffering from physical or psychological disabilities to be used in their localized areas.
- To summarize the scale of irregularity in a continuous manner by storing the final outcomes in the database and also present the pattern of activities with observed anomalies in graphical format.
- To initiate appropriate warning and emergency alert-based notifications for caretakers and doctors during the involvement of the individual in any irregular activities.

1.4 Thesis Contribution

The proposed activity recognition methodologies are directly inspired by the data processing capabilities of deep learning that learn patterns by encoding the input through various layers and sub-sampling operations. Deep learning also helps to increase the level of abstraction by processing the data level by level. The primary objective of the proposed studies is to increase safety measures for individuals under observations by determining the scale of irregularities more effectively. The purpose of the presented solutions is to make the environment smarter that can quickly distinguish irregular behavior so that caretakers and physicians can assess patients as soon as possible. The contributions that will be presented in this thesis are described here:

- Established state of the art healthcare and assistive-care environment by proposing several application-oriented computer vision-assisted monitoring frameworks.
- Demonstrated the applicability of computer vision methodologies to the healthcare domain by being able to classify irregular physical activities.
- Developed a method for storing and retrieving the calculated outcomes from the database for medical or therapeutic assistance.
- Developed a time-sensitive smart alert-based decision-making solution to maintain the sensitiveness of healthcare domain.

1.5 Thesis Outline

The thesis is organized in six chapters. A brief outline of the chapters is given below.

In Chapter 2, we present and evaluate a novel abnormality recognition framework for the individuals suffering from Autism Spectrum Disorder

(ASD). The main objective of the proposed study is to increase safety measures for individuals in the absence of parents or caretakers. The simulation of the proposed study and the results are thoroughly analyzed.

In Chapter 3, we propose computer vision-assisted deep learning-enabled monitoring framework to analyze the physical posture of an individual in their working environment. The motive of the proposed study is to recognize physical irregularities that can cause Generalized Anxiety Disorder (GAD). The outcomes are also analyzed in the chapter.

In Chapter 4, we introduce an edge-analytics assisted monitoring framework to predict health affliction in real-time. The fundamental goal of the proposed study is to provide a medical assessment to the patient in real-time by notifying the respective doctor or caretakers. The performance of the proposed framework is also analyzed in the chapter.

In Chapter 5, we present a hybrid deep learning methodology for recognizing physical movements of the patients in real-time through 1D time series data captured from wearable sensors to deal with the constraints of the above-presented computer-vision assisted monitoring frameworks. The results of the proposed study are analyzed with proper analysis.

In chapter 6, we conclude the thesis highlighting the prime outcomes of the current research of the author and the significant contribution of the thesis. We also notify about the scope for future research in this area.

Chapter 2

Physical Irregularity Recognition of ASD Children

2.1 Introduction

Autism Spectrum Disorders (ASD): ASD is a type of neurological development disorder which affects the individual's social, linguistic and communicative skills ((Lord, Volkmar, and Lombroso, 2002; Association, 2013)). The behavioral issues of these individuals are broadly classified into three categories ((Hutt and Ounsted, 1966; Hutt, Forrest, and Richer, 1975; Dunlap, Dyer, and Koegel, 1983)):

1. **Social interaction:** In most of the cases, an autistic child is less likely to interact with others ((Myers, Mackintosh, and Goin-Kochel, 2009)). Even, they are less capable to express their needs in a proper manner which debilitates their capability of sharing activities and interests with their parents or caregivers.

2. **Communication:** An autistic child may experience issue in the development of communication ability. As indicated by (Griffith et al., 2010), an autistic child face difficulties in nonverbal communication, for example, pointing, eye to eye connection, and hand gestures.

3. **Repetitive behaviors:** Repetitive behavior is considered as the most common issue in autism (Ganz, 2007). Several redundant exercises like

spinning objects, closing and opening doors are some of the examples of repetitive behavior that can cause physical injuries.

Problem Identification: Recent statistics described that by the age of eight-years, 1 out of every 88 children is suffering from autism in the United States (https://www.cdc.gov/ncbddd/adhd/data.html). Similar statistics can also be expected from the other nations as well. It can be noticed that dealing with an autistic child is a completely different experience with more stress for parents compared to the care of a normal child. To analyze their routine based challenges and difficulties, a subjective questionnaire has been conducted:

1. What kind of situations are more critical and difficult for you?
2. Which sort of arrangements is you expecting in your daily routine that can help to make things less demanding for your child?
3. What kind of solutions do you expect to deal with the most common daily problems related to your child?

The questionnaire is concluded into three portions: first, focusing on the daily experiences of the parents with their child. Second, living environment expectations for both parent and child. Third, the expectation of ease in their daily routine by monitoring the child in real-time. Several solutions are proposed to deal with the problem of autism at some level. But according to the best of our knowledge, no specific solution is found that can help parents for child assistance in an independent manner.

Research field: Computer-vision and deep learning are considered as the two most supportive technologies to provide an effective and efficient smart assistive environment. For activity recognition, several handcraft techniques based monitoring solutions have been proposed (Ezzahout and Thami, 2013). But as the amount of data is increasing, these solutions become less capable of real-time data processing and takes more time for decision making (Chu, Song, and Jaimes, 2015). Due to the sensitiveness of the healthcare domain,

the decision making delay caused by these techniques is unacceptable. To overcome the constraint of handcraft features, deep learning is considered as the best solution for real-time data processing and decision making.

Motivation: Smart activity monitoring and irregularity prediction in real-time can provide a reliable assistive environment for individuals suffering from physical or mental disabilities. By collaborating the advanced data processing methods with the monitoring principles of computer vision, the capability of activity recognition has been transformed completely (Ye et al., 2013). It is conceivable to develop a pro-active assessment model by installing visual-sensors in the ambient environment of an individual which can be used to predict and evaluate the physical afflictions in real-time.

Contribution: The goal of our study is to utilize the aforementioned deep learning capabilities to develop a comprehensive irregularity prediction framework as shown in Fig. 2.1. The contribution of the study is described by dividing the proposed framework into four objectives.

1. Primarily focused to screen the irregular physical activities of an autistic child by utilizing the principles of the computer-vision technology.
2. Deep learning based irregularity prediction methodology to calculate the physical vulnerability of the child in real-time.
3. Physical activity-based record generation using the 3D tensor technique for medical or therapeutic purposes.
4. Warning-based alert generation with the deliverance of activity information to caretakers and doctors.

Organization: The chapter is further divided into multiple sections. Section 2.2 deals with the literature survey on various activity recognition methodologies. The architectural detail of the system has been discussed in section 2.3. In section 2.4, the system is evaluated to justify the performance of the system. Lastly, the chapter is concluded with some suitable remarks in section 2.5.

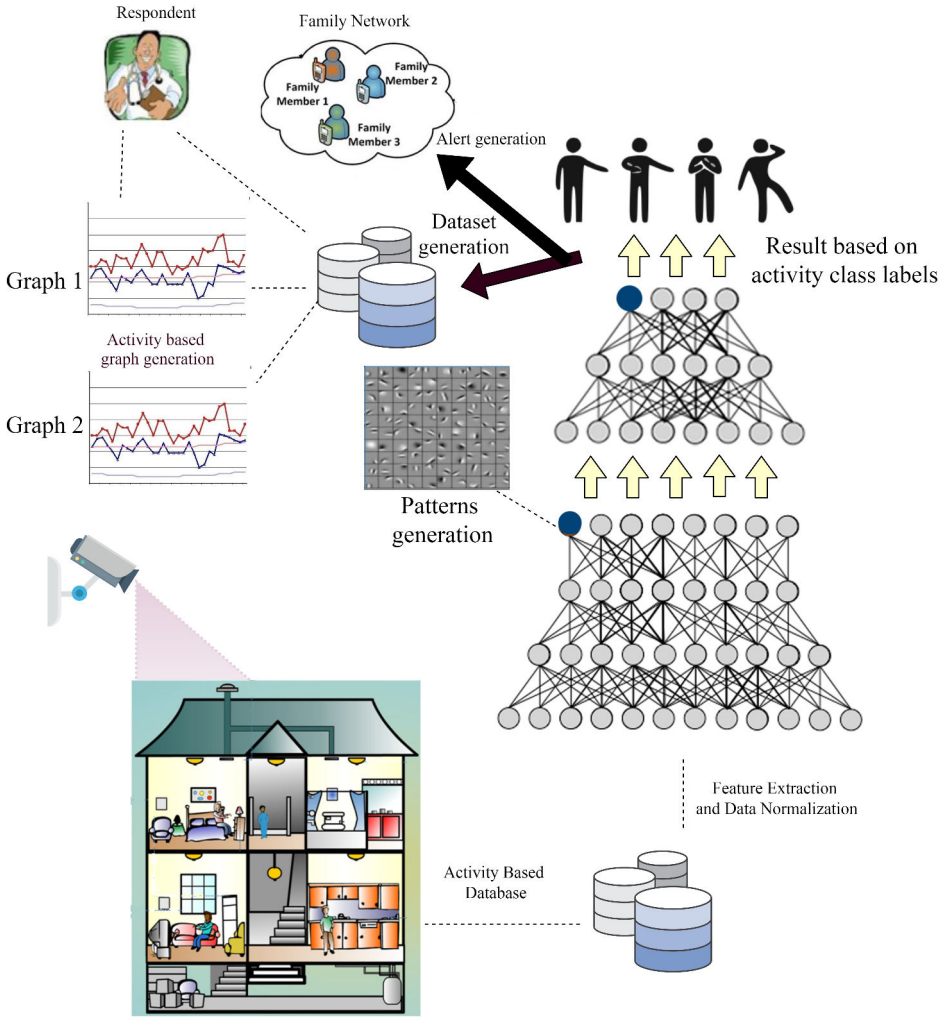

FIGURE 2.1: The conceptual framework of the proposed behavior monitoring system.

2.2 Related works

In this section, many activity recognition solutions have been discussed into two subsections: single-activity recognition solutions and multi-activity recognition solutions.

2.2.1 Single-activity recognition solutions

Noury et al., 2008 characterized the distinct investigations on fall discovery by focusing on the event location. The authors incorporated the identification of post-fall stage of a person. By differentiating the type and methods used to analyze the exercises, Perry et al., 2009 divided the fall detection systems into three classes: single acceleration-based measurement techniques (Lindemann et al., 2005), multi-acceleration based measurement strategy (Noury, 2002) and measurement techniques based on non-acceleration data (Ng et al., 2009). On the other hand, Mubashir, Shao, and Seed, 2013 divided the fall detection systems into three major groups: wearable sensors (Ghasemzadeh, Jafari, and Prabhakaran, 2010; Ghasemzadeh, Loseu, and Jafari, 2010), localized sensors (Nyan et al., 2006; Rimminen et al., 2010) and vision based fall detection system (Shi et al., 2009).

2.2.2 Multi-activity recognition solutions

Several studies have used convolutional neural network (CNN) for activity prediction and achieved better results compared to handcraft features-based solutions. Some other processing methods such as optical flow images, improved dense trajectory (IDT) features, and feature fusion are used along with CNN layers to improve the activity prediction efficiency (Yang, Molchanov, and Kautz, 2016; Xu, Yang, and Hauptmann, 2015; Chéron, Laptev, and Schmid, 2015). Many other studies have used temporal fusion-based solutions with CNN models for feature extraction from the video (Wang et al., 2017b; Feichtenhofer, Pinz, and Wildes, 2017; Sun et al., 2015). But in recent studies, 3D CNN became most popular solution for spatio-temporal feature extraction from the data for activity prediction (Tran et al., 2015).

Ye, Stevenson, and Dobson, 2015 proposed a learning-driven approach to recognize the movements on numerous subjects. The proposed framework investigated the semantics activities and calculated the semantic differences to fragment the data generated by the sensor and divided the data into several groups to differentiate the exercises. In the article (Feichtenhofer, Pinz,

and Zisserman, 2016), the authors have used 2D CNN model for spatial feature extraction from the RGB images and 3D CNN model to extract temporal features from the sequence of optical flow images. They combined these features using fusion methods to represent the final combined output. Liu et al., 2016 introduced a deep architecture to illustrate the human body based on tree-structure and revised the LSTM network to learn the links in a particular sequence. Shahroudy et al., 2016 proposed a Long Short-Term Memory module-based solution to predict physical activities. The model is trained on the variations in human joint structures to learn several body movements.

Song et al., 2017 proposed a spatiotemporal attention model to recognize the human activities. The author proposed a spatial attention module to select discriminative joints and temporal attention module for assigning weights to each frame for activity prediction. Ijjina and Chalavadi, 2017 proposed a human actions recognition approach by analyzing the temporal features of the RGB-D video. The motion-based features are obtained by using RGB and depth image-based two data modalities. Akula, Shah, and Ghosh, 2018 proposed a supervised Convolution Neural Network (CNN) architecture based action recognition system to predict six daily physical actions from IR images. Luo et al., 2018 introduced an activity recognition method to monitor the on-field activities of workers. Convolutional networks-oriented spatial-temporal streams are used to recognize activities. These features are further combined by following a fusion strategy to generate a combined prediction result. Singhal and Tripathi, 2019 introduced an action recognition solution based on local binary pattern (LBP) technique. The proposed approach focused on the LBP feature extraction via spatiotemporal relations to determine the type of action. Chen, Zou, and Zhang, 2019 proposed a network called STMP to predict activities from unconstrained scenes. The authors used 3D ConvNet to encode spatio-temporal features and performed a temporal activity localization method for activity recognition.

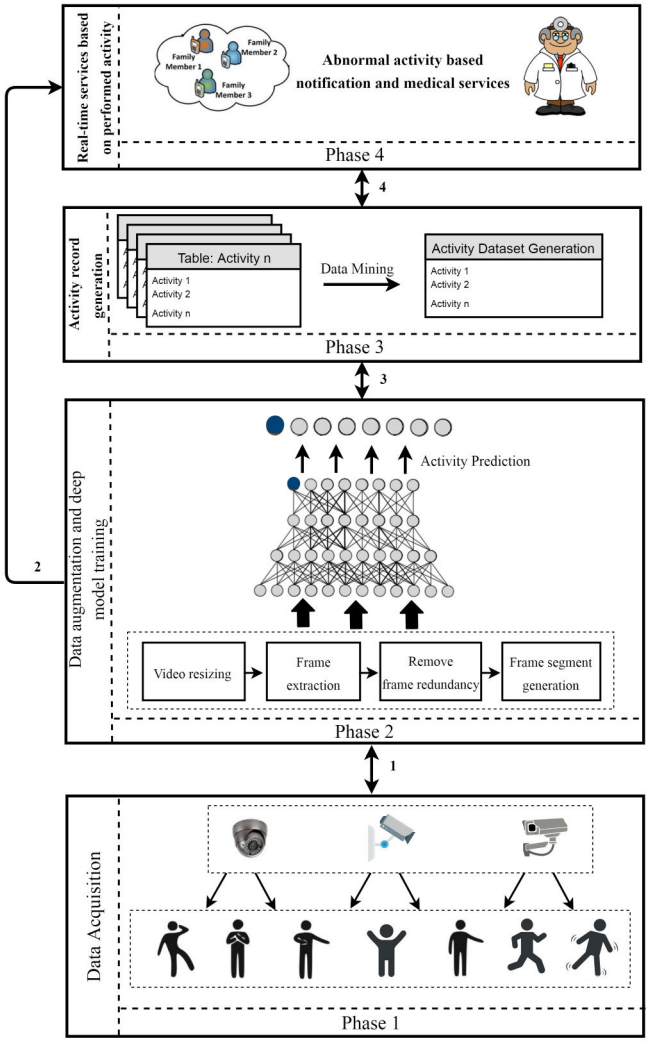

FIGURE 2.2: The layered approach of the proposed system.

2.3 Proposed Model

The novelty of the framework is explained into layered architecture as shown in Fig. 2.2. The layers of the framework performed the task of data acquisition, activity prediction, performed activity-based record generation and smart alert based decision making. The task-oriented layer synchronization process is further explained in subsections.

1. Data acquisition layer

TABLE 2.1: An overview of the dataset.

Dataset	**Technology**	**Sensor Specification**	**Activity classes**
Visual templates	Wide-range visual sensor	Pixel density: 2MP Frame rate: 30 FPS Aspect ratio: 16:9 Resolution: 1920×1080	1. Running away, 2. Pulling hairs 3. Throwing things, 4. Self punching, 5. Fighting, 6. Head beating.

2. Data augmentation and activity prediction layer
3. Activity-based record generation layer
4. Warning notification-oriented smart alert generation based decision making layer.

2.3.1 Data acquisition layer

Visual sensors are the main source of data acquisition in our study. Essentially, the data related to physical activities (regular and irregular) are constantly captured by visual sensors. After performing some pre-processing operations, the data is further transmitted to the next layer of the framework for irregularity prediction. To make an in-depth assessment, the templates must be classified into its respective activity class of the system. Table 2.1 provides an outline of the activities and technology used in this study for data generation.

It is assumed that the dynamics of abnormal activities could be identified by analyzing the series of frames. A frame sequence is sub-divided into non-overlapping frame segments called video templates. Definition 1 explains the process of video template formulation and equation 1 is used to formulate video templates with an equal number of frames.

Definition 1: (Video Template) A video template V is defined as a group of 30 frames (Average time: 1 second) into one segment described as $V_i = \{f_1, f_2, f_3, ..., f_{30}\}$.

$$(V_i, F_S, F_E) = F_N, (F_E - F_S)\ mod\ n = 0 \tag{2.1}$$

where F_S, F_E is the first and last frame in a video template. F_N is the total frames in a sequence and the value of n is set to 30 frames. $(F_E - F_S)\ mod\ n = 0$ is the division process to formulate video templates of 30 frames. If the frame segment is not divisible to 30 or the value of the remainder is not equal to 0, we pad zeros after the last frame of the segment to make it divisible.

The video templates are further compressed to reduce the template processing cost. Equation 2 is used to calculate the compression ratio of the video template.

$$\alpha_i = \frac{V_c}{V} \tag{2.2}$$

where V_c is the size of the template after compression and V is the original size of the video template. The large compression ratio defines the large size of video template which leads to a large compression rate. After finalizing template formulation and compression process, the video templates are transmitted to its second layer for activity analysis.

2.3.2 Data augmentation and activity prediction layer

Before transmitting video templates to activity prediction methodology, the system performs some data augmentation operations to provide activity prediction stability to the framework. The data augmentation operations help to overcome the constraint of overfitting which is considered as the main challenge in deep learning and occurred due to the less amount of data.

Data augmentation

Various operations like cropping, rotating, flipping can be used to translate images and generate horizontal reflections. To increase the classification stability of the system under different conditions, (Krizhevsky, Sutskever, and Hinton, 2012) alter the RGB values of the images. We performed scaling, translation, flipping, rotation, noise and a sub-sampling operation on the dataset to increase the activity prediction capability of the system.

Scaling: The scale variation on an object of interest is the most important aspect of frame diversity. In a real environment, the child can be close or far in the frame. Sometimes, a small part of the object is present in the frame or it covers the entire frame. A scale variable is applied randomly at the subject's position along the sequence of frames. The distributed number ± 0.4 is selected to scale the frames represented as $f \in \{f_{min}(0.6), f_{max}(1.4)\}$.

Translation: The translation operation is used to generate new frames with different subject position by moving the frame to X or Y direction (or both). The main purpose of applying translation operation is to increase the recognition efficiency of the system. A translation vector f_t is computed between two consecutive frames f_{t-1} and f_{t+1} to calculate the new position of the subject.

Flipping: This scenario is basically used to remove the recognition biasness of the system by applying global flipping displacement vector d = (d(x), d(y), d(z)) on the dataset. Only horizontal plane flipping is allowed by assigning $d(z) = \{0\}$, $d(x) = \{-0.5\}$ and $d(y) = \{0.5\}$ values to the displacement vectors.

Rotation: The rotation helps in maintaining the ability of subject detection. We define R rotation angles $\theta = \{\theta_1, \theta_2, \theta_3, ...\theta_i\}$ and the angle transformations $T_\theta = \{T_{\theta 1}, T_{\theta 2}, T_{\theta 3}, ...T_{\theta i}\}$ to perform rotation operation on the frames. $T_\theta R$ denotes the rotation of the sample with an angle of θ. Applying

angle transformations T_θ to all the video templates $V_i = \{F_1, F_2, ..., F_N\}$ can generate new dataset to train the model denoted as: $T_\theta = \{T_\theta V_1, T_\theta V_2, .., T_\theta V_N\}$. The original and augmented samples, i.e., $V = \{V, T_\theta V\}$, are jointly used to train the model.

Noise: We randomly select frames and apply Gaussian noise to the frame to distort the features belonging to the high frequency. The value of mean and standard deviation is always set to 0 and 1. This strategy is used to increase the robustness of the system to handle occluded frames in real-time.

Subsample operation: The size of the dataset is expanded by dividing a sequence of frames into sub-segments. The reasonable length of the segment is needed to perceive the activity. By considering the constraint of segment length selection, the number of frames F_n with the time instance $\Delta T \in \{0, t\}$ are processed to generate multiple frame segments and the first frame F_1 in the frame segment is denoted with s. The segments are generated by following the size of a time module that is preset to 1 second and the number of the frames is set to 30. The resulted subsamples are achieved by equation 3:

$$(F_{\Delta T}, s, n) = \frac{(t-s)}{n} = 0 \tag{2.3}$$

where $\frac{(t-s)}{n}$ is denoted as the value of remainder that must be equal to 0. By following the strategy of Time series based segment generation, we can generate multiple segments by processing each original sequence.

Detailed overview of proposed activity recognition methodology

To measure the irregularity scale from the physical activities of a child at particular time instance ΔT, different factors should be contemplated such as brighten conditions, camera angle, area coverage dimension and current angle of the subject towards the visual sensor. Figure 2.3 demonstrates the irregularity recognition methodology of the proposed system. In the case of sequential activity prediction in real-time, it is essential to include the output from the non-fixed previous viewpoints with the current viewpoints. To

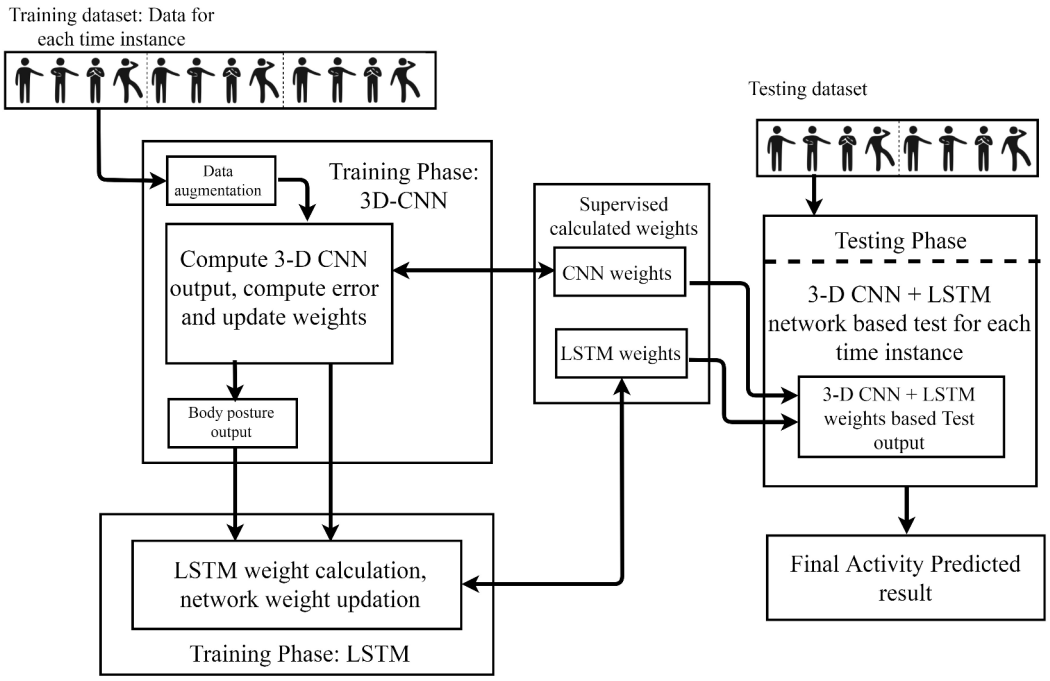

FIGURE 2.3: The proposed methodology for Abnormal Activity Recognition.

achieve this challenge, a two-stage 3D CNN-LSTM model is proposed for sequential activity analysis. 3D Convolutional Neural Network (3D CNN) is used for spatio-temporal features extraction from the preprocessed video templates (Yang, Molchanov, and Kautz, 2016). An LSTM network is used to deal with the time-based temporal feature modelling to calculate the intensity of an irregular event.

Definition 2: (Child Activity Detection) An activity A at a particular time module $\Delta T = [t_i - t_{i+j-1}]$, where the value of t is the time instance and j is the total number of activity instance. An activity A_k can be the best-recognized activity based on a particular video template V at a specific time instance ΔT.

3D CNN architecture: The 3D CNN could be used to process F_i frames of a video. The video templates V_i are used to create an input volume which is

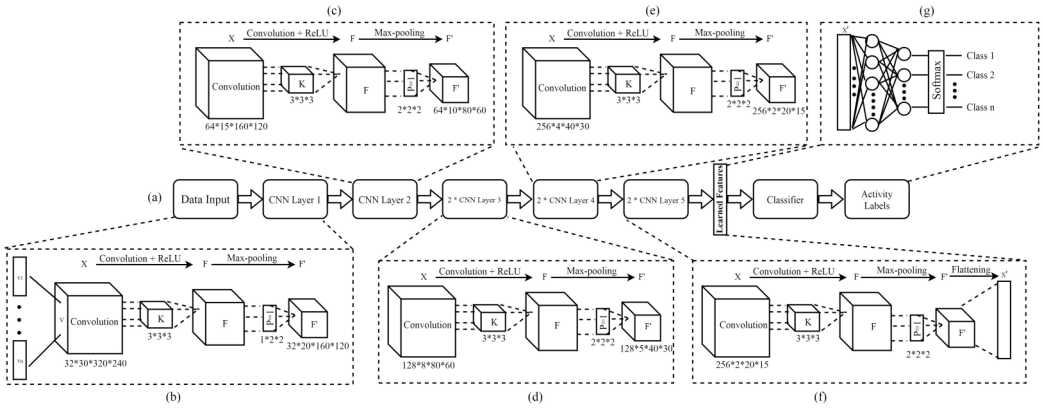

FIGURE 2.4: The proposed 3D CNN architecture for feature extraction and pose classification.

fed to the 3D CNN. The output would be a vector which represents spatio-temporal information from the sequence of frames. Figure 2.4 described the architecture of the proposed 3D CNN network. (a) The proposed 3D CNN architecture contained eight-layers of CNN. (b) The very first layer takes the input of the stack of pre-processed frames, and transfer it to the reduced feature map F'. (c) The second convolution layer takes the input from the first reduced feature map and again transfers it to the second reduced feature map. Similarly, (d) and (e) convolution layers performed the same task of feature reduction as like (b) and (c). (f) the last convolution layer reduced the feature map and performed the flatten operation to the last reduced feature map and convert it into a single x' vector. (g) The vector x' provide the input to the Fully-Connected (FC) layer by following a softmax layer. The final result vectors are generated by the convolution network and transferred to the LSTM module to analyze the dynamics of abnormality from the actions as described in figure 2.5.

Long Short-Term Memory (LSTM) model: Recurrent Neural Network (RNN) is considered as a most suitable technique to deal with the process of sequential modelling (Mikolov et al., 2010). The network contains several loops to allow the information to persist throughout the sequence. But the network usually suffers from vanishing gradient problem (Pascanu, Mikolov,

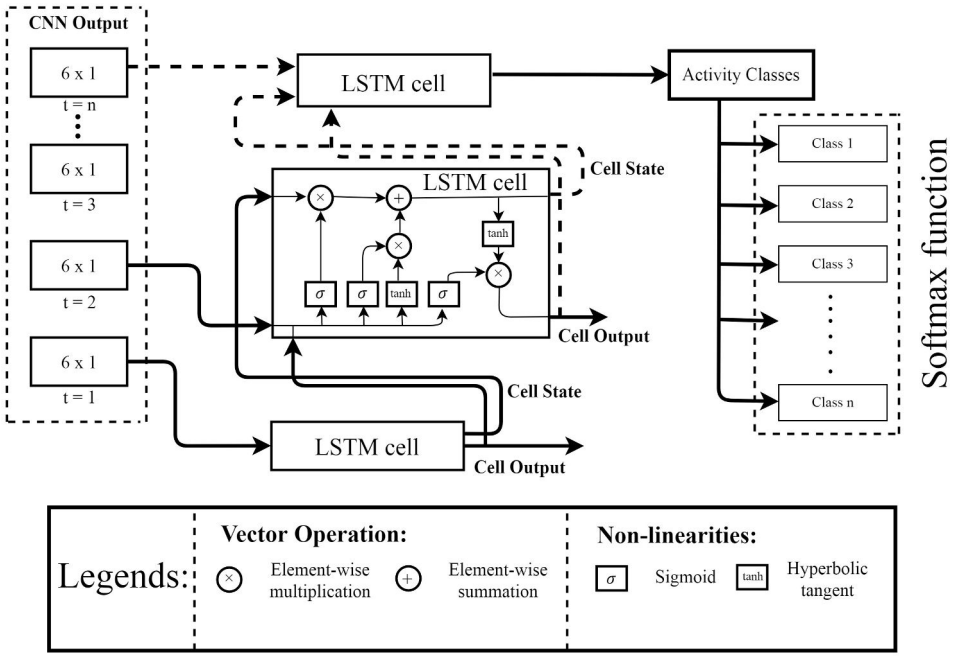

FIGURE 2.5: Activity prediction using LSTM network.

and Bengio, 2013; Graves, 2013). An LSTM is considered as the best solution to deal with the limitation of RNN technique. An LSTM cell contains several gates and memory mechanism to model the features in a sequential manner to predict an activity as described in Fig. 2.5.

An LSTM layer provides several different procedures to evaluate the hidden states of the network. Each memory cell of LSTM is designed with three types of gates where i_t represents the input gate, o_t represents the output gate and f_t represents the forget gate. Each gate performs the task of storing, recovering, and updating the information by considering the current state c_t of the cell. The non-linear activation function is applied to calculate the output of each gate by containing the value range between 0 and 1. To deal with the data based on time series, every cell deals with a distinct time instance to save and generate the output of the input value by observing the value of the gate connected to the cell. An updated memory cell state defines the output of the LSTM layer.

Suppose $v_1, v_2, v_3, ..., v_m$ is a series of output feature vectors generated by the 3D CNN model, where m defines the input length. The final vector v_t is obtained by combining all the output vectors at a particular time instance. Let h_{t-1} be the previously hidden state and c_{t-1} be the previous cell state of the network. Equations 4 to 9 described the process of dynamic feature modelling:

$$i_t = \sigma(W_i[v_t + h_{t-1}] + b_i) \tag{2.4}$$

$$f_t = \sigma(W_f[v_t + h_{t-1}] + b_f) \tag{2.5}$$

$$o_t = \sigma(W_o[v_t + h_{t-1}] + b_o) \tag{2.6}$$

$$g_t = tanh(W_g[v_t + h_{t-1}] + b_g) \tag{2.7}$$

$$c_t = (f_t * c_{t-1} + i_t * g_t) \tag{2.8}$$

$$h_t = (o_t * tanh(c_t)) \tag{2.9}$$

Non-linear activation function is symbolized with σ, element-wise product is defined by $*$ and W_i, W_o, W_f, W_g, b_i, b_f, b_o, b_g, h_{t-1}, c_t are indicating the learning parameters of LSTM network. Here, the size of the hidden layer is defined by N and the output of the network is calculated by h(t).

Hyperparameter selection and optimization: To display the pose identification result based on each video template, a series of result vectors are achieved by the 3D CNN network. The final value for each class is contained in the resulting vectors. The 3D CNN model contained a total of 8 convolutions with 5 max-pool layers. 2 FC layers are followed by 1 softmax to classify the posture of the child with respect to the trained activity classes. The first convolution input measurement is denoted as 32 $\times$ 30 $\times$ 320 $\times$

240 pixels. 32 denotes the convolution channels, 30 denotes the depth of the video template, 320×240 defines the height and width of the frame. During the training process, we applied the re-sizing process on the video templates by applying max-pooling operation for feature map reduction. The size of the kernel for all convolutions is set to $3 \times 3 \times 3$ with the stride value 1. The size of all pooling kernels is set to $2 \times 2 \times 2$. But the size of the kernel in first convolution layer is set to $1 \times 2 \times 2$. Each FC layer contains 2048 units. 3D CNNs are capable to process the spatial and temporal information within a specified receptive field size. The final result vectors generated by 3D CNN model are further transmitted to the LSTM module to classify the activity by performing sequential modelling on spatio-temporal features. The LSTM network essentially helps to represent the activity over time by learning the non-linear combination of segmental features.

Model Traning: The proposed methodology is trained into two parts. The first part deals with the training process of 3D CNN model with MLP model for features learning. Rectified Linear Unit (ReLU) function is used to make the decision related to the conversion of an input signal to the output signal. The stance-based recognition efficiency is completely dependent on the video template with its pre-described label. Total 80 epochs are set to train the 3D CNN network. After 50 epochs, the network predicts high scores for irregular video templates. Up to 58 epochs, the network keeps the scores with minimal variation. So, we applied early stopping operation after 60 epochs to save the model from overfitting.

In the second stage of the training, the MLP module is replaced with LSTM network. Dropout (Srivastava et al., 2014) and l2 regularization are used to regulate the weights of the network to reduce the overfitting of the network. But the dropout is applied only on pooling layers. To optimize the performance of the classification network, Adadelta algorithm (Zeiler, 2012) is used to optimize the learning rate significantly. Maximum 150 epochs are set to train the proposed classification model. Initially, the learning rate of

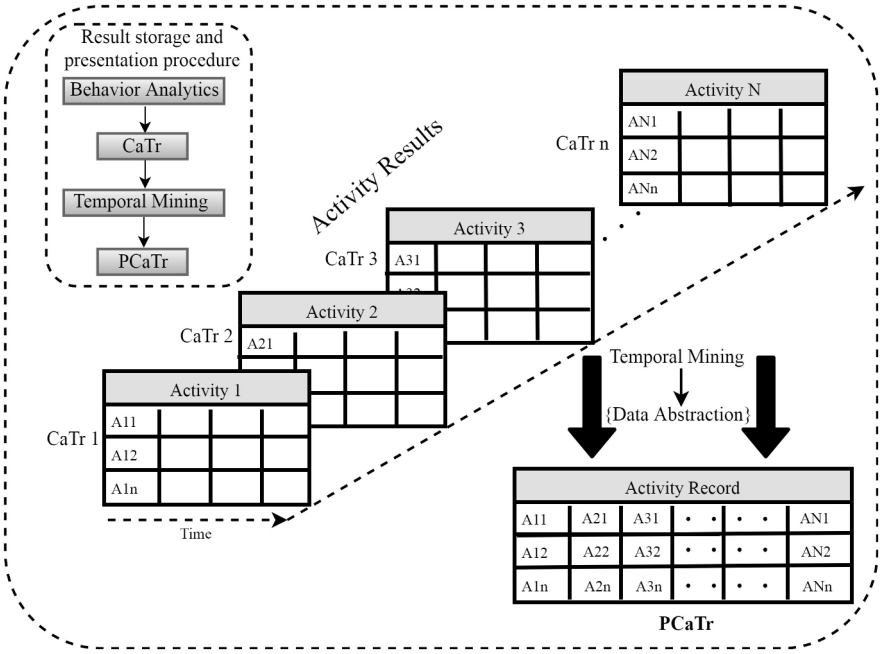

FIGURE 2.6: Activity tensor formation with activity record index generation.

the model is set to 0.001 which is gradually reduced by 0.01 when no accuracy improvement has been observed. A step decay strategy has been used for a learning rate reduction. Several trails were made to train the model by increasing the number of training steps, but every time the model got overfitted.

2.3.3 Activity-based record generation layer

In this layer, the predicted activity scores generated by the LSTM module at a particular time-stamp is stored in the local database of the system. The main challenge of this phase is to store the final activity scores of the performed activity so that effective results can be drawn in a graphical format. A 3D tensor technique is used to store the final predicted score related to the performed activity (Kolda and Bader, 2009). A 3D tensor named Child-activity Tensor (CaTr) is proposed to store the performed activity-based score.

Definition 3: (Child Activities Tensor) Given an unequal temporal activity space of $'n'$ child consists of $'m'$ type of activities. Here the value of $'n'$ always remains constant that is 1. The child's activity data tensor can be defined by a 3D tensor $CaTr\varepsilon(I_c * I_T * I_A)$ where the orders I_c, I_T and I_A correspond to the dimensions of the tensor: child ID, time and type of activity respectively such that $I_c = n$, $I_A = m$, $I_T = max(ti_{i=1}^{n})$, where t_i denotes the number of distinct time-stamps based on the activity performed by the child in a particular tie module ΔT.

In the above definition, the predicted activity score is coined as a temporal activity result where each score of activity is associated with a particular time-stamp. At different time-stamps, different scores are represented in the form of matrices. The predicted score related to each activity differs from each other. The score difference is maintained by the time stamp in which the activity has been performed. Hence, the time-stamp axis does not have an absolute value.

Mathematically, daily CaTr of single activity can be expressed as $CaTr = [A_1, A_2, A_3, ..., A_m]$. Here, A_i represents the prediction scores correspond to m activity class. Here, the matrix formulation procedure is adapted to store and presents the performed activities of the child. Each matrix consists of multiple activity scores related to a single activity class at different time modules. Hence, a combined CaTr of multiple activities is represented as: $CaTr = [AC_1[A1_1A1_2...A1_T], AC_2[A2_1A2_2...A2_T]\ , \ldots\ , AC_N[AI_1AI_2...AI_T]]$. Here $'AC'$ represents the activity class and $'A_i'$ represents the scores of the preformed activity generated by the proposed methodology. It serves the purpose of storing multiple activities in one tensor effectively which helps to generate a combined record of the performed activities.

Processed CaTr formation (PCaTr): In CaTr, all scores are collected in a structured form generated by the proposed classification model. The CaTr provides an effective solution to store the activity-based final scores, but it is not an effective way of the representation. To make tensor more efficient

to present final index, CaTr is converted to Processed CaTr with the help of temporal mining technique as shown in figure 2.6. The processed tensor is represented as processed CaTr (PCaTr).

Temporal mining: Temporal mining technique is used to retrieve the activity scores from the database of the system. The predicted movement scores are stored in the database of the system using the 3D tensor technique. Therefore, the time-series-based mining process became essential for our system. Temporal mining technique is utilized to retrieve a combined record of predicted activities and generates an activity index as shown in Fig.2.6. Algorithm 1 is used to generate the combined record of the performed activities.

Algorithm 1: Activity score recording for medical and therapeutic purposes
Input: *Child physical activity score.* **Output**: *Current child state with relative activity class.*
Step 1: Determine the current predicted activity score calculated by the proposed activity classification phase.
Step2: If (Child_State(i)) lies in the trained classes of the system, goto step 3 else goto step 4.
Step 3: do **Step 3.1**: For the calculated activity score by phase 2, CaTr.add (child ID, current time stamp, activity type); Return CaTr; End for
End if
Step 4: Exit.

2.3.4 Warning-based smart alert generation for smart decision making

In the healthcare and assistive-care domain, warning and emergency signal play an imperative role to notify caretakers or doctors about the current situation of an individual under monitoring. The monitoring process is generally divided into two parts: Continuous monitoring and Alert-based monitoring. In our study, we have combined both monitoring services to make the proposed system more efficient. Continuous monitoring transmits data to the proposed classification methodology in a continuous manner to analyze the performed activity. Alert-based monitoring generates an alarm or notification to doctors and caretakers when an individual is performing any irregular activity. By using these two monitoring procedures, a combined approach is proposed to consolidate the presentation of irregular activities with alert generation.

Notification Generation: The last layer of the proposed system deals with the notification generation services for the custodians. The mechanism of alert generation is completely dependent on the recognition efficiency of the model that is further analyzed by calculating the false positive ratio parameter. Moreover, the proposed irregularity prediction method is less invasive for the child and helps the parent to deal with the current physical state of the child in real-time. Lastly, the system helps the doctor or parental figures to deal with the physical irregularity of the child at the initial stage with the goal of deploying early safety measures. Algorithm 2 described the procedure of determining the child physical state and handle the critical situation by adopting the proposed alarm generation mechanism.

Algorithm 2 is responsible to generate alert-based decisions based on the physical state of the child in real-time. Two physical state measures are calculated to determine the physical state: safe or unsafe. The safe state determines that there is no compelling reason to ascertain any estimation. On the other hand, an unsafe state determines the irregular physical condition

of the child and force the caretaker or doctor to respond. As described in the algorithm 2, if the child's current state lies in any of the activity class, the trigger is activated and the respondent is notified with the early cautioning signal.

Algorithm 2: Child's current activity state determination and notification generation
Input: *Calculated value of the current frame segment at time t.* **Output**: *Child's physical position with alarm generation.*
Step 1: Calculate the current activity status of the child at time instance t using the proposed activity classification methodology.
Step2: If (Child_State) lies in abnormal activity class of the model, goto step 3 else goto step 4.
Step 3: Event Trigger = True **Step 3.1**: Generate warning signal to family members or caretakers in real-time with physical status and transfer the temporal attribute from phase 2 to doctors for handling medical emergencies if the critical situation happens.
Step 4: Repeat Step 1 after a definite time interval based on the performed activity.
Step 5: Exit.

2.4 System implementation and experimental evaluation

The proposed system is enabled with Intel Core i5-8600 Processor (@ 4.30 GHz) with NVIDIA Geforce GTX 980 Ti 6GB Graphical Processing Unit (GPU). The software configuration of the system is as follows: Operating system (Ubuntu (14.04)), Programming language (Python (2.7.11)), and Integrated Development Environment (IDE) (Spyder (3.1.4)). MySQL database system is also utilized to maintain the database for activity score storage. The performance of the system is evaluated in four sub-sections:

1. Data acquisition process,
2. Impact of augmentation techniques on the performance of the system,
3. Evaluation of irregularity prediction efficiency,
4. Statistical measurement of false-positive ratio,

2.4.1 Data acquisition process

By considering the privacy issue, we decided to simulate the proposed system on local machine. We have selected Self-Stimulatory Behaviour Dataset (SSBD) dataset (Rajagopalan, Dhall, and Goecke, 2013) to evaluate the performance of the proposed methodology. It comprises of untrimmed recordings of 6 irregular physical activities, such as falling, pulling hairs, self-punching, head banging, fighting, and throwing things. Variations in the sequence of frames defined temporal diversity in the activities performed by the children. The selected physical irregularities significantly affect the physical safety of the child.

Manual dataset formulation:- To ensure the irregularity classification stability of the system, a complex dataset has been formulated by dividing the dataset into training, validation and testing set. The complexity of test dataset is verified by selecting video templates which are not used for training and are also related to different children.

The original dataset consists of 5,579 video templates and further generate 15,892 templates by applying several data augmentation techniques. The dataset is divided into training and testing set by applying 70:30, 75:25, and 80:20 ratios. The training set is further divided into two parts by applying a 70:30 ratio to formulate validation set. Details of video templates in training, validation and testing set are listed in Table 2.2.

TABLE 2.2: Detail of training, validation and testing set.

Dataset type	**70:30**	**75:25**	**80:20**
Original Dataset	2735 + 1171 : 1673	2930 + 1255 : 1394	3125 + 1339 : 1115
Augmented Dataset	7788 + 3338 : 4767	8344 + 3575 : 3973	8900 + 3814 : 3178

2.4.2 Ratio-based performance analysis:

The best classification performance has been observed on the 80:20 ratio as described in Table 2.3. According to the 80:20 ratio on original dataset, the system is trained on 3125 video templates and 1339 templates are used to validate the system to finalize the weights of the classification methodology. The performance of the system is evaluated using 1115 video templates of the test dataset. The same process is applied to the augmented dataset. Table 2.3 demonstrate the effect of data augmentation techniques on the recognition performance and the calculated results are obtained from the 80:20 ratio.

TABLE 2.3: Performance comparison of original dataset and augmented dataset with each data augmentation technique.

Augmentation Methods	**Original Dataset**	**Augmented Dataset**
Without augmentation	86.26 %	—
Scaling	—	87.52 %
Translation	—	87.95 %
Flipping	—	89.56 %
Noise addition	—	90.25 %
Rotation	—	91.89 %
Sub-sampling	—	90.51 %
All combined operation	—	92.89 %

The effect of every data augmentation technique on the final recognition performance of the model is demonstrated in Table 2.3. It can be observed that scaling, transformation, flipping, rotation and noise operation improved the recognition stability of the model. On the other hand, the sub-sampling

operation reduced the accuracy of the model. This penalty is due to quality reduction and the size of the training dataset introduced by this process. But sub-sampling with noise, scale, translation, rotation and flipping operation helped in increasing the variability in the dataset by increasing the size of training data. It has outperformed all the previously calculated results.

2.4.3 Evaluation of the system performance for irregularity prediction

This section evaluates the performance of the system by analyzing several irregular events. The implementation environment is selected inspired by the environment of the house. As specified before, the proposed system is divided into distinctive stages. The exploratory execution is performed to perceive the two fundamental objectives and calculate results to describe the overall applicability and utility of the system in the assistive-care domain.

- Statistically regulate the activity recognition efficiency of the proposed system to provide real-time assessment services.
- Justification of the proposed activity recognition methodology by comparing the activity classification performance with the other state-of-the-art methodologies.

Activity prediction efficiency

The system efficiency for activity recognition is concerned with the analysis of the current physical state of the child. Two most relevant deep learning methods named 3D CNN and LSTM with data augmentation techniques are incorporated to train and test the prediction efficiency of the system. Fig. 2.7 presents the qualitative approach of the system. Plots in graphs (a) - (c), (e) and (f) shows the video templates with irregular events. The proposed classification methodology successfully detect the irregularities by producing high probability scores for each frame with the irregular stance. (d) represents no irregular event by producing minimal probability score (near to null) when

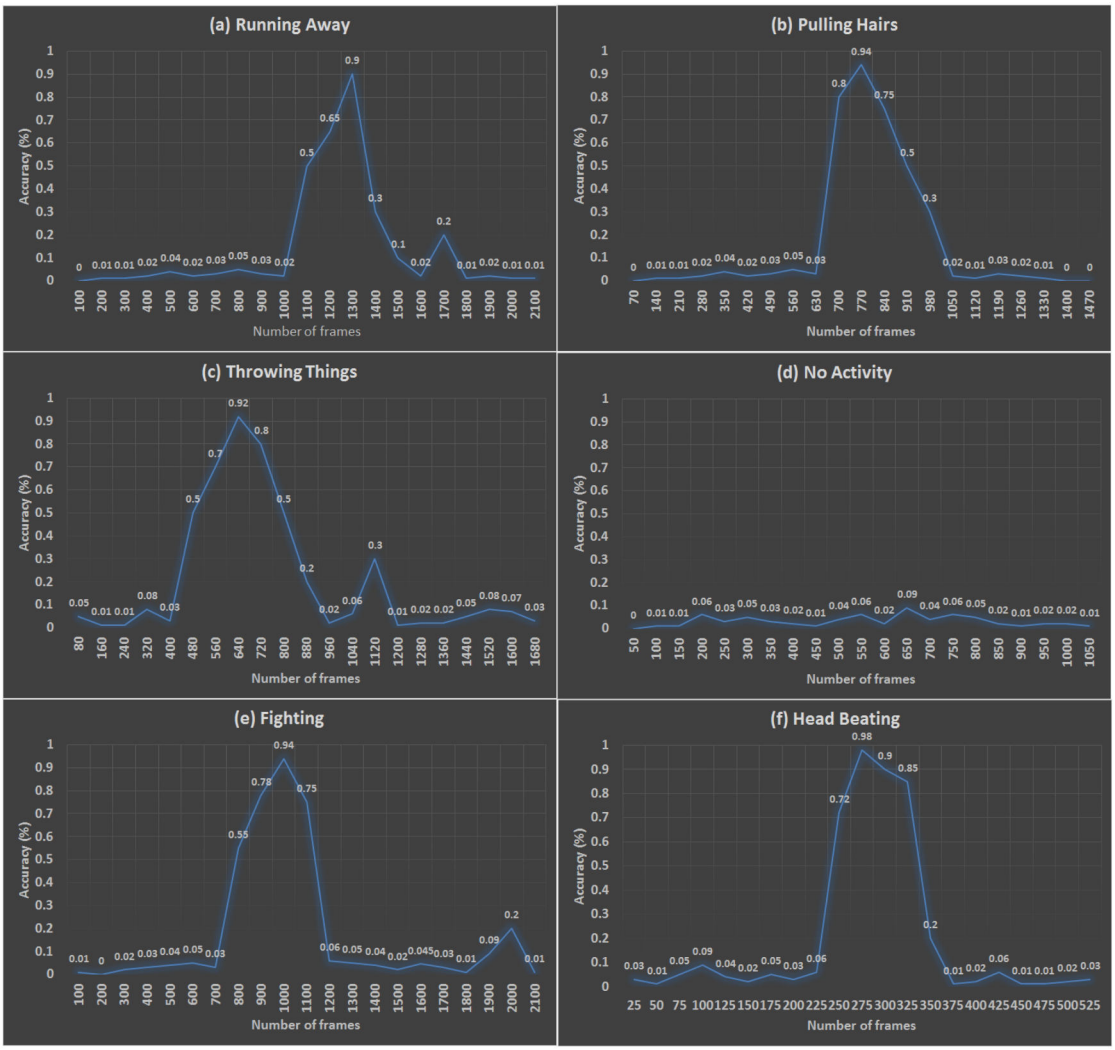

FIGURE 2.7: Qualitative results of the proposed methodology for testing videos. (a), (b), (c), (e) and (f) contained severe conditions with accurate detection. (d) Shows no event happening in this video template.

no irregularity happens.

Several statistical parameters, such as accuracy, precision, recall, specificity, and f-measure are calculated based on the recognition results to justify the efficiency of the proposed system as shown in table 2.4.

Accuracy defines the closeness of the measured value to its actual value

TABLE 2.4: Activity class oriented system performance analysis.

Irregular Activities	**Accuracy**	**Precision**	**Recall**	**F-measure**	**Specificity**
A. Running away	90.23%	92.08%	87.52%	89.15%	87.22%
B. Pulling hairs	94.37%	93.32%	90.31%	92.28%	89.03%
C. Throwing things	94.02%	95.35%	92.47%	93.22%	92.17%
D. Self punching	85.12%	86.64%	83.42%	83.17%	82.27%
E. Fighting	98.49%	99.26%	93.27%	95.78%	92.42%
F. Head beating	95.22%	95.88%	92.21%	93.89%	91.39%
Mean accuracy	**92.89%**	**93.75%**	**89.86%**	**91.76%**	**89.22%**

and our proposed system achieve the overall accuracy of **92.89%**. **Precision** measurement is used to calculate the exactness of the system. The system provides a higher precision of **93.75%**. **Recall** is a part of applicable events that is recovered from the total number of significant events. The proposed model generates a higher recall of **89.86%**. The proposed methodology also contribute higher value of **F-measure** and **Specificity** by achieving **91.76%** and **89.22%** respectively.

Confusion matrix: In the testing phase of the system, the precision is figured by detecting the activities based on the predetermined time instance. The observed outcomes demonstrate the normal execution of the proposed technique. The average class-based activity prediction accuracy of 92.89% is given in Fig. 2.8.

The confusion matrix (Fig. 2.8) explained that the proposed approach

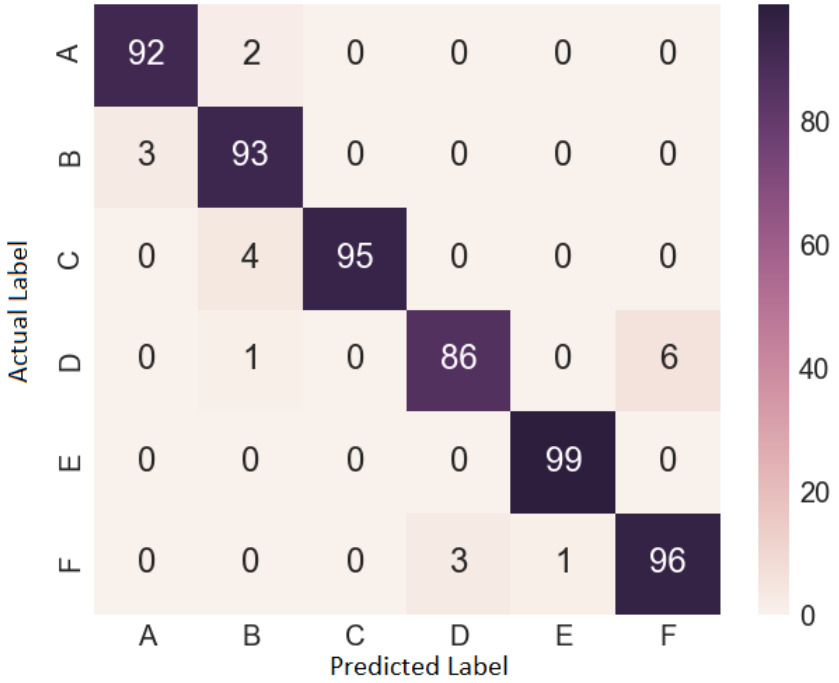

FIGURE 2.8: Confusion matrix of abnormality recognition.

has effectively classified all the activities. Especially two activities "Self-punching" and "Head beating" has less physical exploration and the proposed methodology is also successfully classified these two activities with better prediction margins. However, the system still needs improvement in the prediction of "Self-punching" activity. Less number of video templates in the training phase can be a reason of miss-classification which can be resolved by increasing the number of video templates of that particular activity in the training process of the system.

State-of-the-art comparison result

Different state-of-the-art methodologies are considered to compare the performance of the proposed framework. During the implementation of the state-of-the-art methodologies, it is essential to mention that only the classification methodology is changed without modifying or adjusting the system parameters.

TABLE 2.5: Activity recognition performance results of different state-of-the-art classifiers

Method	Accuracy	Precision	Recall	F-measure	Specificity
Song et al., 2017	83.71%	84.41%	81.85%	83.38%	82.10%
Shahroudy et al., 2016	84.58%	85.62%	84.18%	84.82%	87.63%
Liu et al., 2016	88.75%	89.51%	86.56%	87.82%	86.28%
Proposed Model	**92.89%**	**93.75%**	**89.86%**	**91.76%**	**89.22%**

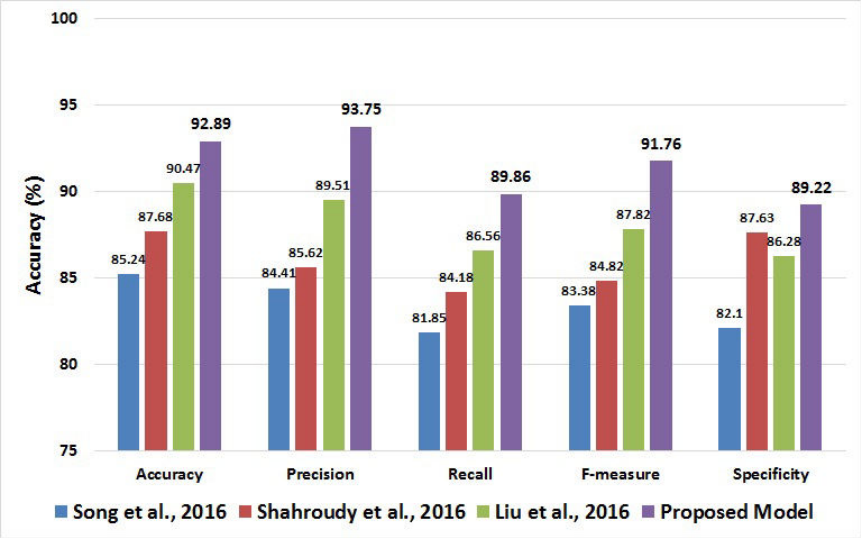

FIGURE 2.9: State-of-the-art comparison.

Table 2.5 and Fig. 2.9 shows the experimental results on the dataset. Based on the results drawn from testing dataset, we can conclude that the proposed irregularity prediction methodology is far more efficient than other state-of-the-art methodologies in terms of statistical evaluation. The proposed prediction methodology outperformed the other selected approaches namely, (Song et al., 2017; Shahroudy et al., 2016), and (Liu et al., 2016) on average accuracy. Our method achieved 4.14% higher recognition accuracy than the competitive state-of-the-art prediction methodology, (Liu et al.,

2016). For specificity, we have registered 89.22% for the proposed methodology which is higher than the other state-of-the-art prediction models taken into consideration. Moreover, in the case of precision, recall and f-measure, the proposed methodology outperforms other classifier models with 93.75%, 89.86%, and 91.76% accuracy respectively.

2.4.4 Statistical measurement of false-positive ratio for alert generation:

A child with irregular physical activity requires an immediate assessment. In this case, physical state prediction and information deliverance time are considered as the important parameters in an alert generation. In the proposed system, the efficiency of an alert generation mechanism is evaluated based on the time gap in the irregularity prediction and information deliverance. The information deliverance delay can be calculated as:

$$Delay = T_{Prediction} - T_{Deliverance(V_i,\Delta T)} \tag{2.10}$$

where $T_{Prediction}$ is the event prediction time and $T_{Deliverance(V_i,\Delta T)}$ is the deliverance time at which state information based on the current video template V_i in a particular time module ΔT is delivered to the caregiver.

This section also determines the total number of "false positive" alerts based on the true alerts generated in the testing phase of the system. The above-calculated performance evaluation parameters are used to evaluate the ratio of the alerts generated to the concerned specialists or custodians. The false positive ratio is calculated by:

$$false\ positive\ ratio = \frac{Total\ number\ of\ false\ positive\ events}{Total\ number\ of\ negative\ events} \tag{2.11}$$

Table 2.6 explained that only 2.79% of the alerts are covered under the examination of false positive events. The parameters like Mean Absolute

TABLE 2.6: Statistical descriptors of the proposed system.

Sr. No.	Parameters	Accuracy
1.	Accuracy	92.89%
2.	Precision	93.75%
3.	Recall	89.86%
4.	F-measure	91.76%
5.	Specificity	89.22%
6.	Mean Absolute Error	3.15%
7.	Relative Absolute Error	7.80%
8.	False Positive Ratio	2.79%

Error and Relative Absolute Error performance parameter also justify the accuracy of alert generation based on true activity prediction results.

2.5 Conclusion

This chapter presented the development of an irregularity prediction model to evaluate the physical activities of autistic children. The proposed 3D CNN model performed the operation of spatio-temporal feature extraction and LSTM model is used to perform temporal time-sensitive sequential feature modelling to calculate the scale of irregularity for further assessment. The 3D tensor technique is used to store the predicted activity results in the database which can be further used for medical or therapeutic purposes. Temporal mining technique is used to extract the values from the database related to the requested time module for the overall performance calculation of the child. To maintain the domain sensitiveness, an alert generation mechanism is proposed to generate warning signals by transmitting the current physical state of the child to the caretaker and concerned doctor. The calculated outcomes prove the proficiency of the proposed system in the assessment domain. Hence, it can be concluded that the proposed system is effective and well efficient for irregularity prediction and it shows the considerable potential for use in healthcare and assistive-care domain.

Chapter 3

Stance Monitoring for GAD Detection

3.1 Introduction

Continuous development in smart healthcare became a strong reason for the high expectations of a healthy lifestyle. But ever-increasing workload became a primary reason for poor health and high stress. Several surveys depicting that the person with the heavy workload is suffering from several physical and mental issues. According to the American Psychological Association (APA), individuals suffering from mental illness experiencing constant stress and anxiety over standard exercises or scheduled events. The mental illness can cause several physical side-effects, such as a migraine, inconvenience dozing, stomach upset, headache, neck pain and many others.

In a report conducted by World Employment and Social Outlook, more than 59% of the population is working and 33.33% of the population is working for more than 48 hours in a week (Office, 2015). Moreover, 51.9% of cases were registered with severe mental and health issues (Yu et al., 2013). Analysis of these statistics suggested that longer working hours can pose severe health issues causing mental imbalance and physical abnormalities to body. By analyzing these adverse effects caused by this under-considered problem, there is a strong need for a system to monitor the physical state of an individual for early prediction of physical abnormality to handle health adversities.

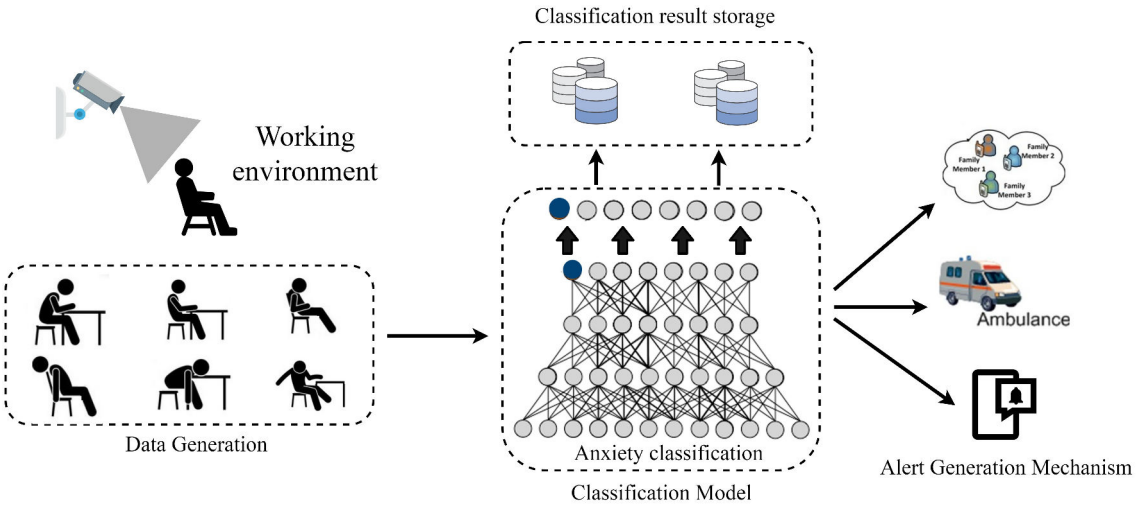

FIGURE 3.1: Anxiety prediction process.

In recent couple of years, the continuous advancement in computer vision and deep learning technology has provided a major contribution in several monitoring applications and services (Buch, Velastin, and Orwell, 2011; Gao et al., 2018). Improved technology such as high definition (HD) visual sensors and advanced deep learning methods can provide superior data acquisition and processing capability for real-time monitoring (Mabrouk and Zagrouba, 2018). By utilizing the advancements of computer vision and deep learning, an advanced dimension of intelligent monitoring can be added to healthcare or assistive-care domain.

Novelty Aspect: By considering the advancement of computer vision (Dawn and Shaikh, 2016) and the data processing efficiency of deep learning (Hinton and Salakhutdinov, 2006), a novel activity prediction system (Fig.**??**) is proposed to predict the scale of physical abnormality related to Generalized Anxiety Disorder (GAD). The primary findings of the proposed study are:

1. To design a computer vision assisted smart activity monitoring system to analyze the scale of physical abnormality.

2. To provide deep learning facility for abnormality prediction from the physical activities of an individual in an indoor environment.
3. To generate activity-based records for analyzing the physical state of an individual for medical and therapeutic purposes.
4. To maintain the sensitivity of the healthcare domain by proposing an alert generation mechanism for the doctors and caretakers.

Chapter structure: The chapter is arranged into following sections. Section 3.2 provides an overview of imperative works in the field of computer vision-based activity monitoring. In section 3.3, several strategies related to the proposed system have been discussed during implementation. Section 3.4 discussed the utility of the system by analyzing the performance of the system for different cases. The study has been concluded in section 3.5.

3.2 Related works

An overview of some essential methodologies in the field of activity monitoring has been discussed in this section. The literature is divided into two categories: handcraft features based activity recognition and deep learning based activity recognition.

3.2.1 Handcraft features based activity recognition:

Over a recent couple of years, research of Human Activity Recognition (HAR) is predominantly focused on utilizing activity-based video recordings for activity analysis. Using handcraft techniques, the majority of methodologies are dealing with a common viewpoint for activity prediction. Li, Camps, and Sznaier, 2012 represented the dynamics of tracklets to predict actions. The proposed method recognized the actions by dealing with the Hankelets which are invariant to viewpoint changes. Li and Zickler, 2012 connected the initial and final view by assuming a virtual path in these views. They consistently inspected a number of points with the virtual path. In this manner, they used a linear transformation function to consider each point in the form

of a virtual view. Zhang et al., 2013 enhanced the same methodology by using infinite linear transformation function. Despite the fact that these strategies can work without feature-to-feature correspondence among the initial and final view, they required the training samples of the final view. Zheng and Jiang, 2013 proposed a technique for building a transferable dictionary. The proposed technique basically paired the same action based videos of different viewpoints to calculate the sparse coefficients. But, the proposed technique expected video or frame level based feature-to-feature correspondence in the training process, thereby restricting the applications. Wang et al., 2014 proposed a cross-view activity recognition methodology by finding 3D Poselets and learned the geometric relations among various viewpoints. They used linear SVM solver to learn different transformations between various views. For activity recognition, every learned transformation is utilized to make coordination and the AND-OR Graph (AOG) technique is used to join the outcomes. Gupta et al., 2014 proposed a Non-linear Circular Temporary Encoding method for locating the best match in long mocap succession. In article (Cai et al., 2018a), the authors used a hidden conditional random field (HCRF) and an improved sparse Gaussian process latent variable model (GPLVM) for action prediction. They extracted the features from the sequence of actions by fusing the skeletal information and the human body motion characteristics. The improved sparse GPLVM algorithm reduced feature dimensions to handle computation cost with better visualization.

3.2.2 Deep learning based activity recognition:

Several classification models have been proposed based on the singular or combinational techniques of deep learning (Simonyan and Zisserman, 2014a; LeCun, Bengio, and Hinton, 2015). Donahue et al., 2015 introduced a first CNN + LSTM (LRCN) method based combined architecture to recognize the physical activities of a person from the RGB visual templates. In article (Yue-Hei Ng et al., 2015), the authors proposed an LSTM based activity classification method to model temporal features generated by the CNN model from the sequence of frames to predict the type of activity. Veeriah, Zhuang, and

Qi, 2015 proposed an activity classification and recognition model by utilizing a differential gating scheme for an LSTM. After performing normalization techniques on the data, the system utilized 3D coordinates of skeleton map as input to classify the activities. Wang, Qiao, and Tang, 2015 utilized deep convolutional technique to learn discriminative features. They aggregated extracted features using trajectory-based pooling layers.

In article (Du, Wang, and Wang, 2015), authors introduced a Tanh Bidirectional Recurrent Neural Networks (Tanh-BRNN) based architecture which converged in an LSTM. The architecture contained a single layer of fully-connected cells to decompose the input to generate output. Neverova et al., 2016a proposed a system to identify the physical movements from the temporal patterns generated by the gyroscope and accelerometer sensors of smartphones. The system utilized the combination of Convolutional Neural Network and Recurrent Neural Network to train and classify human identities with their physical movements. In article (Neverova et al., 2016b), authors exhibited a strategy to identify the gestures and confinements based on multi-modular deep learning systems. Liu et al., 2016 proposed an architecture to represent the skeleton in the form of a tree structure. Furthermore, a modified LSTM architecture is proposed to visit each skeleton coordinate in a sequence to identify the physical movements.

Wang, Farhadi, and Gupta, 2016 proposed a Siamese network for action modelling. They represent actions as transformations by utilizing high-level feature space. Transformations demonstrate the precondition and after-condition state of an action to predict the type of activity. Li et al., 2016 proposed an architecture to calculate the initial and the last point of activity by utilizing three layers of LSTM. The architecture follows a fully connected layer to perform the classification task. Wang et al., 2018 recognized the activities by proposing a model based on CNNs and trajectory maps. In article (Carreira and Zisserman, 2017), the authors used 3D convolutions to deal with a spatiotemporal relationship in videos. The deep architecture effectively shows the appearance but failed to demonstrate long-term motions.

In article (Feichtenhofer, Pinz, and Wildes, 2017) and (Wang et al., 2017b), the researchers develop a temporal fusion methodology for spatiotemporal feature extraction from the videos using basic CNN technique. Krishna et al., 2017 identify events from the video by utilizing attention model-based language model and data analysis protocols (DAPs). The proposed model also described the predicted events in a natural language simultaneously.

In article (Das et al., 2018), the authors utilized LSTMs to model human dynamics for activity prediction from a video by calculating geometry features of the subject. They had effectively encoded long-term motion dynamics which is an imperative perspective for perceiving exercises. In article (Cai et al., 2018b), authors utilized improved convolutional neural networks (CNN) to predict human activities. They extracted depth sequence features by utilizing depth motion maps from the frame segments to calculate the pose variability. Akula, Shah, and Ghosh, 2018 have proposed an automated 2D-CNN architecture to predict ADL activities of an individual for assistive care. The authors used infrared imaging modality based dataset to predict terminal actions like falling or standing. Luo et al., 2018 introduced an activity prediction system to predict abnormal activities of workers in a continuous manner. The authors used two-stream convolutional neural network to encode spatio-temporal features from the given input video and fusion technique is proposed to generate a combined prediction result.

3.3 Proposed Model

The fundamental aim of the proposed system is to predict several physical abnormalities of an individual in an indoor environment and provide the required healthcare services on time. Figure **??** demonstrates the modular approach of the proposed system. The overall system comprises of four modules, namely Data Acquisition, Anxiety Classification, Activity-based Record Generation and Alert-based Decision Making. Each of these modules perform a pre-specified task and provide required services to its adjacent modules. Every parameter used in the proposed system is detailed

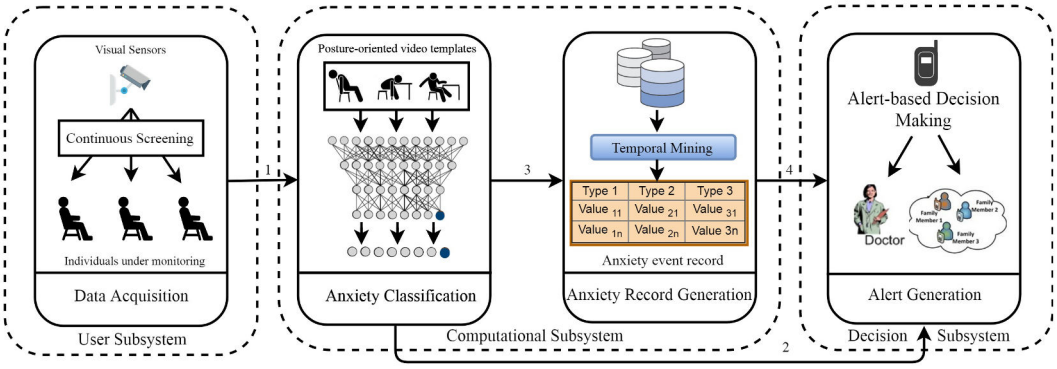

FIGURE 3.2: Proposed Architecture of Anxiety Prediction System.

ahead:

3.3.1 Visual Sensors Embedded Environment:

Visual sensors monitor physical activities of an individual and the captured sequence of frames serve as a source of data in the proposed study. Usually, there are two forms of activities namely, regular activities and irregular activities. Activities of Daily Living (ADL) such as standing, walking, sitting, laying and many others are considered as regular activities. On the other hand, falling, unconsciousness, nausea, staggering are some of the examples of irregular activities. This phase performs the initial task of data acquisition by analyzing the physical activities of a person in their working environment. Table **??** provides an overview of sensor specification and the activities which are used to determine the scale of Generalized Anxiety Disorder (GAD).

Data Pre-processing: Before transferring the data to its next module, the first module performs some pre-processing operations. The visual sensor node formulates the frame segments by combining the fixed number of frames into a group. These frame segments are further compressed by performing a down-sampling operation which helps to reduce the data processing cost. The activity prediction efficiency is directly proportional to the quality of

TABLE 3.1: Description of Dataset.

Type of data	**Data capturing technology**	**Sensor specification**	**Type of activities**
Physical Postures	High-Definition Visual Sensors	Sensor density = 2 mp Aspect ratio = 16:9 Resolution = 1920×1080 Frame rate = 30 fps	1. Staggering 2. Headache 3. Stomachache 4. Backache 5. Neckache 6. Nausea

data. The prediction efficiency of the system is tested on multiple ratios. The compression ratio of a frame segment F_i is calculated as: $\alpha = \frac{F_c}{F}$, where F_c denotes the size of compressed frame segment and F denotes the size of original frame segment.

To select the best compression ratio without sacrificing the efficiency of activity prediction, Quantization Parameter (Chen et al., 2017) is utilized for data quality measurement. The high value of Quantization Parameter defines the low quality of the video and vice versa. The quantization parameter (q) is fitted with respect to the compression rate as follows:

$$q(r_s) = \frac{1}{c_2} \log \left(\frac{r_s}{c_1} \right) \tag{3.1}$$

where $q(r_s)$ denotes the Quantization Parameter correlated to the compression rate r_s, $c_2 \leq 0$, and $c_1 \geq 0$.

3.3.2 Anxiety Classification Methodology

The process of abnormality prediction is presented in figure **??**. The 3D Convolutional Neural Network (3D CNN) and Gated Recurrent Unit (GRU) based combined approach is proposed to predict the physical adversity. The stance-based actions are predicted by extracting spatial and temporal features from the frame segments using 3D CNN model. The stance variability

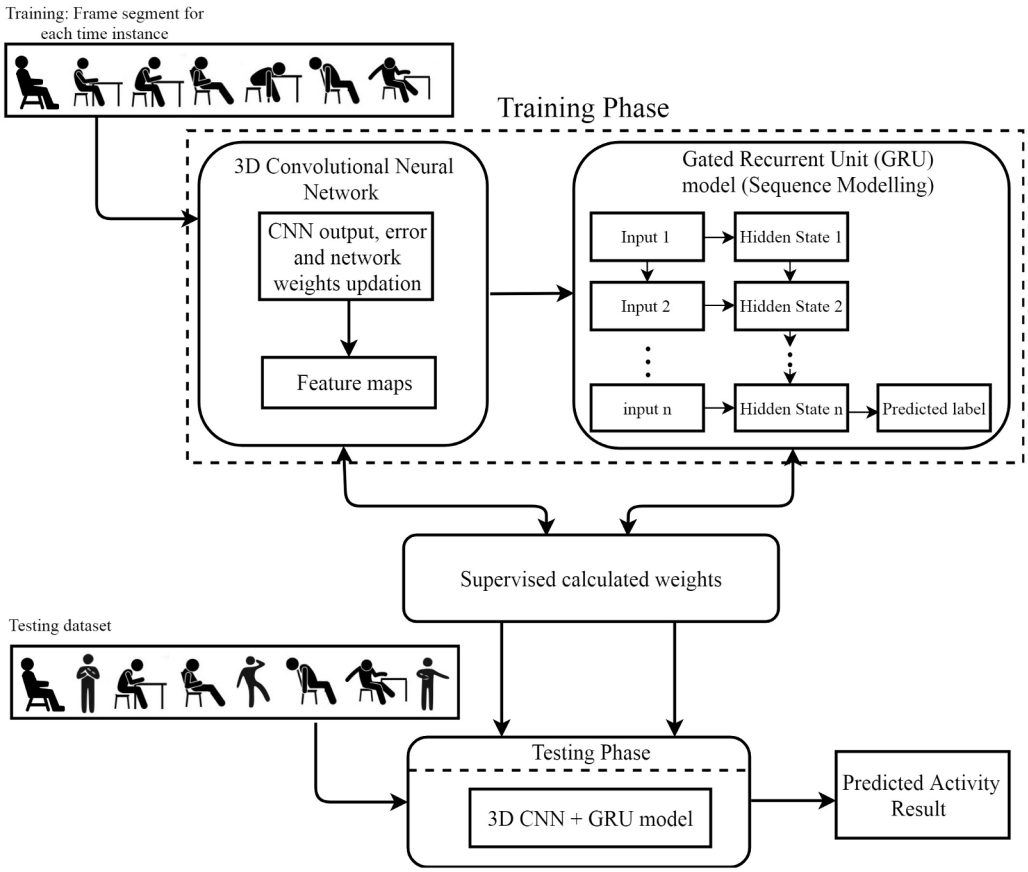

FIGURE 3.3: Abnormality prediction and classification.

is determined by calculating the temporal relations from the extracted features of the input frame segments using the GRU model.

Definition 1: (Activity Prediction) An irregular activity IA_j related to current frame segment FS_i can be the best predicted abnormality at a particular time module ΔT. A time module ΔT is represented as $\Delta T = |t_S - t_E|$, Where t_S denotes the initial time instance and t_E denotes the last time instance of the time module.

Stance classification

The stance prediction model extracts spatial and temporal features from the input frame segments to determine the action of the person. To train the data, a well known pre-trained VGG-16 (Simonyan and Zisserman, 2014b) model is used. Among the different CNN architectures, the VGG-16 Net considered as the most basic structure of the CNN model. It is favorable to implement VGG-16 network with some structural changes and can accomplish reasonable action prediction stability between the prediction exactness and the prediction time.

Fig. **??** shows the architecture of VGG-16 Net and explained the parameters of the model. The kernel size of all convolutional layers of VGG-16 net is set to 3×3 with stride value 1 which convolve each pixel of an image and helps to reduce the feature parameters. Less value of stride helps to prevent the important patterns of an image. Two consecutive convolutional layer with the kernel size of 3×3 is also applied in the network without applying a pooling layer. The results of the combination of two consecutive layers provide the actuality of 7×7 kernels. Non-linear function "relu" is used to assemble three convolutional layers consecutively, where non-linear activation functions help to make the features more discriminative. To handle the problem of overfitting, a dropout layer is added between the two FC layers of the model by setting the ratio value of 0.7. The model is trained on 0.0001 learning rate. The value of weight decay is set to 0.0005 with the momentum size of 0.9. The batch size of the input is set to 16. The action classification ability of the network does not diminish as compared to the other CNN models. The features generated by the VGG-16 model is further utilized by Gated Recurrent Unit (GRU) model for context learning.

Context Learning

The scale of an abnormality is determined by calculating the high-dimensional temporal features in a sequential order. Therefore, an optimized Gated Recurrent Units (GRU) model is used to calculate the dynamics of actions from

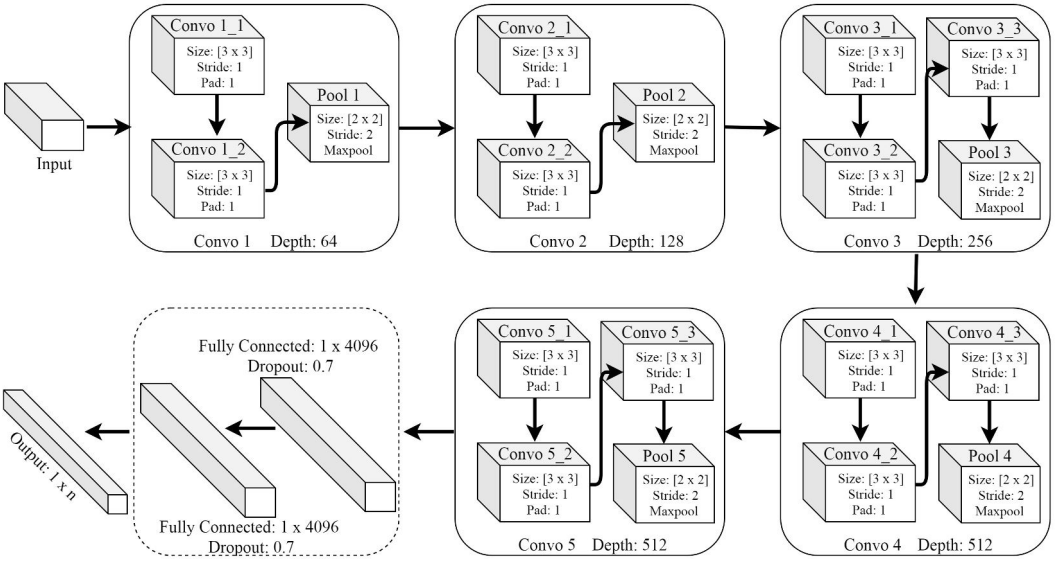

FIGURE 3.4: VGG-16 architecture for stance classification.

the temporal features squeezed by 3D CNN model. This empowers the proposed system for temporal feature learning to determine the type of an activity at a particular time module ΔT. The GRU model is an improvised version of Long Short-Term Memory (LSTM) technique (Cho et al., 2014). The working process of GRU is quite similar to Long Short-Term Memory (LSTM) but it contains some structural differences. GRU contains only two gates namely, reset and update gate. These gates are responsible to deal with the sequential dependencies in the temporal features of the input frame segments. Update gate of the GRU model is a combination of forget and input gate of LSTM. The reset gate of GRU is directly applied to the hidden state of the model. The update and reset gate of GRU model is further split by forget gate as shown in Fig.**??**. GRU model is also capable to deal with the less amount of data and take lesser time for training because of its fewer parameters as compared to LSTM. To handle the input features generated by 3D CNN model, the activation process is represented as:

$$\begin{pmatrix} z \\ r \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \end{pmatrix} \quad \left(W \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix} \right) \tag{3.2}$$

$$h = tanh\left(W\begin{pmatrix} x_t \\ r \odot h_{t-1} \end{pmatrix}\right) \tag{3.3}$$

$$h_t = ((1-z) \odot h + z_t \odot h_{t-1}) \tag{3.4}$$

where σ denotes the sigmoid function. The reset gate, update gate, candidate activation state, and previously hidden state is denoted with r, z, h_t and h_{t-1} variables and the flow of the data from one gate to another is explained in Fig.**??**. At the final stage of activity prediction, the long-term context features are classified using softmax function and produced a probability score for each feature matrix. Overfitting is also considered as one of the major problems in deep learning. A modified dropout method (Moon et al., 2015) is used to handle overfitting and to improve the prediction efficiency. This method sets the output value to zero with the probability of 0.5 to reduce the problem of overfitting in the proposed network.

3.3.3 Physical abnormality-based record generation

Final predicted scores generated by proposed methodology is transmitted to the local database of the system which can be used for medical or therapeutic purposes. The predicted probability value describes the type of an activity performed by the person in a particular time module. The probability value of the predicted activity defines the sensitivity of the abnormality.

Definition 2: (Time Series for Irregular Activity (TSIA)) Given a frame segment FS_i based predicted irregularity IA_i at a particular time instance t_i, the Time Series for Irregular Activity (TSIA) is described as the prediction time t_i consumed by the activity prediction module for abnormality prediction is represented as: (t_i, IA_i).

1. Definition 2 utilized to assign a specific time instance value to every predicted abnormality by the proposed methodology.

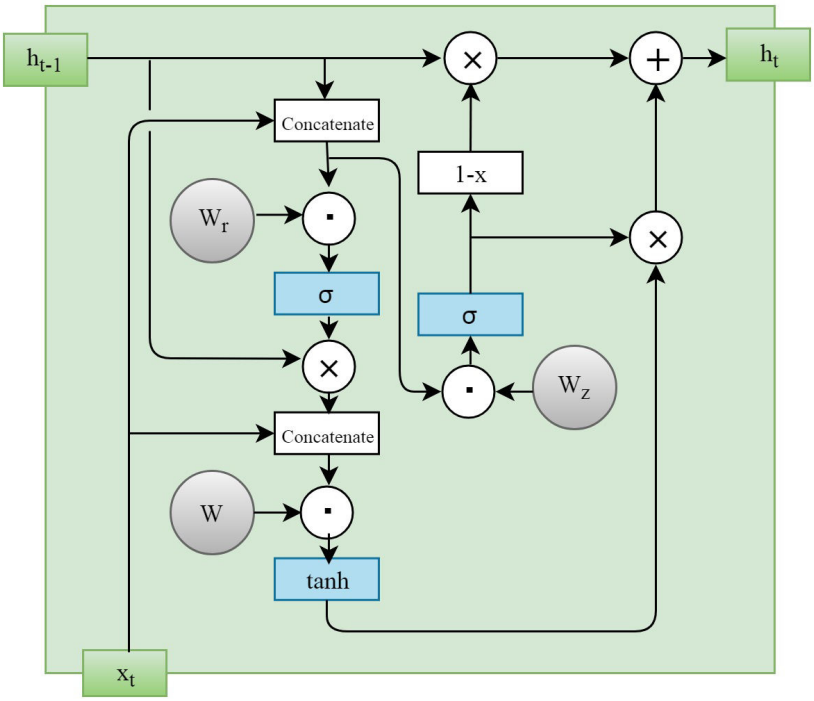

FIGURE 3.5: GRU cell architecture.

2. In this manner, several abnormal activities are correlated with each other based on a specific time module.

Result storage: To deal with the parameter of temporal diversity in physical activities, the physical movements need to screen in a continuous manner so that the predicted probability scores can be stored. In the proposed system, the predicted activity results are stored in the database continuously in the form of the temporal instances. By storing the predicted probability scores, we obtained a combined activity record to analyze the condition of an individual. Algorithm 1 explained the process of result storage in the database.

Physical state analysis: Mining layer is responsible to extract the predicted activity results from the database of the system to analyze the physical state of an individual. The requested information is fetched in the form of a continuous time series pattern using Temporal Mining Technique (Sacchi et al.,

2007). Temporal mining based data abstraction method helps to present predicted activity results in a common format which helps the doctor or caretaker to analyze the physical state of an individual for medical or therapeutic purposes. The process of information retrieval from the database is described in Fig. **??**.

Algorithm 1: Prediction score based record generation
Step 1: Calculate the scale of an activity based on the segment of the frame at a particular time instance t.
Step 2: If (Activity_Occurrence = True) Goto Step 3 else goto Step 5.
Step 3: Perform the process of result storage: Do **Step 3.1:** Create a check point at time t_i to save the predicted result. **Step 3.2:** Generate space to database storage by deleting the initial log value. **Step 3.3:** Record formation for predicted result by using sliding window.
Step 4: ADD current activity score.
Step 5: Return to step 1.
Step 6: End

Definition 3: (Temporal Granule Series (TGS)) TGS of predicted activities based on the requested time module ΔT is described as a combined form of abnormal activities occurred in between a sequence of time instance $|t_s - t_e|$.

Temporal Granule Series (TGS) is responsible to present the sub-portion of performed activities to the concerned caregiver or medical specialist based on the requested time module for state analysis. Furthermore, with the help of TGS, the doctor or caregiver become more capable to analyze the physical condition of an individual. Summarized information can be utilized by the hospitals, medical agencies, and government organizations for designing new work policies to deal with the issue of GAD and also for survey purposes.

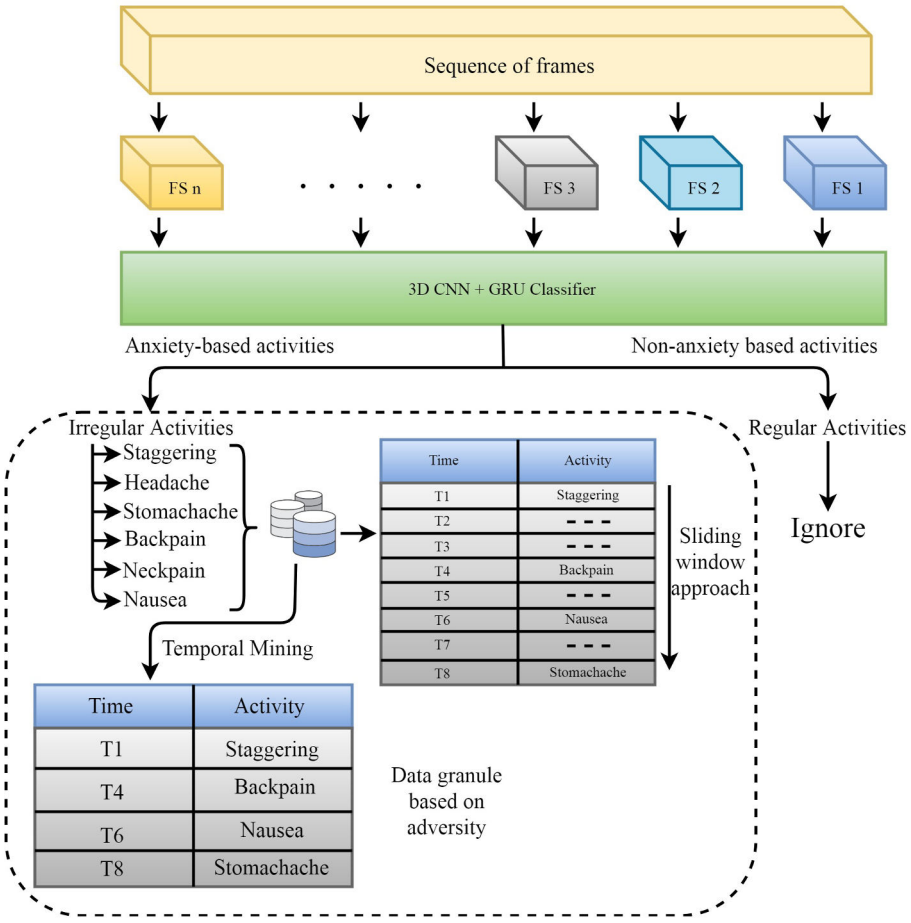

FIGURE 3.6: Temporal data granule based on abnormal activities.

3.3.4 Smart decision making

The smart decision-making module is responsible for generating alerts along with the current physical status of an individual. Activities can be monitored by using two basic procedures, Alert-based monitoring and Continuous monitoring. In the proposed study, alert-based monitoring is considered as the best option to notify a doctor or caretaker to deal with health adversity. Current abnormality-based requested information delivered to the doctor upgrades the utility of the system in the smart healthcare domain.

Definition 4: Scale of Abnormality (SoA) Given an irregular activity IA_i

belongs to its activity class, the calculated prediction probability score for SoA defines the adversity of individual's health.

This definition helps to evaluate the health sensitiveness by determining the current physical condition. The SoA is essentially partitioned into two conditions, severe and non-severe. The adverse impact calculated in Definition 2 shows the discomfort related to health caused by the occurrence of an irregularity IA. The detailed procedure for the measurement of adversity and alert generation is explained in Algorithm 2.

Algorithm 2: Alert generation with physical status deliverance
Step 1: **Input:** Current frame segment F_i
Step 2: Activity-based probability calculation at time instance t_i as described in definition 2 to determine the scale of anxiety.
Step 3: If (*Activity_Score* (i)) not lies in system's trained classes, goto step 4 else goto step 5.
Step 4: *Physical_State* = *Not_Severe* goto step 6.
Step 5: *Physical_State* = Severe **Step 5.1:** Generate medical alert signal to family member and responder with the current physical status to handle medical emergencies if any adverse situation prevails.
Step 6: Repeat step 1 for definite time.
Step 7: End.

The proposed alert generation procedure is followed in two phases. In the first stage, the physical activity is analyzed by calculating the adversity using proposed activity prediction methodology. In the second stage, a real-time alert is generated to the responder for providing required medical or assistive services. The process of an alert generation with the delivery of the current physical status of an individual increase the novelty of the system. Fig. **??** explained the flow of real-time decision making.

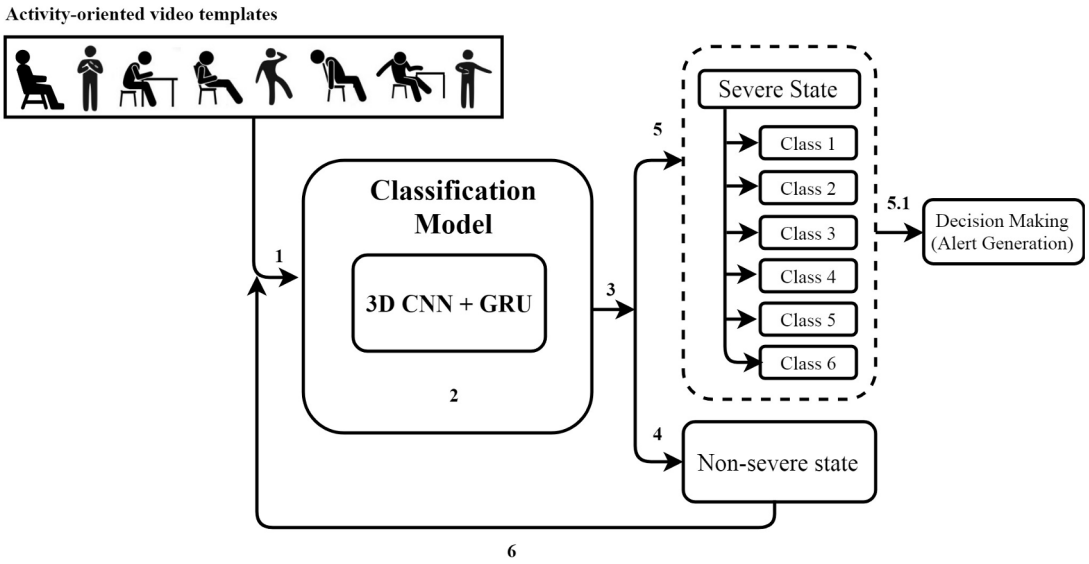

FIGURE 3.7: The procedural flow of decision making.

3.4 Experimental evaluation and performance analysis

The proposed system is implemented on a workstation with the following configuration: CPU: Intel i5-6600 processor, RAM: 16 GB, GPU: NVIDIA GeForce GTX 980 Ti, GPU Memory: 6GB GDDR5, CUDA Cores: 2816. Several tools and libraries used to implement the proposed methodology includes Operating System: Ubuntu (14.04), Programming Language: Python (2.7.11), IDE: Spyder (3.1.4), Tensorflow (1.10.0), Keras (2.2.0), and OpenCV (3.0.0). Mysql database management system is also used to store the predicted activity results and helps to generate activity records.

Dataset: The proposed system's performance is assessed on several health activities as listed in Table **??**. To maintain the variability and to increase the robustness of the system for activity prediction, several health activities are selected from NTU RGB+D dataset (Shahroudy et al., 2016) including staggering, headache, stomachache, backache, neck pain, and nausea. These vital conditions are considered as the most common physical abnormalities

in GAD. The NTU RGB+D dataset is considered as one of the largest publicly available activity dataset. The activities of this dataset are broadly divided into 40 routine-based activities, 9 health-oriented activities, and 11 mutual activities. The dataset consists of more than 56,000 videos with 4 million frames. The dataset contains four different data modalities captured by the Kinect sensor: RGB frames, skeleton data, IR sequences, and depth maps. In this study, RGB modality has been used to train an test the efficiency of the system. Fig. **??** show the examples of abnormal activities which are used to train and test the prediction performance of the system.

Performance analysis: As discussed above, the system involves four noteworthy modules. In the first step, video templates related to an individual's physical activities are captured utilizing visual sensors. In the second step, every activity is predicted and classified based on the severity level into its respective activity class using the proposed methodology. In the third step, the predicted activity scores are stored in the database of the system and the sub-portion of results based on the requested time module are extracted from the database using temporal mining technique. To satisfy the goal of the smart assessment, an alert mechanism with current status deliverance is proposed. In lieu of these perspectives, the execution of the proposed modules are evaluated as following:

1. Hyperparameter selection and video pre-processing time analysis.
2. Abnormality prediction efficiency of the system.
3. Performance analysis of the system with state-of-the-art outcomes.
4. Alert-based decision making efficiency of the proposed model.
5. Average training and abnormality prediction time analysis.
6. Prediction performance validation on the public datasets.

FIGURE 3.8: Examples of different abnormalities from the dataset.

3.4.1 Hyperparameter selection

The proposed methodology is tested on different learning rates to choose the optimum prediction model. The entire model is tested on the different learning rates from 0.1 to 0.0001. In the first phase, the VGG-16 based 3D CNN model is used to classify the body postures of the person. By using max-pooling operation, the output feature based on the context region is pooled

together and provide input to GRU. The GRU model is tested on all possible combinations of N units, where N = 68, 128 and 256. The second phase deals with context learning by using the extracted temporal features. The proposed model takes 60 to 110 min for training. The best outcomes based on different learning rates are displayed in Fig. **??**(a). It can be observed that the comparative outcomes of the proposed methodology can be achieved on 0.01 and 0.001 learning rates as compared to 0.0001 learning rate. The average accuracy of the proposed system is also tested by varying the number of units in the GRU model. The calculated results on the different number of GRU units are presented in Fig.**??**(b). We conclude that the GRU model with 128 number of units at the learning rate of 0.001 improves the prediction performance over other variants.

Time complexity analysis for data pre-processing: Time complexity is considered as one of the most sensitive and influential parameters of the study which helps to verify the data processing cost complexity of the system. The performance of the proposed study is analyzed by calculating the time taken by the system for video pre-processing at the experimentation stage. The basic need to pre-process the video data is to maintain the performance of the system by decreasing the data processing cost. The pre-processing time is calculated for Video to Frames, Frames to Segment, and Segment Compression operation.

The time duration for segmentation generation and the number of frames in a frame segment are fixed. The frame generation rate is also uniform. The specified constraints help to evaluate the performance in an accurate manner. Fig. **??** explains the computation time taken by the system for Video to Frames, Frames to Segment, and Segment Compression operation by consuming the time of 43 ms, 78 ms, and 62 ms, respectively. As the time and cost complexity is directly proportional to each other, we can conclude that the data pre-processing cost of the proposed system is highly efficient.

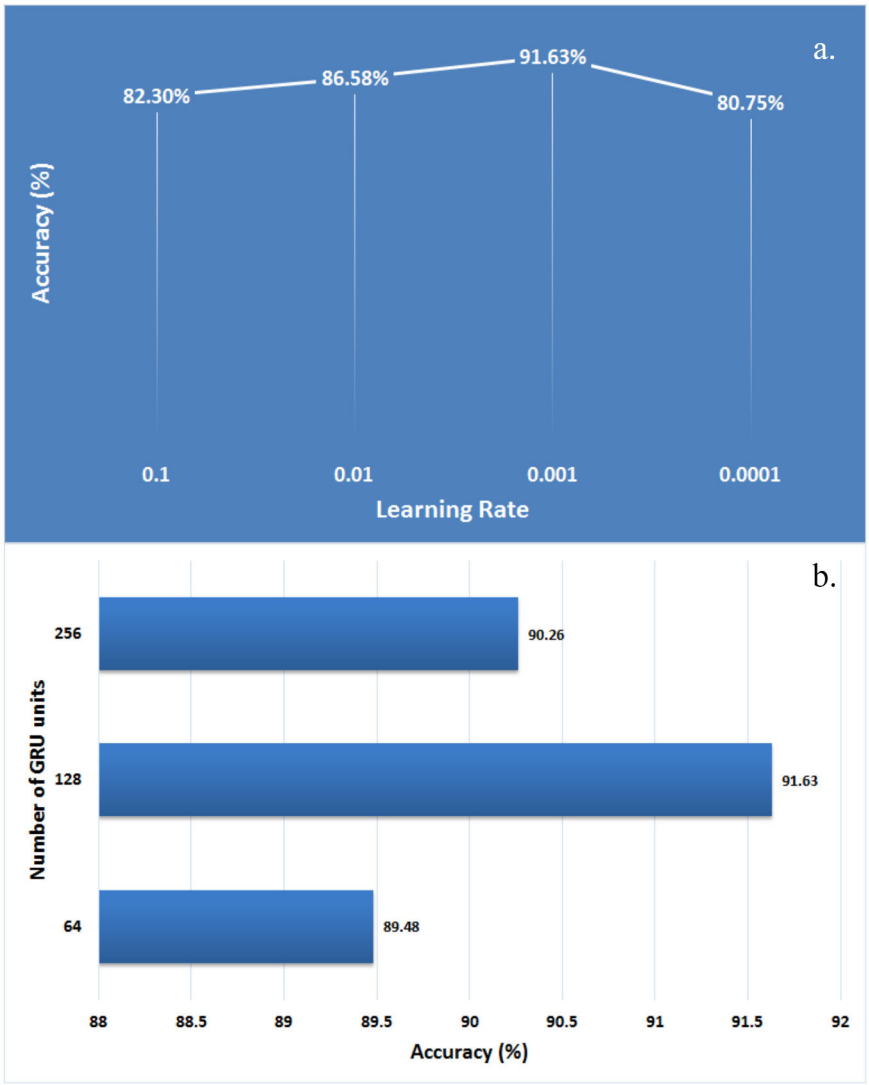

FIGURE 3.9: System performance analysis (a) Learning rate analysis, (b) Number of GRU unit selection.

3.4.2 Abnormality prediction efficiency

The efficiency of abnormality prediction is determined by differentiating the preformed activity into pre-specified activity classes of the system. The abnormality prediction efficiency is calculated into three approaches:

1. Frame-based early abnormality prediction.

2. k-fold cross-validation based average prediction accuracy.

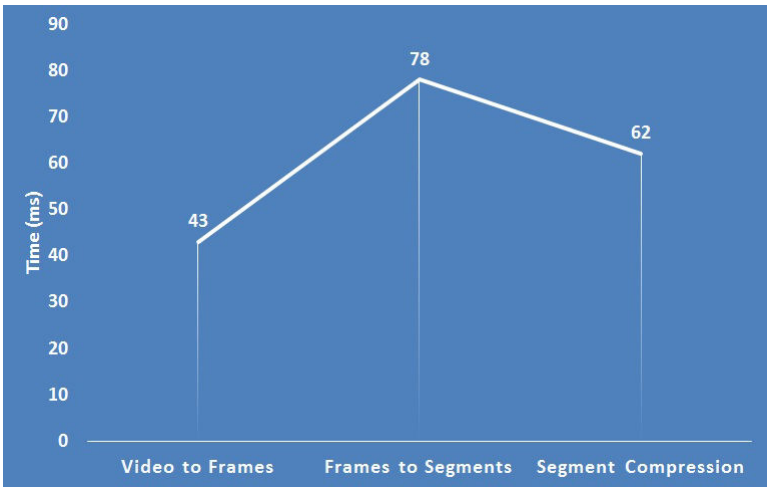

FIGURE 3.10: Time complexity analysis for video pre-processing.

3. Class-based mean accuracy of each physical activity.

Frame-based quantitative approach

To justify the effectiveness of the proposed vision-based solution for smart healthcare or assistive environment, we focused more on near abnormality detection for early abnormality prediction. We present a quantitative approach to prove the prediction efficiency of the proposed system. Table **??** depict the frame level prediction performance on 20 testing videos.

Where S_i represents the frame segment number, +ve represents the anomalous frame segment, -ve represents the frame segment without anomaly or fake events. Table **??** depict that the proposed solution successfully localized the anomalous frames by using Intersection of union (IoU) parameter. If the value of IoU is greater or equal to 0.7, the analyzed frame is considered as a true localized frame. If the methodology cannot localize it or localize it with less value of IoU, that frame will be considered as a false frame which does not contain any physical abnormality. Here, we present the analysis of early abnormality prediction for 20 testing videos (12 has positive frame segments, 8 has negative frame segments). By analyzing the recognition results,

TABLE 3.2: Frame-based early abnormality detection

Segment Identifier	**Segment Status**	**Groundtruth/Total frames**	**Localization**	**Non_Localization**
S_1	+ve	15/242	15	0
S_2	-ve	0/257	0	0
S_3	-ve	0/164	0	0
S_4	+ve	12/268	10	0
S_5	+ve	9/223	7	0
S_6	+ve	24/285	21	0
S_7	-ve	0/238	0	0
S_8	+ve	18/263	18	0
S_9	+ve	25/228	22	5
S_{10}	-ve	0/246	0	0
S_{11}	-ve	0/236	0	0
S_{12}	+ve	6/257	6	4
S_{13}	-ve	0/324	0	0
S_{14}	+ve	35/277	29	0
S_{15}	+ve	17/258	17	0
S_{16}	-ve	0/242	0	0
S_{17}	-ve	0/274	0	3
S_{18}	+ve	8/252	6	0
S_{19}	+ve	10/208	8	0
S_{20}	+ve	6/316	3	0

we can say that the proposed methodology has achieved real-time prediction performance at a high frame rate (30 fps) for 224 $\times$ 224 image resolution and have an overall competitive performance for near abnormality prediction.

K-fold cross-validation

The dataset contains 6 health-oriented physical activities including staggering, headache, stomachache, back pain, neck pain, nausea or vomiting conditions. Physical activities like staggering and stomachache with similar physical trajectories make the dataset more challenging for exact activity prediction. In this study, a 5-fold based rigorous cross-validation technique is used to evaluate the prediction performance of the system. 4-folds are used to train the system and remaining fold is used to test the prediction efficiency. In this manner, each fold has a chance to be a test set.

Fig. **??** shows an average prediction of each activity for 5-folds. The average prediction results described that most of the time the confusion occurs between staggering and stomachache. Total 10% activities of staggering are misclassified as a stomachache. Given the fact that the less number of training templates can be a reason for this problem that can be resolved by increasing the activity samples related to that particular activity in the training phase. As we can see from Fig. **??**, headache and neck pain activities shared almost similar physical trajectories. But the large training samples with different angles made the system more capable to predict these activities correctly with sufficient margin. For the other activities, the confusion rate is below to 5%. To improve the prediction efficiency for these activities, there is scope to enhance long-term movement representation technique by either increasing the number of GRU units or combining the other data modality for prediction stability. Table **??** also explains the aggregated mean accuracy of each fold with a standard deviation of the proposed approach.

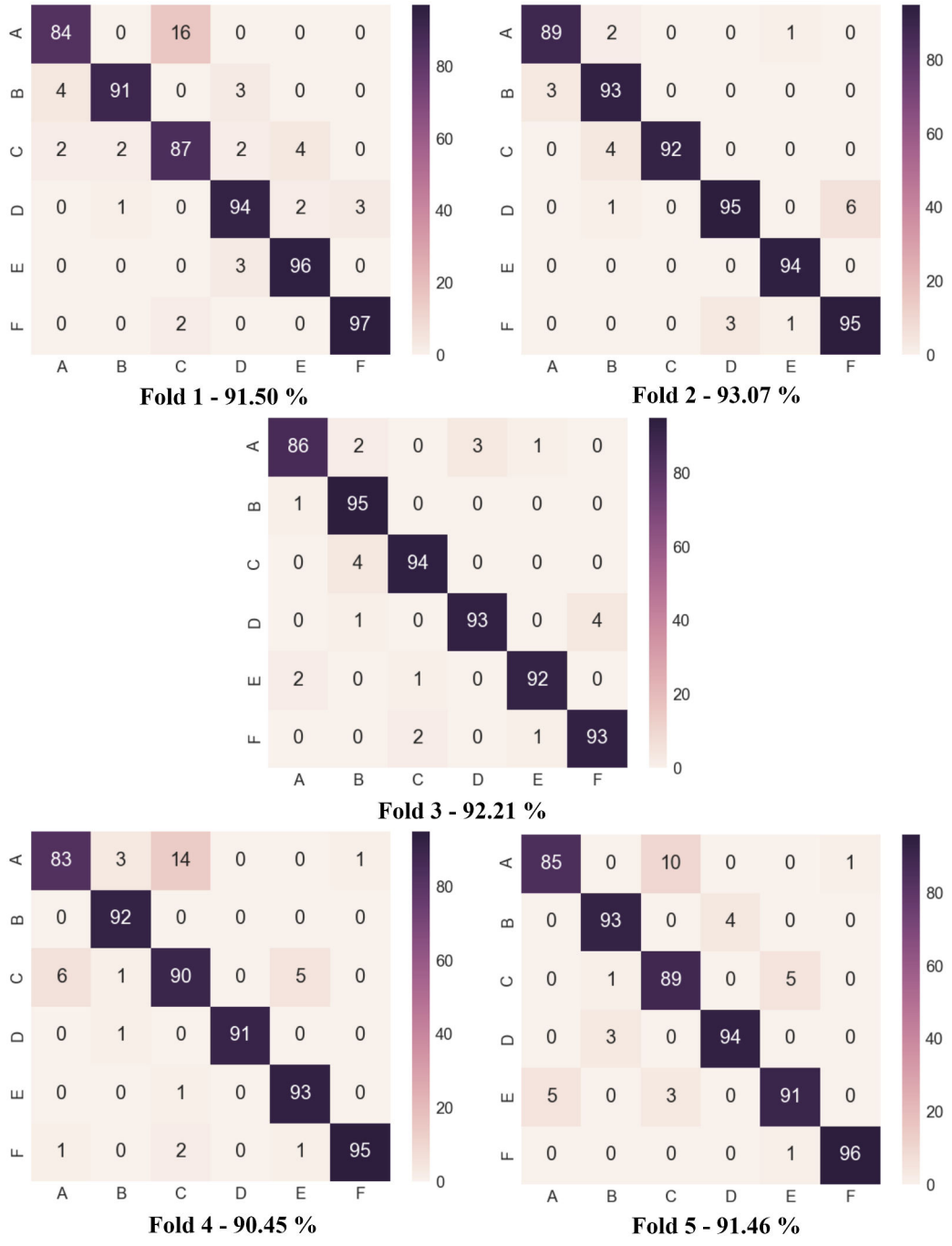

FIGURE 3.11: 5-fold based confusion matrices representation

TABLE 3.3: Individual 5-fold cross validation.

k-fold (5-fold)	Accuracy
Fold-1	91.50%
Fold-2	93.07%
Fold-3	92.21%
Fold-4	90.45%
Fold-4	91.46%
Mean	91.88%
Standard deviation	$\pm$0.94

TABLE 3.4: Class-based mean accuracy scores.

Performance Measures	**A (%)**	**B (%)**	**C (%)**	**D (%)**	**E (%)**	**F (%)**
Accuracy	85.54	90.61	87.27	93.72	96.17	97.25
Specificity	83.68	90.57	89.85	91.36	90.89	92.72
Sensitivity	80.68	88.75	87.55	89.63	88.98	89.75
F-measure	82.18	89.51	88.78	90.09	89.24	91.31

Mean accuracy of class-based activity

The performance of the proposed methodology is assessed by calculating four different performance matrices including accuracy, specificity, sensitivity, and F-measure score. The mean accuracy of each activity is presented in table **??**.

Based on the calculated outcomes, we can presume that the proposed 3D CNN + GRU model is more productive, even for the small-scale dataset. Mean outcomes related to other performance matrices are also presented graphically in Fig.**??** for better understanding. The proposed model acquired the classification accuracy of 91.88%. For sensitivity, the system has registered better accuracy of 87.62%. Specificity defines the exactness in activity prediction state and the proposed classifier achieve significantly higher

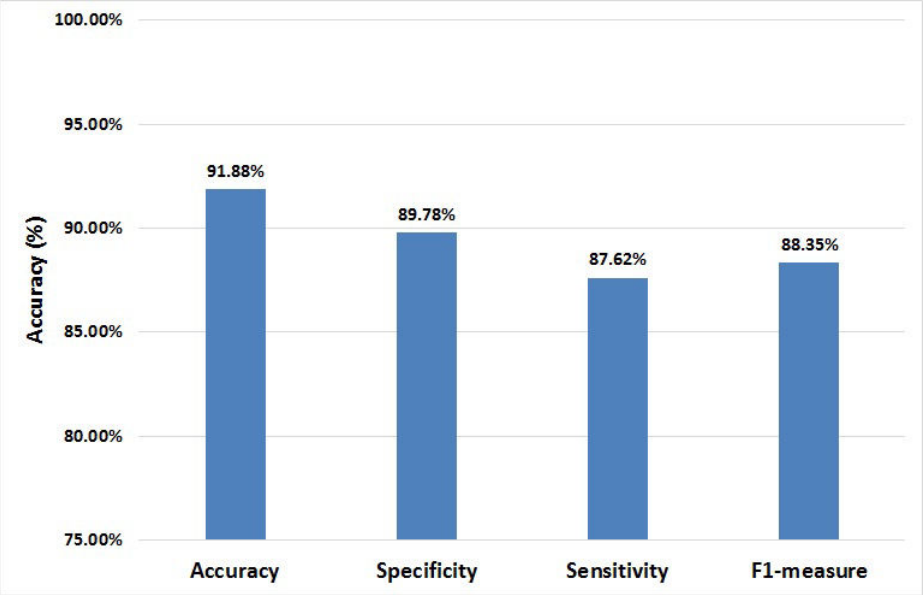

FIGURE 3.12: Comparative outcomes of classification efficiency.

specificity numerating to 89.78%. In the case of f-measure, the proposed system also registered better performance by numerating to 88.35%. After analyzing the results acquired in the current scenario, the proposed deep learning based activity prediction methodology is considered as highly effective for physical abnormality prediction in real-time.

3.4.3 Comparison of the proposed methodology with state-of-the-art methodologies

The activity prediction performance is tested with four CNN-based deep learning methodologies named Temporal stream (Simonyan and Zisserman, 2014a), LSTM (Yue-Hei Ng et al., 2015), TDD (Wang, Qiao, and Tang, 2015), and Transformations (Wang, Farhadi, and Gupta, 2016) as reported in section 2 for fair comparison. Mean accuracy of each activity is compared and presented in Table **??**. The other performance matrices are also calculated to analyze the performance of the system and illustrated in Table **??** which are

TABLE 3.5: Comparative analysis of class-based mean accuracy.

Methods	**A (%)**	**B (%)**	**C (%)**	**D (%)**	**E (%)**	**F (%)**
Simonyan and Zisserman, 2014a	86.82	70.91	89.50	83.33	64.40	87.30
Yue-Hei Ng et al., 2015	89.38	83.27	84.51	67.72	88.45	93.75
Wang, Qiao, and Tang, 2015	87.70	84.62	83.62	78.18	89.82	90.64
Wang, Farhadi, and Gupta, 2016	82.94	87.54	84.36	88.58	90.22	91.52
Proposed	**85.54**	**90.61**	**87.27**	**93.72**	**96.17**	**97.25**

further presented graphically in Fig. **??** for better understanding.

From Table **??**, it can be concluded that LSTM (Yue-Hei Ng et al., 2015) outperforms the proposed and other methodologies for stagerring prediction by achieving the average accuracy of 89.38%. For stomachache prediction, the Temporal stream (Simonyan and Zisserman, 2014a) outperforms the other methodologies by achieving 89.50% accuracy. But for the other activities, the proposed methodologies outperforms the state-of-the-art methodologies. From figure **??**, it can be noticed that the proposed methodology outperforms the state-of-the-art methodologies in the average accuracy of activity prediction by achieving the higher accuracy of 91.88% with the closest competitor being Transformations (Wang, Farhadi, and Gupta, 2016) with the average accuracy of 87.53%. The proposed approach also achieved the best specificity by numerating to 89.78%. The same improvement can be seen for sensitivity and f-measure by achieving the accuracy of 87.62% and 88.35%. By analyzing the outcomes, we can say that the proposed methodology is highly efficient for abnormality prediction compared to other state-of-the-art methodologies.

TABLE 3.6: Comparison of performance matrices with state-of-the-art methodologies

Methods	Accuracy	Specificity	Sensitivity	F1-measure
Simonyan and Zisserman, 2014a	80.41%	79.68%	77.52%	78.49%
Yue-Hei Ng et al., 2015	84.51%	82.93%	80.56%	81.46%
Wang, Qiao, and Tang, 2015	85.76%	84.12%	81.48%	82.94%
Wang, Farhadi, and Gupta, 2016	87.53%	86.29%	84.62%	85.47%
Proposed Methodology	**91.88%**	**89.78%**	**87.62%**	**88.35%**

3.4.4 Average training and anomaly prediction time comparison

The operations of feature extraction, system training, and testing are performed on NVIDIA GeForce GTX 980 Ti GPU. The total time taken by the 3D-CNN model for feature extraction is approximate 0.09 seconds for a single frame. In this study, we gave the input of a frame segment which contains 30 frames for taking an advantage of the processing capability of GPU. The GPU takes approximate 0.68 seconds for feature extraction of the frame segment. In the second step, the approximate time taken by the GRU for temporal feature modelling is 0.33 seconds. The total activity prediction time taken by the last layer of the GRU model is 0.12 seconds. The aggregated activity prediction time taken by the system for a single frame segment is 1.13 seconds. The same process of activity prediction is also performed on CPU and the CPU-based processing time complexity is compared with the GPU-based activity prediction and presented in Fig. **??**.

Fig. **??** demonstrates that the average running time for the proposed deep

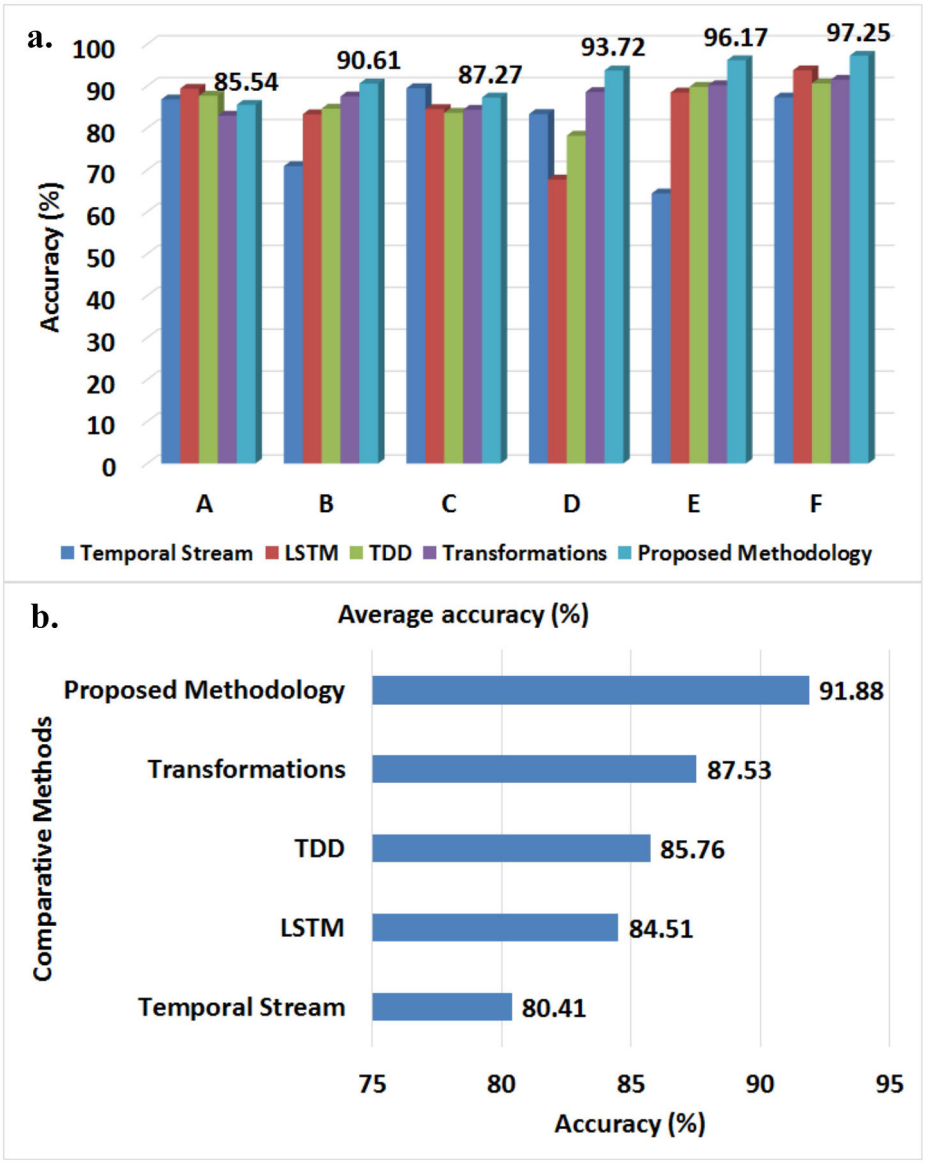

FIGURE 3.13: Comparative analysis of activity prediction efficiency with state-of-the-art outcomes. (a) Activity class based performance analysis, (b) Overall prediction performance analysis

learning approach is highly satisfactory, which guarantees the real-time prediction performance of the model. After analyzing Table **??**, we can conclude that the time consumed by the proposed model for activity prediction takes

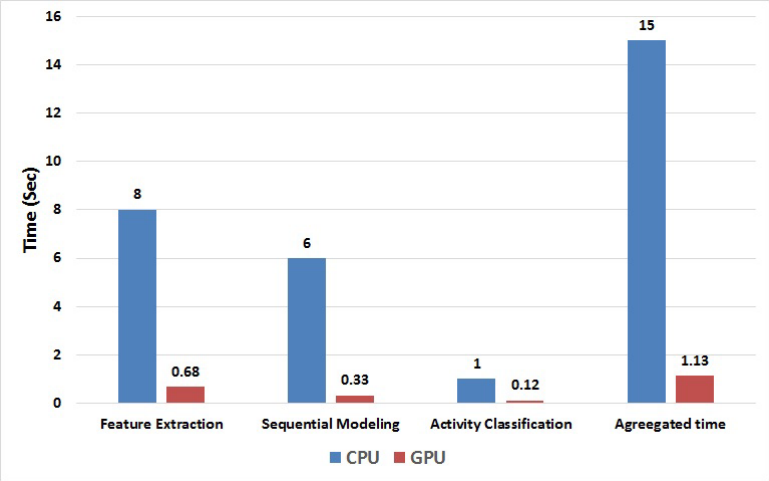

FIGURE 3.14: Activity prediction time on CPU and GPU.

TABLE 3.7: Activity classification scores.

Methods	Training time	Real-time anomaly prediction time
Simonyan and Zisserman, 2014a	1.58 s	1.43 s
Yue-Hei Ng et al., 2015	1.45 s	1.18 s
Wang, Qiao, and Tang, 2015	1.37 s	1.29 s
Wang, Farhadi, and Gupta, 2016	1.41 s	1.21 s
Proposed	**1.28 s**	**1.13 s**

less time compared to the time taken by the state-of-the-art methods for activity prediction. We can also state that the GPU-based activity prediction solution is well efficient for real-time activity monitoring compared to CPU-based monitoring solutions.

3.4.5 Alert based decision making efficiency

By analyzing the modular approach (Fig **??**) of the proposed system, we can say that the alert based decision making efficiency is directly dependent on

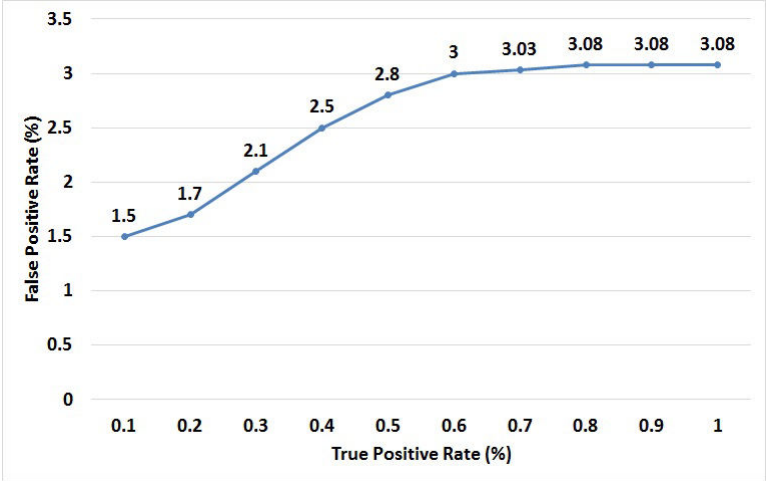

FIGURE 3.15: True-positive rate versus false-positive rate.

the activity prediction efficiency of the proposed methodology. The performance of the proposed alert mechanism is evaluated by calculating "false-positive" alerts and compared to the total calculated "true-positive" rate for the dataset.

Figure **??** explained that only 3.08% of the alarms are covered under the examination of false positive. The other performance matrices like accuracy (91.88)%, specificity (89.78)%, sensitivity (87.62)%, and f-measure (88.35)% defines the correctness of the activity prediction of the system. It can also be seen that the proposed model obtained less false-positive rate with high true-positive accuracy.

3.4.6 Performance validation on public dataset

The activity prediction performance of the proposed system is justified by validating the prediction efficiency on the two benchmark datasets including HMDB51 (Kuehne et al., 2011) and UTD-MHAD (Chen, Jafari, and Kehtarnavaz, 2015). The calculated outcomes are compared with the results of the state-of-the-art methodologies.

MHAD Dataset: MHAD dataset is considered as one of the challenging datasets for activity prediction by involving 27 different exercises. Similar trajectory and motion-based exercises like "draw a circle clockwise", "draw a triangle" and "draw circle counter-clockwise" makes it more troublesome for an activity prediction system. Performance evaluation protocol explained in (Chen, Jafari, and Kehtarnavaz, 2015) is utilized to train and test the methodology. Odd subjects are utilized to generate training samples and a total of 430 samples are selected from even subjects to test the model.

HMDB51 Dataset: The HMDB51 is another most challenging dataset which contains activities related to physical movements, facial interaction, physical exercises, sports, and dealing with objects. Total 6849 activity samples are contained by the dataset which is further categorized into 51 activity categories. Each category contains more than 100 video samples. The samples related to single activity has been taken from different sources with different viewpoints and illumination which makes the dataset more challenging. The average prediction accuracy of best in class methodologies lies under 60%.

In the proposed study, we predicted the activities by utilizing the high-level features extracted by 3D CNN and the intensity of activity is calculated by performing sequential modelling operation on temporal features using GRU. The combination of these two techniques helps to improve the prediction accuracy of complex activities. The comparison of the proposed methodology has been done with previously reported methodologies and the results are presented in Table **??**.

After analyzing Table **??**, we can say that the proposed prediction methodology achieved an average accuracy of 94.28% for UTD-MHAD dataset and 70.33% for HMDB51 dataset. We compared the performance of the proposed system with the CNN based methods including Temporal stream (Simonyan and Zisserman, 2014a), LSTM (Yue-Hei Ng et al., 2015), TDD (Wang, Qiao, and Tang, 2015), and Transformations (Wang, Farhadi, and Gupta, 2016). The

TABLE 3.8: Comparative analysis of the classification scores.

Methods	Accuracy (%)		Specificity (%)		Sensitivity (%)		F1-measure (%)	
	MHAD	HMDB51	MHAD	HMDB51	MHAD	HMDB51	MHAD	HMDB51
Simonyan and Zisserman, 2014a	85.67	60.76	83.78	58.38	81.14	56.29	82.76	57.43
Yue-Hei Ng et al., 2015	89.27	58.45	86.32	57.83	84.59	55.91	84.92	56.64
Wang, Qiao, and Tang, 2015	90.74	63.38	89.68	62.21	88.86	60.84	89.21	61.49
Wang, Farhadi, and Gupta, 2016	92.14	65.27	90.45	63.92	88.47	61.26	89.18	62.74
Proposed Methodology	94.28	70.33	89.32	69.06	90.84	62.34	90.68	65.53

proposed system is outperformed the Temporal stream (Simonyan and Zisserman, 2014a) method by achieving 8.61% improvement in the accuracy for UTD-MHAD. Total 9.57% of the improvement has also been reported for HMDB51 dataset. The other calculated statistics also represents the viability of the proposed framework for activity prediction on UTD-MHAD and HMDB21 dataset.

3.5 Conclusion

In this chapter, a computer vision-assisted monitoring system is proposed to address several physical abnormalities by considering upper body stance of an individual. Specifically, the proposed model consolidate two essential perspectives, namely (a) continuous monitoring with abnormality prediction and (b) alert based decision making based on the predicted abnormality. Both these aspects justify the efficacy and utility of the system for healthcare and assistive-care domain. Activity prediction over the abnormality scale shows the probabilistic estimation of the individual's health. The deep learning approach is consolidated as a novel aspect of the system to classify the anxiety-oriented activities. The activity prediction methodology is also responsible to generate alerts in real time to notify caregivers about the current physical state of an individual. Moreover, activity-based probability scores are mined to figure out the physical status of an individual for medical or therapeutic purposes. From the exploratory outcomes of the proposed system, it can be inferred that the proposed model outperforms the comparative approaches for abnormality prediction with less error rate by accomplishing 91.88%, 94.28%, and 70.33% mean accuracy for predicting the irregular and regular video templates from the NTU RGB+D, UTD-MHAD, and HMDB51 datasets, respectively. The comparative analysis justifies the activity prediction performance of the system. Additionally, numerical assessment legitimizes the usage of deep learning for domain sensitiveness. Hence, it can be reasoned that the proposed system is exceedingly viable and capable of providing an appropriate smart healthcare or assistive environment.

Chapter 4

Predicting Health Afflictions

4.1 Introduction

Internet of Multimedia Things (IoMT) with video analytics has lead to the development of wide range applications for several domains. It has the potential to be used in several remote applications like public surveillance, event recognition, and behavior monitoring in real-time. IoMT-based real-time video analytics can provide healthcare benefits, e.g., patient monitoring, irregular stance classification, physical activity based smart health suggestions. So, IoMT has the potential to increase personal independency and health satisfaction by transmitting an emergency alarm or notification to handle physical severity (Ghasemzadeh and Jafari, 2011; Alshurafa et al., 2014). But the deliverance of data to the cloud server for processing can cause delay which is unacceptable for healthcare domain. To overcome the data processing constraints of IoMT, the Edge-of-Things (EoT) technology (El-Sayed et al., 2018) has been introduced as an emerging technology. Edge computing is a platform which provides a local computational and communicational environment for data processing to handle video processing tasks efficiently (Garcia Lopez et al., 2015).

Essentially, the expression visual analytics defines the combination of several scientific solutions, such as, applications, services, and procedures (Cook and Polgar, 2014; Amor, Su, and Srivastava, 2016). Real-time video analytics provide several advantages relative to wearable sensing solutions

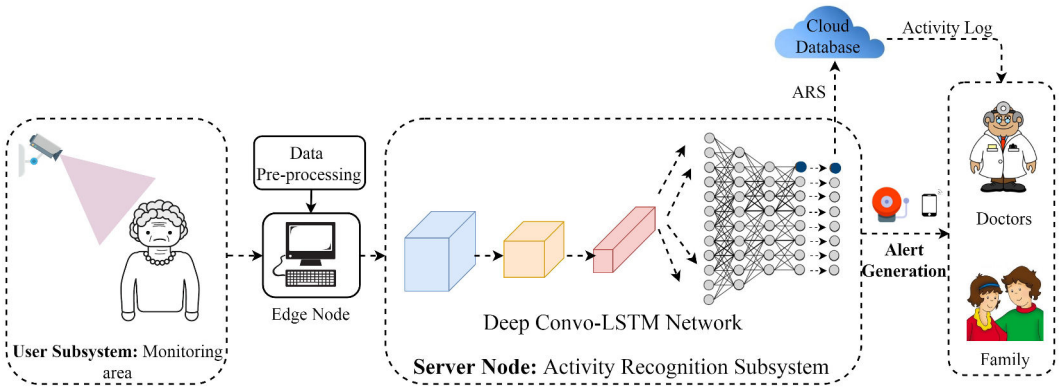

FIGURE 4.1: The base model of the proposed system.

such as wide area coverage, high quality of data, and low hardware cost. The advance data processing methods enable several monitoring opportunities by extracting more information about the event (Lin et al., 2016). The passiveness of video capturization is considered as the best advantage for the healthcare domain. In video analytics, the patient only need to be visible in the range of the visual sensors for analyzing physical activity without wearing any special kind of wearable device.

Considering previously discussed advanced factors of edge computing and video analytics, the proposed system is primarily focused on analyzing the physical postures of an individual to predict health afflictions in real-time. The idea of combining the features of edge computing with video analytics is to increase the hardware proficiency by minimizing the cost of the data transmission to the remote servers. The fundamental aim of the edge computing in the proposed study is to provide superior computational capabilities by reducing delay and computational cost of the remote IoT servers. Fig. **??** described the contribution of the study by explaining the novelty of the system into four objectives:

1. Establishing visual sensor-enabled smart environment for individual monitoring.

2. Edge computing assisted deep learning based activity analysis in real-time.

3. Preformed activity-based record storage on the cloud for medical care.

4. Warning and Emergency alert based smart decision making with the deliverance of physical status of an individual to medical representative and caretakers.

The rest of the chapter is sorted into multiple sections. A review is made in section 4.2 by discussing conventional and modern activity recognition systems. The necessity of edge computing is discussed in section 4.3. Section 4.4 provides the detail about the material and methods used to implement the proposed model. The results of the system are evaluated in section 4.5. In section 4.6, the chapter is concluded with important points related to the proposed work.

4.2 Related works

In this study, we specifically focused on real-time activity recognition systems by utilizing the principles of edge computing and video analytics. we have discussed several activity recognition system by dividing into two parts, handcraft techniques based conventional activity recognition systems and deep learning based modern activity recognition system.

4.2.1 Conventional activity recognition system

Chaaraoui, Climent-Pérez, and Flórez-Revuelta, 2013 configure contour points to characterize the human body on human silhouette. To calculate the distance between contour points, authors utilized Euclidian distance. They globally represented each pose of human body using ecludian distance technique. Further, they employed k-NN algorithm to calculate distance between the sequence of key body postures. Authors in (Tran and Sorokin, 2008) also used the human silhouette to determine the body pose. They have formulated a local descriptor by combining the optical flow technique with human silhouette. The frames are divided into 15 frames per segment to calculate the summary of the motion. Final local descriptor features are used as the

input of 1-Nearest Neighbor classifier with Metric learning to recognize human activities.

In the article (Weinland, Özuysal, and Fua, 2010), the authors have converted the sequence of frames into histogram blocks to calculate 3D spatiotemporal gradients. After formulating the blocks, the distance between the blocks have been calculated to provide input to the classifier. Multi-local classifiers are used to predict human activity. The calculated results are further combined using product rule to get final output. Pehlivan and Duygulu, 2011 also used 3D information for activity recognition. The sequence of images are used to construct the circular model to encode 3D pose. The features of circular model are divided into three parts: (i) total number of circles, (ii) area of an outer circle, and (iii) area of an inner circle. Distance metric algorithm is used to calculate the distance between the circle keypoints for activity recognition.

4.2.2 Modern activity recognition system

Recently, Deep learning has incorporate feature extraction and movement analysis that attracts more consideration from researchers involved in computer vision domain. In article (Karpathy et al., 2014), authors have used three diverse CNN structures called late, early, and slow fusion to train activity recognition model from the RGB video recordings. (Ji et al., 2013) proposed a novel 3D-CNN model to recognize the activities of a person. 3D convolutions are used to extract the spatio-temporal features from the sequence of frames in this proposed model. Wang et al., 2017a proposed the system to calculate the decisions over element level fusion based on RGB features. Zhu et al., 2016 proposed human activities by generating the frame segments of key postures. The authors defined the movements by utilizing the maximal entropy markov model to characterize human activities without mentioning the starting and ending point of activity. Cippitelli et al., 2016 proposed a model to extricate key postures to create a single feature vector to classify

multiple activities by using multiclass Support Vector Machine. A CNN-GRNN based hybrid model is proposed by (Zhang, Shao, and Luo, 2018) to improve the recognition accuracy of an image. Two models named CNN and General Regression Neural Network (GRNN) are used to extract multilayer features from an image. The latest study has proposed a two-stream CNN model to deal with spatial and temporal features by using different CNN based network to perceive activity from video templates (Han et al., 2018).

4.3 Role of Edge Computing in the proposed system

Edge is a middle layer (Fig. **??**) between the user subsystem and cloud layer that supplement the benefits of distributed computing by providing data processing services for the emerging requirements of IoMT. The fundamental aim of edge computing in the proposed system is to increase the response rate by reducing the system's architectural complexities. Edge computing decrease the latency rate in an alert generation. The main components of the proposed system are:

1. **Visual Senors node:** The physical activities are captured through visual sensors installed in the local environment of an individual. Several preprocessing operation has been performed on the sequence of frames such as compression, segment generation. After preprocessing, the segments are transmitted to the edge node through wireless medium for analysis.

2. **Edge Node:** Mobile device with an adequate computational and storage capabilities to perform video analytics operation to predict health afflictions from the segment of frames.

3. **Cloud System:** Cloud platform provides the facility to store the predicted results by enhancing the capability of the long term storage.

As shown in figure **??**, the edge analytic offers improved data processing services to upgrade the framework from multiple dimensions. The primary

advantage of the edge layer in the proposed system is described in the following section.

4.4 Proposed Work

The main objective of this framework is to monitor anomalous physical activities of an individual to predict health afflictions using edge analytics technology. Moreover, the features of edge computing like mobility support, scalability and real-time interactive services can serve as an ideal decision for the healthcare domain. Fig. **??** explains the novelty of the study by dividing the framework into four phases, namely Data acquisition and preprocessing phase, Edge analytics-based activity prediction phase, Activity record generation phase, and Real-time alert-based notification generation phase. All phases perform their operations independently and provide an effective operational environment to its next phase. The implementation methods are discussed in the following sub-sections:

4.4.1 User subsystem: Data acquisition and preprocessing

In initial phase of the framework, physical patterns need to retrieve ubiquitously from visual sensors embedded in the individual's surrounding. The visual sensor can commanded through wired or wireless medium and are capable of capturing and transmitting data in real-time. The individual's physical patterns can be collected in a sequence of frames and need convert into compatible format before sending it to the edge module for further processing.

Maximum realistic situation-based activities are used for the training purpose of the system. NTU RGB+D dataset is used to incorporate train and test the preformance of the proposed framework. A set of 6 activities has been considered to determine the health afflictions of an individual as listed in

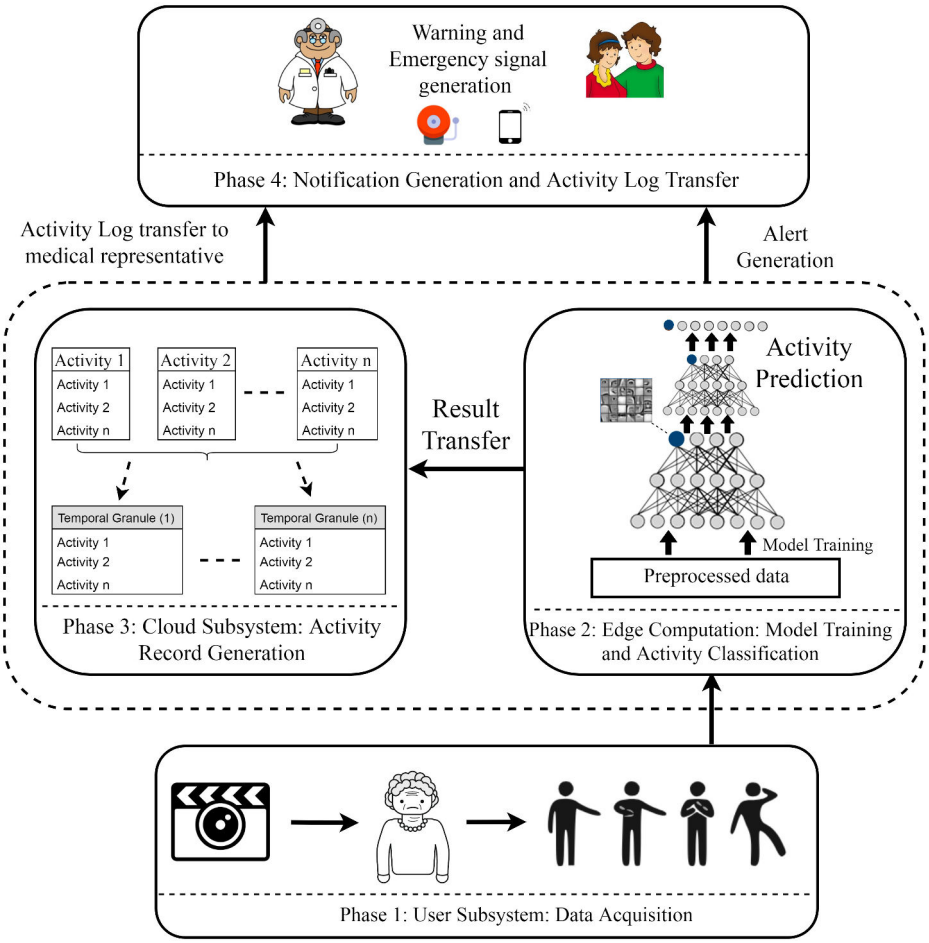

FIGURE 4.2: The division into phases of the proposed framework

Table **??**. The NTU RGB+D dataset provides a large collection of human activities and we have utlized RGB dataset modality to test the activity recognition performance (Shahroudy et al., 2016).

Data augmentation techniques has become exceptionally useful to deal with the problem of over-fitting in deep learning. Over-fitting is a problem which usually occurs due to the limited training data. Several data augmentation techniques such as, scaling, cropping, rotation has been performed on the dataset to increase its size and variability in the dataset to remove the recognition biasness of the system. To perform crop operation, a frame is

TABLE 4.1: Activity patterns measurements

Activity Classes	Definition
1. Cough or Sneezes	Symptoms of cold.
2. Straggling	Type of abnormal walk.
3. Back Pain	The problem in any part of the spine can cause back pain
4. Nausea	Nausea is a type of stomach discomfort and the sensation of vomit.
5. Unconsciousness	Feeling uncomfortable due to high temperature or unconsciousness.
6. Falling	Towards the ground without intending to or by accident.

TABLE 4.2: Dataset description

Type of Dataset	IoT Technology	Sensor Specification	Description
Video Dataset	Visual sensor (Wide Angle)	Pixel density: 2 mp, Aspect ratio: 16:9, Frame rate: 30fps, frame resolution: 1920×1080	Physical posture oriented real-time video.

randomly selected and cropped the region by selecting the value between 0.06 to 1. The value is selected based on the subject location in the original frame. After performing the crop operation, the scale operation has been performed by selecting the scale value of 0.75 to 1. At last, the scale operation is performed again on the augmented frame to convert the size to its original size. The color jittering-based data augmentation operation is performed to provide spatial strength to the classification model. Table **??** gives an overview of the dataset with sensors specifications used in this study to capture physical activities.

Data Pre-processing: It is assumed that the visual sensors used in this study is connected through a wireless medium. The visual sensors node is responsible for segment generation (Definition 1) and transmission. Before segment transmission, the frame segments are compressed by performing

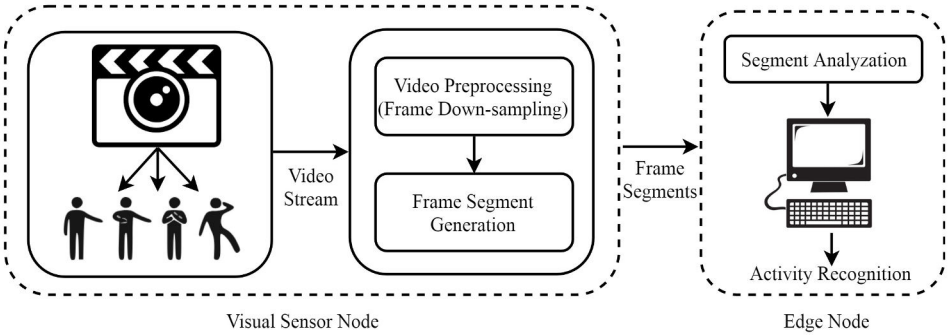

FIGURE 4.3: Visual sensor node: Data Preprocessing.

downsampling operation. The compressed segments are transmitted to its associated edge node as described in Fig.**??**.

Definition 1: Frame Segment (FS) A frame segment F_S contains the number of frames F = $\langle f_{i-1}, f_i, ..., f_{i+n+1} \rangle$, where the frames are pre-initialized to 30.

The downsampling ratio is characterized as the size of the frame segment after performing the compression operation on the original frame segment. For example, the downsampling ratio of frame segment FS_i is defined as:

$$FS_i = \frac{FS_c}{FS} \tag{4.1}$$

where the original size of the frame segment is denoted as FS and the downsampled frame segment is denoted by FS_c. As we know the quality of video directly affects the activity recognition rate of the system. To find the most relevant downsampling rate for video compression, quantization parameter (Chen et al., 2017) has been used to calculate the quality of the video and describe the relationship between the downsampling rate and activity prediction accuracy. The higher value of the quantization parameter points to the low quality of the video. After frame segment generation and compression, the edge nodes perform segment analysis in order to predict health afflictions.

4.4.2 Edge analytics: health affliction prediction

We have assumed that visual sensors and edge node share a wireless communication medium and the time of segment transmission is divided into fixed slots lasting several milliseconds (Cao et al., 2016). In long-term monitoring, there is a possibility of Internet unavailability hence leading to the failure of cloud based system. On the other hand, edge analytics provides the facility of segment analysis near to the IoMT layer instead of transmitting frame segments on to the cloud and awaiting response. The edge layer also provides the local data storage facility in the case of Internet unavailability and perform data synchronization with the cloud storage Internet connection is restored. Hence, the response time of the edge-based system for health-based irregularity prediction is much less and more reliable as compared to only cloud-based data processing.

Anomalous Activity Detection:

A deep learning-oriented multistage (3D CNN-LSTM-FC) activity classification methodology is proposed to analyze the health affliction as shown in Fig. **??**. In our methodology, 3D Convolution Neural Network (3D CNN) is utilized to extract features from the frame segments. After extracting the features, Long-Short Term Memory (LSTM) network is used to compute the continuous movement progression and the final prediction has been made by proposing a fully connected (FC) layer. The FC layer increases the activity prediction rate by analyzing more discriminative features from the output generated by the LSTM module at a particular time module ΔT.

Feature Matrix Generation (3 Dimensional Convolutional Neural Network (3D CNN)): 2D CNN has considered the best solution to detect objects from the images by extracting spatial information. On the other hand, 3D CNN extracts spatial as well as temporal features for temporal activity modeling from the sequence of frames by performing convolution and pooling

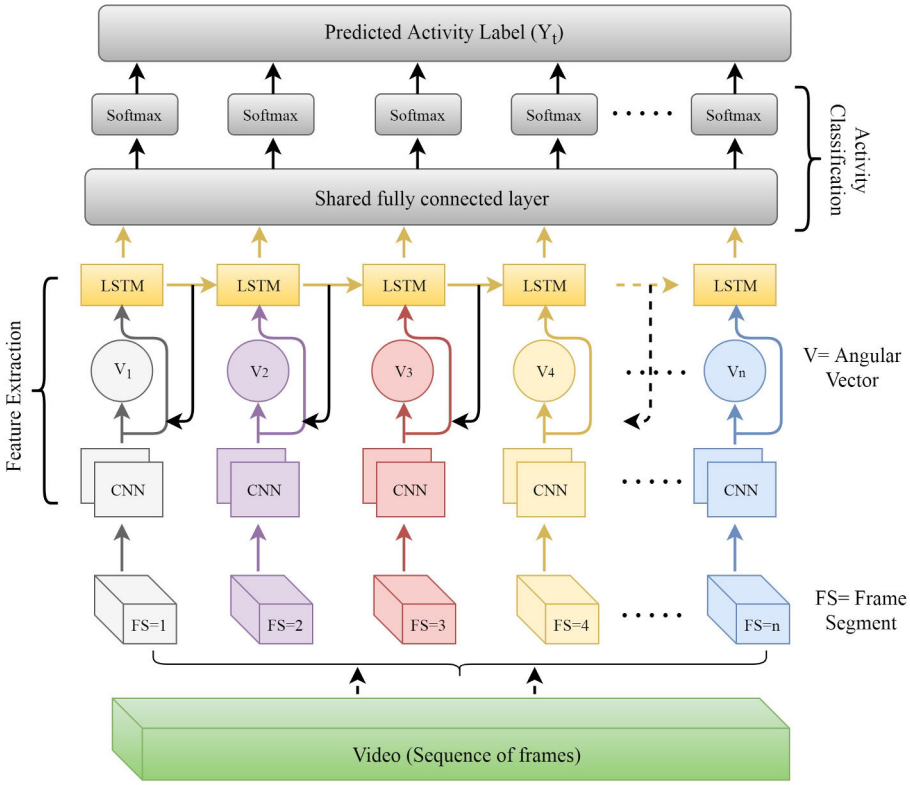

FIGURE 4.4: Proposed system for activity analyzation.

operations (Tran et al., 2015). By considering the advantages of temporal feature modeling, 3D CNN has attracted considerable attention in recent years to deal with sequence of frames (Shou, Wang, and Chang, 2016; Tran et al., 2015; Chatfield et al., 2014; Wang et al., 2017a).

Our focus is to design a 3D CNN classification architecture to detect abnormal physical stance by extracting spatio-temporal features using 3D CNN model. The whole architecture is demonstrated in Fig. **??**. Based on the previously discussed architectures of 3D CNN, the size of the convolution kernels $3 \times 3 \times 3$ has achieved best results (Tran et al., 2015). Hence, the receptive field size is fixed to $3 \times 3 \times 3$ with stride 1 for feature extraction. The frame segment size is represented as: $c \times l \times h \times w$, where the total number of channels are represented by c, l represents the depth of the frame segment, h represents the height of the single frame, and w represents the width of the

frame. These dimensions are analyzed by 3D CNN model for feature extraction. The batch-normalization technique has also been incorporated in the network to deal with problem of internal covariate shift. The sequential layer of the 3D CNN model can be described as:

$$CNN(u;W) = f_{maxpool}(\sigma(f_{BNorm}(u * W + b))) \tag{4.2}$$

Where W and b are the learning weights of the CNN network, $f_{maxpool}$ represents to max-pooling operation which is used to reduce feature map size, f_{BNorm} represent the operation of batch-normalization, and σ denotes the ReLU activation function.

Hyperparameter optimization for 3D CNN model: The 3D CNN model is initially trained for up to 40 epochs on 1×10^{-4} learning rate with the batch size of 64. The value of weight decay and momentum is set to 1×10^{-1} and 0.9. ReLU activation function is used to calculate the Non-linear complex functional mappings between the inputs and outputs. For batch normalization, the value of α, β and γ is set to 1.0, 0.0, and 1×10^{-5}. α parameter defines the scale, β represents the shift parameter of the model and γ represents the regularization parameter to provide numerical stability to the model. These parameters are trained during the training process of the model. The training process has been stopped after 25 epochs. The CNN model does not surpass one hours of training on Graphical Processing Unit (GPU).

Temporal Dynamics-Oriented Sequential Feature Modeling (Long Short-Term Memory-Fully Connected (LSTM-FC) Module): Recurrent Neural Networks (RNNs) have the ability to learn activity based temporal activity dynamics from the sequence of frames. RNN uses hidden states to map the data for sequential modeling. During the long activity representation, the RNNs is not well efficient solution to deal with long temporal reasoning. The main reason of this limitation is its similar feature representations over time. Therefore, several studies demonstrated the limitation of utilizing a CNN and RNN directly (Pan et al., 2016; Abu-El-Haija et al., 2016).

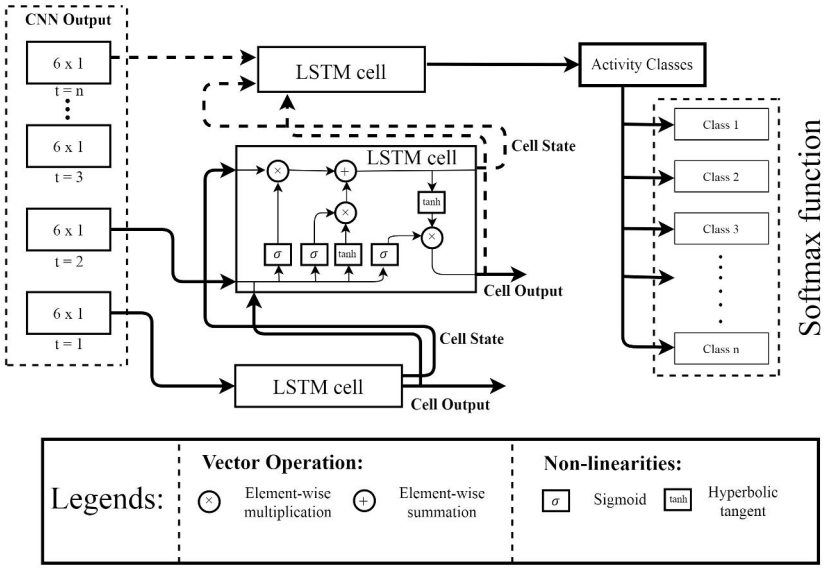

FIGURE 4.5: LSTM model for sequential activity prediction.

To overcome this limitation of RNN and to learn distinct features of an activity from the video, we have proposed a segment formulation technique by dividing the sequence of frames into multiple groups. We extract high-level spatio-temporal features from each frame segment at a particular time module ΔT from 3D CNN model. Max-pooling layers are used to pool the high-level features and transfer to LSTM network to predict the activity label as described in Fig. **??**. By transferring the spatio-temporal feature segments to LSTM cells, the feature dynamics across each frame segment has been calculated and have significantly boosted the prediction accuracy. The LSTM network-based segmental representation procedure is described mathematically to define the process of temporal dynamics-based feature modeling as follows:

$$\begin{pmatrix} i_t \\ f_t \\ o_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \end{pmatrix} W \begin{pmatrix} u_t \\ h_{t-1} \end{pmatrix} \tag{4.3}$$

$$c_t = f \odot c_{t-1} + 1 \odot tanh(W_t u_t + W_t h_{t-1}) \tag{4.4}$$

$$h_t = o_t \odot tanh(c_t) \tag{4.5}$$

where i_t for the input gate, f_t for the forget gate, and o_t for the output gate, respectively. c_t represents the current state of the cell and h_t represents the final state of the network. We use $\odot$ for element-wise multiplication and Backpropagation Through Time (BPTT) algorithm is used to learn the weights.

Hyperparameter optimization of LSTM model: After the completion of 3D CNN training process, LSTM model has been trained on 5×10^{-5} learning rate. The weights of the 3D CNN model and LSTM network is optimized using ADAM optimizer. The combined network is trained on 90 epochs. But the learning of the proposed network has been stopped upto 70 epochs when no learning accuracy was observed. The early stopping technique is used to stop the training process which helps the system to save the model from overfitting. The model took one and half hour for training on GPU. The multiple stage training process provides more efficient results as compared to the single stage model training process. Several trails have been made to increase epochs, but every time the network got over-fitted.

To decode the temporal feature vectors, the output of LSTM network is transferred to a shared two 4096-dimensional fully connected layers. The final layer defines the lower dimension of the system correspond to the number of classes of the system. The final activity label Out_t is generated by using softmax function followed by batch normalization and activation layer:

$$Out_t(u) = \sigma(f_{BNorm}(u * W_{fc} + b_{fc})) \tag{4.6}$$

4.4.3 Cloud subsystem: Activity Score Recording

Cloud layer plays a vital role to aggregate the predicted anomalous scores generated by the classification model at the edge layer as shown in Fig. **??**. Moreover, the summarized records are requested by the doctors or caretakers

to analyze the current physical state of an individual and to provide required medical services in real-time.

Definition 2: Activity Recognition Score (ARS) Activity Recognition Score (ARS) is a predicted activity score for current frame segment FS_i transmitted to the edge layer at a particular time instance [t^i, t^{i+j-1}], where j defines a total number of time instances, the activity $'A'$ can be best-recognized activity based on the current frame segment at a particular time instance t.

Activity Record Generation

The predicted scores must be stored in an adequate format which can be further used to analyze the health status. Algorithm 1 describes the complete process of activity record generation.

Algorithm 1: Physical state determination with activity record generation in the proposed methodology.
Input: *Activity scores generated by phase 2.* **Output**: *Individual's activity record related to each activity class.*
Step 1: Analyze the current frame segment and determine the ARS on edge device.
Step2: If $(ARS(i) = Network\_Trained\_Class(i))$ Then goto Step 3 else goto step 4.
Step 3: do **Step 3.1**: Set the current event record buffer. **Step 3.2**:For the calculated activity score by phase 2, IAR.add (activity type, current time stamp); Return IAR; End for
End if
Step 4: Exit.

Temporal Mining for Record Analysis:

The temporal instance based predicted results are stored in a private cloud. Since the health activities are completely time dependent, Temporal Mining technique is used to retrieve the required activity-based information based on the requested time module in a sequential time series.

Definition 3: Time instance based Frame Segment (FS) Time instance based Frame Segment (FS) t^i can be defined as an ordered list of frame-segments in a particular time module ΔT of the transmission $\Delta T = \langle FS_1 t^1, FS_2 t^2,$ ....., $FS_n t^n \rangle$.

Corollary 1.1 Time Series (TS) can be characterized by the arrangement of m predicted abnormality over a specific time module: $\{ < t^i, ARS_i >, < t^{i+1}, ARS_{i+1} >, ..., < t^{i+j-1}, ARS_{i+j-1} > \}$.

1. In **Corollary 1.1**, starting time is denoted with i and numbers of time instances are denoted as j which is used to retrieve the activity scores.

2. At particular time instance ΔT, the activity score corresponds to the currently performed activity. For example, $< t^i, ARS_i >$ corresponds to current activity score at $< t^i$ time unit.

3. Sliding window approach is used to describe the initial and last time instance. The sliding window length is pre-fixed and represented as $\Delta T = |t_{end} - t_{start}|$, where t_{end} represents the last time instance and t_{start} represents the initial time instance of the window.

During the monitoring process, various regular and irregular activities are performed by individuals. In our study, we are considering non-continuous activities based on health afflictions. So, it is imperative to analyze activity scores in sequential time pattern for effective decision making. Temporal Data Mining performs the data abstraction process to form sequential time series.

Definition 4: (Irregular Events (IRE)) Given an Irregular Activity 'IA' performed at a particular time instance t^i, then Irregular Events (IRE) is defined as the recognized irregular activity IA_i at time t^i. It is represented as: (IA, t^i).

Definition 4 provides the irregular activity score related to each activity class correlation with continuous time instance $< t^i$ of current time module ΔT_n. Otherwise, the value is reset to null. Activity scores and time stamp plays an important role in determining the individual's condition.

Temporal Granulation (TG): The temporal granulation technique provides an abstraction view of the activity record for a specific time module ΔT as shown in Fig. **??**. The final predicted activity scores are stored for future references to make smart home monitoring more effective for doctors and caretakers. The concept of temporal granule and data abstraction can be achieved by using Map and Reduce technique (Dean and Ghemawat, 2004). To process the large data in distributed cloud environment, the map reduce function is considered as a best option and now it is commercially available. The temporal granule based temporal activity logs will help to:

1. Produces the continuous information about the individual's physical activities.
2. Make efficient decisions related to the individual health by the caretakers.
3. Extract the previous scores from the record based on the requested time window.

4.4.4 Alert-based monitoring process

Edge layer is responsible for determining the physical state of an individual into two states: safe or unsafe. Unsafe physical state of an individual requires

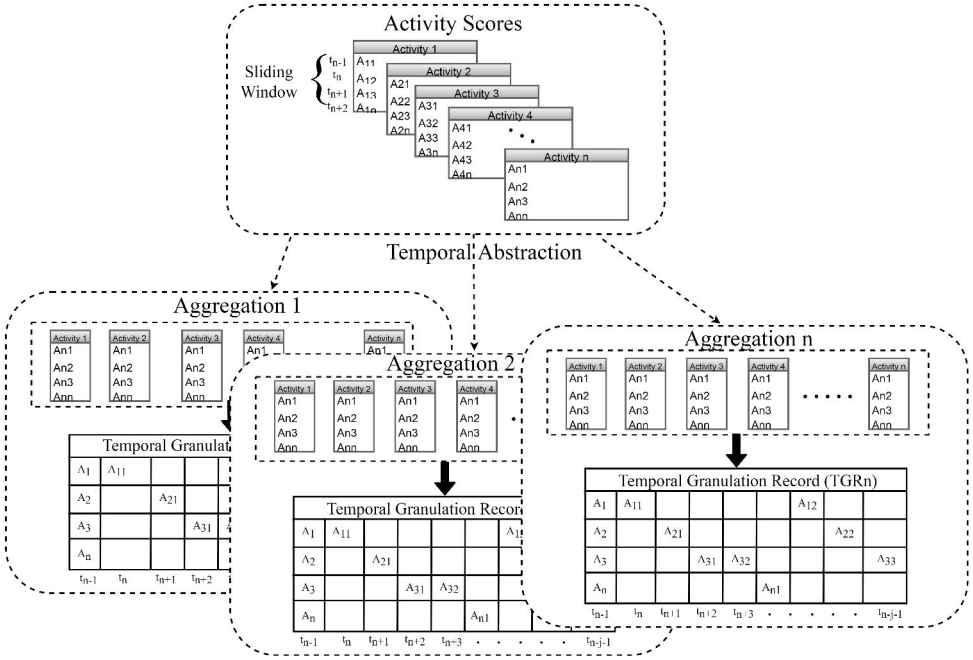

FIGURE 4.6: Temporal granule based activity record formation.

an immediate response from the medical representative. Edge node generates real-time warning or emergency alerts along with the deliverance of the log of activities to doctors and caretakers. The system provides continuous monitoring while fulfilling the time sensitive requirements of the healthcare system.

The effectiveness of an alert generation completely depends upon the efficiency and accuracy of the activity prediction module as shown in Fig. **??**. If the predicted activity belongs to unsafe class, the warning signal has to be sent to the family member or caretaker. The warning signal is generated based on the predicted activity value ARS_i is defined as follows:

$$Warning = Class\{ARS_i | i \in FS_i\} \tag{4.7}$$

where *Class* defines the type of activity classes on which the proposed deep network is trained. ARS_i is the current predicted activity score belonging to the current frame segment. On the other hand, emergency signal is

generated by calculating the health severity level as follows:

$$Emergency = IoA = \left(\frac{ARS_i}{ARS_1 \cup ARS_2 \cup ARS_3 \cup ..... \cup ARS_n} \right) \tag{4.8}$$

where ARS_i denotes the predicted activity score. ($ARS_1 \cup ARS_2 \cup ARS_3 \cup ..... \cup ARS_n$) denotes the total recognized activities in particular time module. If the Index of Abnormality (IoA) is greater than the predefined threshold, the emergency alert is generated for the medical emergency to minimize health severity. Algorithm 2 explains the warning or emergency signal based decision-making process.

Algorithm 2: Alert generation based on the performed activity. **Input**: *Activity scores generated by phase 2.* **Output**: *Current physical state of the person with the alert generation.*
Step 1: Determine the current activity value generated by the phase 2.
Step2: If (ARS(i) = (System Trained Class(i))
Step 3: do **Step 3.1**: Send the warning signal to family members or caretaker and compute IoA. End for
Step 4: if (Index of Abnormality (IoA) $>$ Pre-defined Threshold) **Step 4.1**: Real-time alerts to medical representative with the deliverance of Current Log of Activity.
Step 5: Else **Step 5.1**: No alert generation.
Step 6: Exit.

Here, the normalized threshold value is set to 0.65 to handle the health severity by providing emergency services to an individual. The probable value of the threshold is set by consulting several medical experts in healthcare environment. In addition, alert signals are delivered along with the event logs to medical representatives or specialist to initiate appropriate procedure. Moreover, an emergency signal is also generated by the proposed

mechanism if the health situation become highly severe.

4.5 Implementation detail and experimental evaluation based on resource optimization

In this segment, we evaluate the execution performance of the proposed framework to determine the response for every frame segment. The two edge nodes has been arranged and both the nodes are responsible to perform the task of feature extraction and activity analysis from captured video sequences. To maintain the simplicity, we assume that the edge nodes share a common link of wireless channel from visual sensor node to edge node. The capacity of the link from visual sensor node to the edge node is equal as the capacity of the link between two edge nodes. Both edge nodes are enabled with Intel i5-6600 processor and 24 GB memory with NVIDIA GeForce GTX 980 Ti GPU. The system is coded in Python programming language based on the Ubuntu 14.04 Operating System. The programming related to video processing such as Pillow, OpenCV and Numpy are used to perform preprocessing tasks.

Experimental evaluation based on resource optimization: Resource optimization is a critical process of achieving the required data processing capability by reducing computational cost. Under the limitation of computational power and network bandwidth, the system needs to optimize between the amount and the quality of the data. Edge node contributes to the bandwidth optimization by processing the data at local environment instead of sending the data on the cloud. Camera node significantly reduces the computation cost by downsampling the data. To understand the evaluation process more easily, The system is evaluated according to execution flow described in Fig.

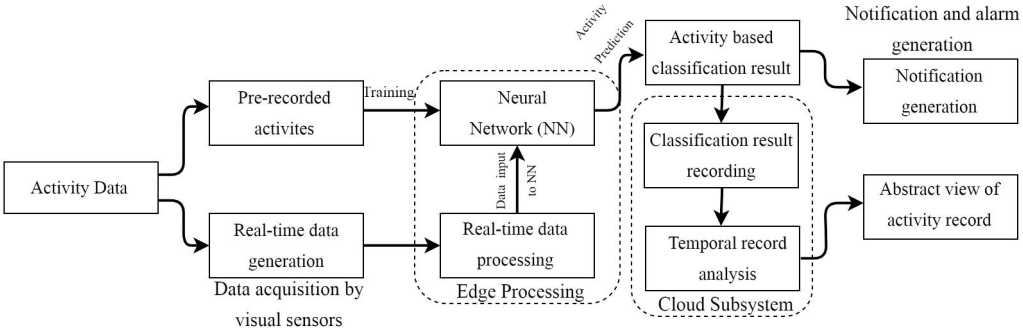

FIGURE 4.7: The procedural flow of the proposed system.

??. The performance evaluation of the system is divided into four subsections namely (1) Computational resource optimization (2) Activity Prediction Efficiency (3) Comparative performance analysis (4) Alert-based decision making efficiency with temporal granule processing and log deliverance.

4.5.1 Computational resource optimization

The visual sensor node is responsible to preprocess the sequence of frames transmitted to the edge node over a wireless channel. We measured three fundamental delay sensitive measurements: (1) The time consumption for segment delivery, (2) Task computation on edge node, and (3) the total time consumption from segment generation to final decision making under different network bandwidth ratio. Fig. **??**a presents that the delivery rate frame segments are influenced by the capacity of the network. But the segment processing time is not affected by the rate of segment delivery. Basically, we can conclude that the network bandwidth plays an imperative role to calculate the performance of the system for activity based decision making in real-time. The time taken by the system for segment delivery became a decisive factor when the network bandwidth is limited. On the other hand, the task computation time became a decisive factor during an adequate network capacity.

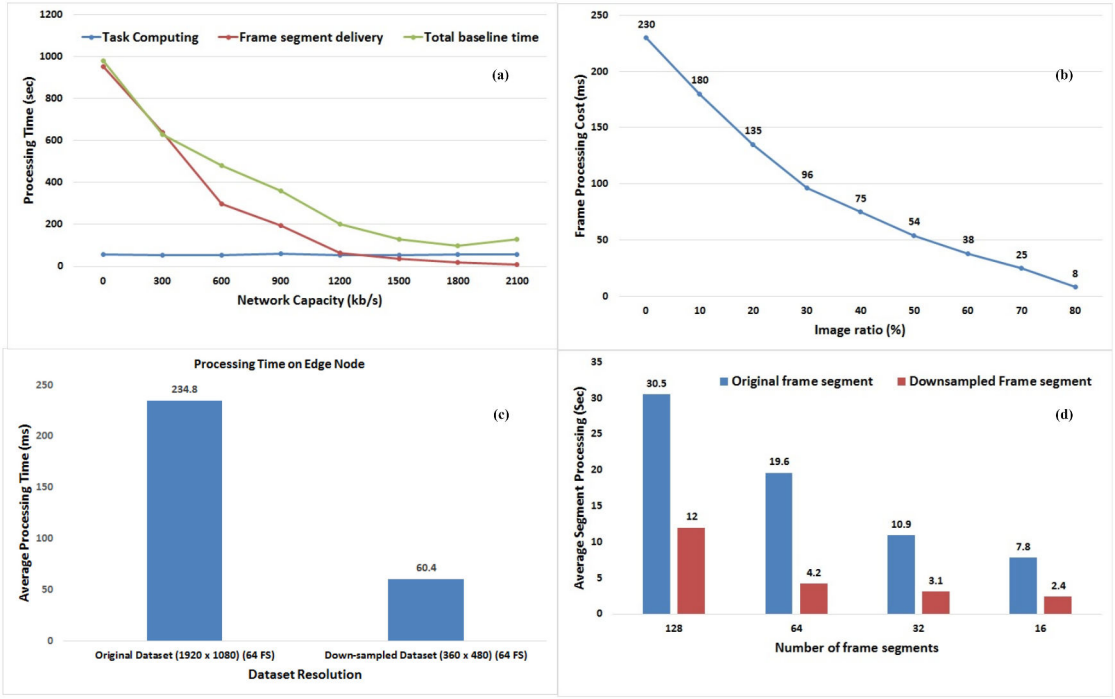

FIGURE 4.8: Processing time on Edge node.

Fig. **??**b represent the processing speed improvement on different scale ratio. It demonstrate that the downsampled data increased the data processing rate and reducing the transmission cost. Fig. **??**c shows that model can run 4 times faster on downsampled data compared to the original data. The model takes 60.4ms to process downsampled data which indicates that the system can support frame processing speed over 30fps. As visual node produces compressed frame segments, Fig. **??**d demonstrate the average segment-based batch processing cost on original frame segments and downsampled frame segments. The system has been tested on downsampled data with the resolution of 480×360, a resolution at which the model can identify physical exercises. A single batch of frame segments are loaded into memory first and then computational time has been recorded.

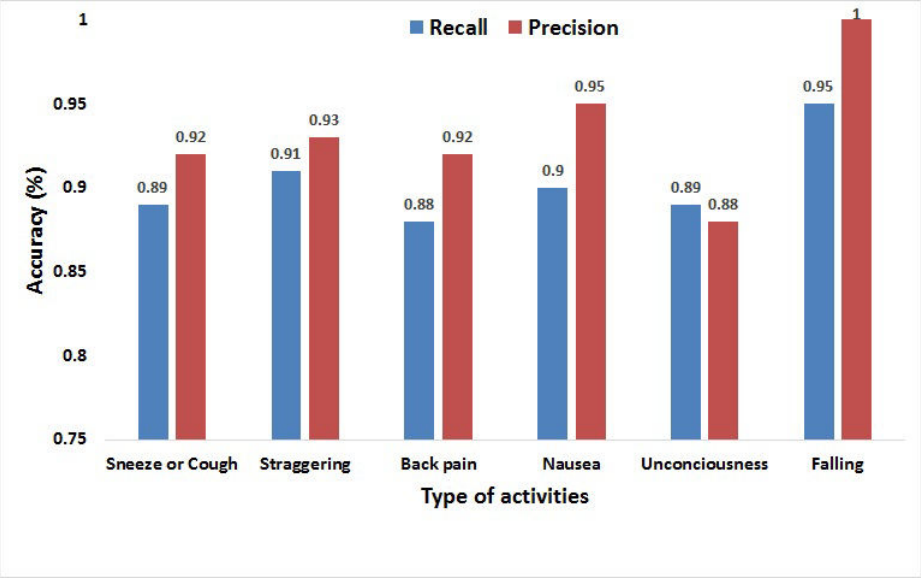

FIGURE 4.9: Classification Analysis.

4.5.2 Activity Classification Efficiency

The proficiency of the activity recognition module is evaluated by performing various experiments on testing dataset. We evaluated the efficiency of the system by manually dividing the dataset in two portions: 80% for training set and 20% for testing set. In experimental phase, the number of time steps $\Delta T = |T_{start} - T_{end}|$ are considered to give input to the model. The input lifetime ΔT for frame segment transmission is selected from three levels, i.e., ΔT = 10, 20 and 30 time steps, with equal probability of 0.5, 1.0 and 1.5 seconds.

After the completion of the training process, a sequence of results are obtained for each frame. The results demonstrate the exactness of the body posture, which depends on the system's activity class. The calculated probability of each frame defines the posture class of the system and the variability scale can be analyzed by analyzing the sequence of probability value generate for each frame. Fig. **??** shows the results of irregularity prediction for each class.

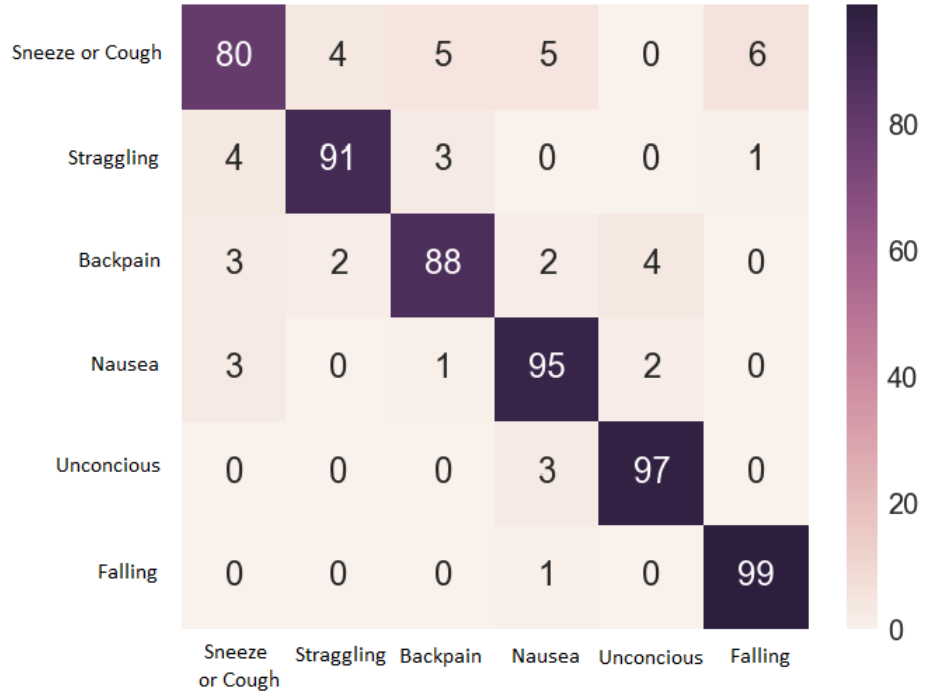

FIGURE 4.10: Confusion Matrix.

The confusion matrix is used to find the Inter-relations between the physical posture related to different activities. The Deep CNN-LSTM acquired a more reliable classification accuracy. It can be observed that, the majority of the activities are classified correctly as shown in Fig. **??**. But the major misclassifications is with respect to 'Sneeze or cough' and 'Back pain'. This is understandable as while Sneeze or Cough, there is a chance that the upper portion of the body is bent in similar motion as a person is bending in the case of back pain, thus leading to false classifications which can be solved by increasing the size of learning dataset with respect to that particular activity class.

4.5.3 Comparative analysis

To validate the system performance, we have compared the results of the proposed system with the state-of-the-art approaches using the same data.

The parameters and system configuration remains same during the implementation of comparative methodologies.

TABLE 4.3: Comparative results on captured dataset with Modern approaches

Methods	**Recall (%)**	**Precision (%)**	**Specificity (%)**	**F-measure (%)**
Weinland, Özuysal, and Fua, 2010	83.12	85.92	87.42	84.65
Tran and Sorokin, 2008	85.90	88.92	90.34	86.88
Chaaraoui, Climent-Pérez, and Flórez-Revuelta, 2013	87.46	91.12	92.27	88.56
Pehlivan and Duygulu, 2011	89.15	92.24	93.57	89.75
Proposed model	**90.33**	**93.45**	**94.28**	**90.87**

Table **??** illustrate the comparative measurements of different methodologies on the same dataset. Our proposed method achieved better results compared to the state-of-the-art methodologies by modeling temporal features extracted from 30 frame based frame segments by achieving the accuracy of 91.67%. The system have registered the high precision of 93.45% which is comparatively far better than the comparative methodologies. Specifically, the proposed classification methodology explore temporal segments and demonstrate that the LSTM cells and fully connected layers trained on fixed frame segments perform better compared to the state-of-the-arts by achieving higher recall and F-measure numerating to 90.33% and 90.87%. Fig**??** represent the resulted plots for better understanding.

4.5.4 Alert-based decision making efficiency

In real-world environment, a majority of the activities performed by an individual are normal. Based on the performed activities, a robust abnormal

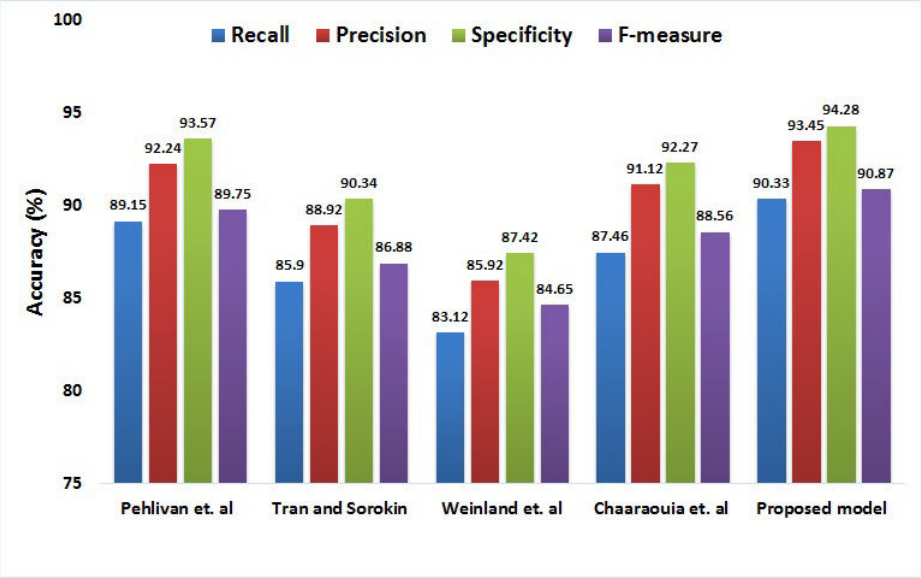

FIGURE 4.11: Classification Analysis.

activity detection method should have low false alarm rates. The principal motive of the statistics-based analysis is to evaluate the 'false positive alert' measurement to determine the true alerts based on the total number of generated alarms.

Table **??** describes the comparative analysis of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) calculated from the proposed model and comparative models. The results shows that the proposed model outperform other methodologies in activity recognition with less error rate and achieving less false positive alerts. The proposed model achieve less value of Mean Absolute Error (MAE) and Root Mean Squared Error (%RMSE) numerating to 1.52% and 4.142% which defines the high stability of the system. Low rate of error in activity classification measurements defines that, only 3.24% of alarms comes under false positive ratio.

During the alert generation, a log of performed activities is also transferred to medical representative. The deliverance efficiency of activity logs is depends upon two factors: (1) the time taken by temporal mining technique

TABLE 4.4: Comparative analysis with other methodologies

Methods	**MAE**	**%RMSE**
Weinland, Özuysal, and Fua, 2010	2.23	10.920
Tran and Sorokin, 2008	2.15	11.010
Chaaraoui, Climent-Pérez, and Flórez-Revuelta, 2013	2.02	10.224
Pehlivan and Duygulu, 2011	1.56	7.088
Proposed Model	**1.52**	**4.142**

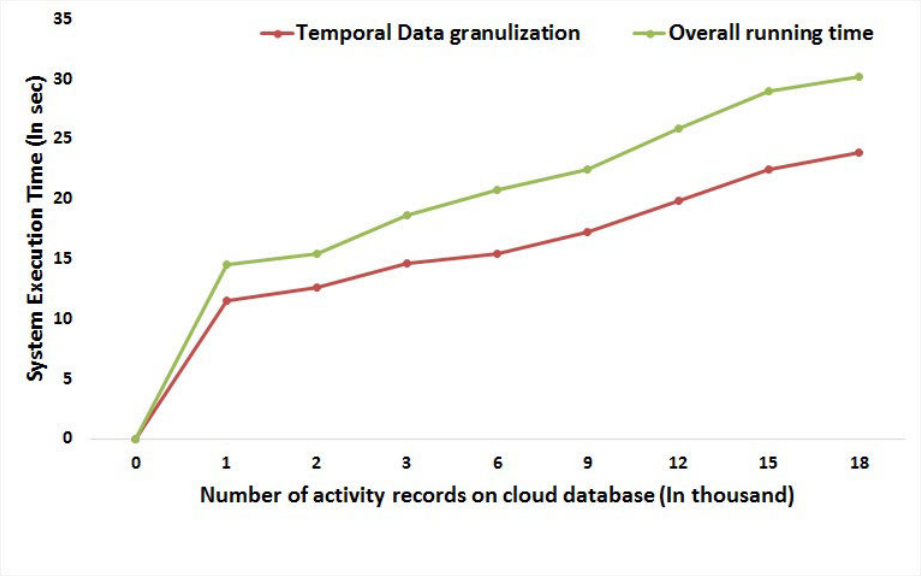

FIGURE 4.12: Decision making analysis.

to retrieve activity scores to form granule abstraction. (2) The time taken by the alert generation mechanism to calculate abnormality index in a particular time interval ΔT. The overall performance of the system is calculated based on the total time taken by the edge node to predict the performed activity with the formulation of temporal granules. Fig. **??** explains the process of log formation shows small increment when number of records are increased.

4.6 Conclusion

In this chapter, an edge analytic-based smart monitoring framework is introduced to monitor physical exercises of an individual. The main purpose

of the study is to find health afflictions in home or medical environment by combining the principles of edge computing with video analytics. The computational speed of the framework is accelerated by utilizing the GPU enabled edge devices. Particularly, three major perspectives have been considered, (i) Stance detection and activity recognition to determine health-based physical irregularity (ii) activity record generation to generate health suggestions and provide medical services, (iii) alert generation mechanism with activity log deliverance to handle various critical conditions. All these objectives are consolidated in the model to upgrade the general adequacy and utility of the system. The edge computing assisted real-time abnormality recognition mechanism has leads to a smart-assessment with better reactive healthcare. Results depict that the proposed methodology provides better activity classification rate as compared to the conventional activity classification methodologies. We have found that edge computing is far more efficient for maintaining the sensitiveness of the healthcare domain by increasing activity response rate and decreasing decision making delay compared to cloud-based platform. Therefore, it can be concluded that the proposed system is highly efficient in live monitoring by combining the several latest technologies like Cloud-of-Things, Edge-of-Things, and Internet-of-Things.

Chapter 5

Motor Movement Recognition in Smart Monitoring

5.1 Introduction

Physical inactivity causes common but severe health issues, such as high blood pressure, obesity, anxiety, and depression (Jefferis et al., 2012). It can also lead to critical health problems like diabetes, lipid disorders, osteoporosis, heart-related diseases and colon cancer. Activity recognition has turned out to be one of the developing domains of research to deal with these issues. Wearable sensors are considered as one of the most feasible and cost-effective solutions of sensing the ADL exercise of an individual in real-time (Arif and Kattan, 2015; Mannini et al., 2013). It can also be used to distinguish several irregular exercises based on the strange occurrence in frequencies (Yin, Yang, and Pan, 2008). By considering the advanced data sensing principles, severe health factors caused by physical inactivity triggers the need for an effective monitoring solution.

In recent years, Cloud-of-Things (CoT) technologies have led to many advanced applications for healthcare. Advanced and effective monitoring solutions have already changed the way of professionals to interact with patients (Shi et al., 2016). The combination of IoT sensors and cloud-based data processing techniques have also improved the reliability of delivering healthcare services. In conventional cloud-based smart healthcare systems,

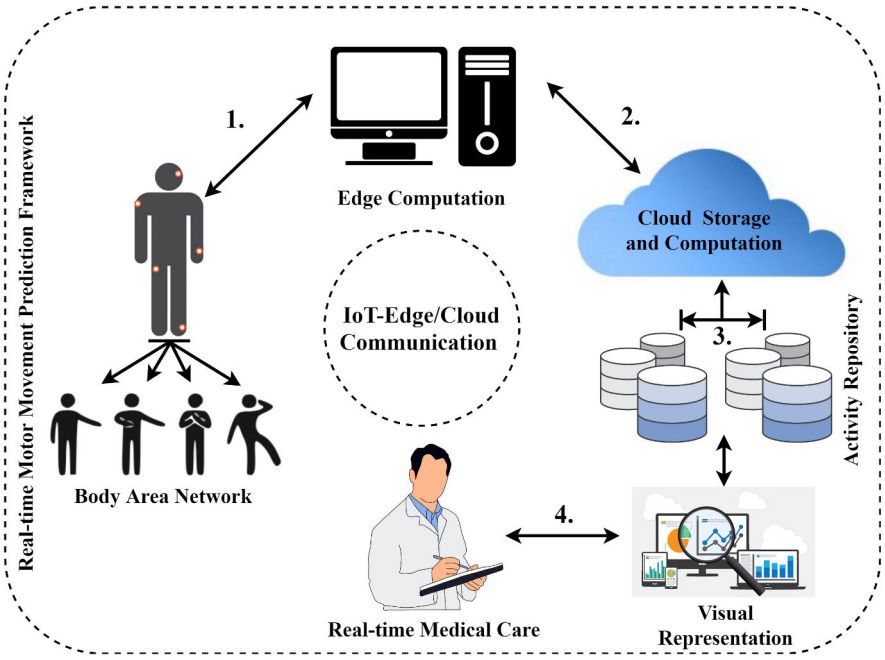

FIGURE 5.1: The element description of the proposed motor movement recognition system.

the sensed attributes are transferred to the cloud for processing causing high data processing cost with high delay in the response time. The delay caused by the these systems sacrifices the sensitivity of the healthcare domain.

Edge computing is an advanced local data processing environment that helps to overcome the challenges of high delay and also helps to reduce data processing cost. The devices used in edge computing provide sufficient storage with high computation to process data before transmitting to remote IoT servers. Even, in the Edge environment, the computational resources are also deployed close to the sensors and end devices for decreasing the latency rate during data transmission to enhance Quality of Service (QoS) (Yin, Yang, and Pan, 2008). By considering the recent successful developments of Edge computing, we proposed an Edge-assisted deep learning enabled motor movement recognition framework presented in Fig **??**.

Contributions: The aim of the proposed framework is to monitor motor

movements of the patient in real-time to analyze the physical activeness. The proposed study focuses on real-time data processing on GPU enabled high computational edge servers in the smart monitoring environment. Moreover, the proposed system recognizes the scale of physical inactivity concerning the patient. The contributions of the proposed study are divided into four objectives described as follows:

1. Physical activity-based data generation for calculating the scale of inactiveness.
2. Edge analytics-assisted deep learning-based motor movement recognition for current physical state analysis.
3. Cloud-based activity record generation for future references.
4. Routine-based physical activeness analysis and auto-suggestion generation for patients in real-time.

The rest of the chapter is explained into multiple sections as follows: In section 5.2, the literature review in context to motor movement recognition is discussed. In Section 5.3, the modular approach of the proposed system is thoroughly explained. The performance of the proposed system is validated by comparing the exploratory outcomes with other machine learning and deep learning technology-assisted approaches in section 5.4. At last, the chapter is concluded by summarizing the accomplishments of the study in section 5.5.

5.2 Related works

In this section, we have reviewed the most significant solutions in the field of IoT based physical movement recognition.

5.2.1 Physical movement-based activity recognition systems

Most recently, IMU sensors have played an imperative role in perceiving human exercises (Khan et al., 2010), (Minnen et al., 2005; Giansanti, Macellari,

and Maccioni, 2008; Narayanan et al., 2008; Marschollek et al., 2008). Majority of the activity recognition solutions are dependent on the procedure of feature extraction such as skewness, standard deviation (SD), and mean from the raw activity signals. These features are further analyzed to recognize the type of activities of an individual (Baek et al., 2004; Gjoreski et al., 2016). Majority of the studies found diverse solutions to explore single accelerometer data for activity determination (Minnen et al., 2005; Giansanti, Macellari, and Maccioni, 2008). In the article (Minnen et al., 2005), authors investigated single accelerometer data to discover recursive activities. In articles (Giansanti, Macellari, and Maccioni, 2008; Narayanan et al., 2008; Marschollek et al., 2008), authors found a machine learning-based solution to detect fall event for elders in a smart home environment. In articles (Khan et al., 2010), (Minnen et al., 2005; Giansanti, Macellari, and Maccioni, 2008), authors proposed solutions to recognize an individual's daily activities from a network of multiple sensors embedded on the human body. The solutions produced considerable activity recognition results such as laying, walking, and running etc. But in some cases, these systems misclassified some of the transitional exercises, such as, stand to sit, stand to lie, and vice versa.

5.2.2 Movement recognition methodologies

As previously discussed, the effectiveness of movement recognition approach is dependent on the capability of data processing. Most existing strategies for activity recognition embrace machine learning classifiers to recognize activities. Three basic steps are followed in every traditional machine learning-based solutions: (1) preprocessing of the raw data, feature extraction, and pattern recognition (Zhu, Chen, and Brown, 2018). Several traditional classification methods, such as Hidden Markov Model (HMM), K-Nearest Neighbor (KNN), Decision Trees, Naïve Bayes, Artificial Neural Networks (ANN), Multiple Instance Learning (MIL), Support Vector Machine (SVM), and Random Forests are widely used for model training on labelled data. The majority of activity recognition models are working on handcraft techniques.

In article (Atallah et al., 2011), authors considered the wrist of a human body is the best location to calculate the low intensities of the transitional activities. Machine learning classifiers (Bayesian classifier and k-nearest neighbor (KNN)) are used to classify the activities into multiple classes and found the best activity class related to a particular activity. In article (Berndt and Clifford, 1994), authors utilized the Dynamic Time Warping (DTW) algorithm to calculate the similarities between the signal sequences. Several studies combined the data captured from accelerometer sensors installed on different body positions and compared the classifier performance. A study proposed by (Cleland et al., 2013) used accelerometer data captured from different body locations (wrist, hip, thigh, foot, lower back and chest) to train support vector machine (SVM) classier using feature fusion method. Analyzing two different locations-based combined data provided a noteworthy activity recognition enhancement compared to single accelerometer sensor-based data.

In article (Bao and Intille, 2004), authors applied several learning classifiers on the accelerometer data collected from the ankle, hip, thigh and lower-upper arm. In article (Olguın and Pentland, 2006; Kern, Schiele, and Schmidt, 2003; Gjoreski, Lustrek, and Gams, 2011), authors have also reported significant improvement in the performance of activity recognition by consolidating at least two or more accelerometer locations. In article ("Elastic Motif Segmentation and Alignment of Time Series for Encoding and Classification"), the authors proposed an effective time-domain-based feature extraction algorithm named Elastic Motif Segmentation and Alignment (EMSA). EMSA algorithm performed several operations such as segmentation, shrinking and stretching to update the length of different time series based same motifs. In this manner, a set of synchronized subsequences are generated for feature extraction and classification.

Recently, the strategies of deep learning have got more consideration by researchers of several domains. The primary deep learning system named Deep Belief Network (DBN) was utilized to perceive patterns (Kiranyaz,

Ince, and Gabbouj, 2015). Furthermore, Convolutional Neural Networks (CNN) has turned out to be well known because of its improved discriminative power. CNN is a most popular technique of deep learning consisting the capability of feature extraction. Basically, CNN is formed by stacking convolutions in a sequence to create a hierarchy of abstract features. Several parts of CNN such as convolution, pooling, tangent squashing, rectifier, and normalization (Deboeverie et al., 2016) helps to extract features from the data and generate classification results based to the requirement of introduced application. CNN is highly efficient to analyze static patterns from the image to recognize the presence of objects. CNN is the most pertinent technique for pattern recognition from a single image instead of taking time-sequential information for determining temporal feature from the data. Therefore, Recurrent Neural Network (RNN) (Gers, Schraudolph, and Schmidhuber, 2002; Graves, Mohamed, and Hinton, 2013) is considered as the best choice for sequential feature modelling as compared to CNN and DBN.

5.3 Proposed Model

Edge computing is considered as a crucial change in the traditional cloud architectures by providing hierarchical data processing capability near to IoT layer. Edge layer helps to provide more reliable communication resources by reducing network congestion. Fig. **??** illustrates the modular approach of the proposed Edge-Cloud assisted motor movement recognition framework to recognize physical activities in real-time.

5.3.1 Data-Acquisition Stage

The initial stage of the proposed framework is responsible for acquiring sensory data related to the physical activities of the patient. Wearable sensors generate data based on acceleration, velocity, and position. Table **??** provides an overview of the dataset, type of motor movements, and devices used to

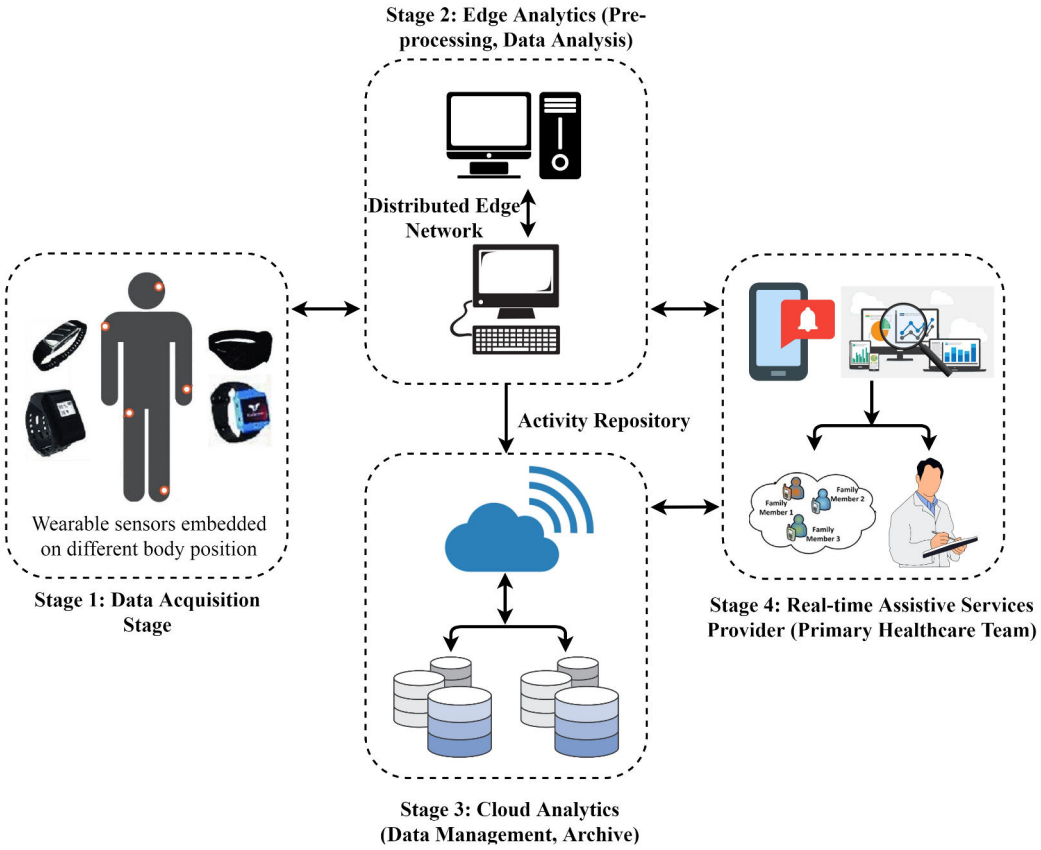

FIGURE 5.2: Modular approach of edge-cloud motor movement recognition framework.

generate activity-oriented data. Definition 1 gives an overview of sensor-based event formulation.

Definition 1: (Sensor-based Event Formulation) Let the type of movement M with the sequence of atomic movements $\{m_1, m_2, ..., m_n\}$ are collected from wearable sensors. A movement M_j belongs to a set of activity classes $A = \{a_1, a_2, ..., a_j\}$ where a_j is the total number of classes on which the system is trained.

The data is collected in sequential order and streamed to the edge node to analyze in real-time. Since the correct limits of the succession of exercises in the data stream are not known in advance, a fixed time interval technique

TABLE 5.1: Dataset attribute.

Type of Dataset	**Motor Movements**	**Wearable Devices**	**Communication channel**
Multi-Sensors based motor movements	Sitting, Standing still, Walking, Waist bends forward, Knees bending (crouching), Lying down, Frontal elevation of arms, Jogging, Cycling, Running	ActiGraph wGT3X-BT (Sterling and Laughlin, 2015)	Bluetooth (Rate: 2.1 Mbps, Band: 2.4 GHz, Distance: 20-200 m, Network nodes: 8, Security: 128 bits AES, Power: 1-100(mW)

is used to divide a continuous signal into blocks/matrix to provide input for movement recognition. The recognized activity results are available based on a fixed time interval technique.

Time-based Event Window: This approach divide sequential data in equal lengths of time modules. The viability of time-based sequence process has been acknowledged for its basic working process (Wang et al., 2012). The correct length of the window plays an imperative role in the recognition of an activity. The small window size may contain insufficient features of the activity and long size of the window may incorporate more than one exercise in a single window (Krishnan and Cook, 2014).

Definition 2. (Time Window based Activity Sample (TWAS) Formation) An TWAS is a four-tuple event window $EW = (t_s, S_s, S_m, t_e)$, in which S_s is the type of sensor that reports the atomic movement S_m. The t_s and t_e is the initial and last time instance of a particular time module ΔT. The data is captured with respect to time window $\Delta T = |t_e - t_s|$.

By definition 2, the activity samples are generated to define the temporal change of the movement. The data samples are transmitted to the edge node in a continuous manner for movements recognition.

5.3.2 Edge Analytics

In the proposed framework, the Edge layer aimed to examine physical exercises of an individual in the given time space. Edge layer provides an effective yet extensible data analysis environment. Edge nodes are also responsible for suggestion-based notification generation to provide a remote diagnosis in real-time. Table **??** illustrate several advantages of edge layer over cloud layer (Ahmed et al., 2017).

TABLE 5.2: Advantages of Edge Computing over Cloud Computing.

Parameters	**Edge Layer**	**Cloud Layer**
Distance (No. of hopes)	Single hop	Multiple hop
Service location	In the edge network	With in Internet
Jitter	Very low	High
Latency	Low	High
Geo-distribution	Distributed	Centralized
Location awareness	Yes	No
Target user	Mobile user	General Internet user
Mobility support	Supported	Limited
Hardware	Limited capabilities	Scalable capabilities
Service scope	Limited	Global
Route attacks	Very low probability	High probability

The process of motor movement recognition based on edge layer is performed into five steps as follows:

1. **Data Collection:** Data related to physical activities of the patient is collected in the initial state of the framework.

2. **Connection Establishment:** After completing the task of data collection, the connection is established to its nearby optimal edge node and the data is uploaded on the edge node.

3. **Data Analysis:** The current physical activity of the patient is recognized by the proposed deep learning approach. The proposed methodology also calculates the time difference between the previously performed activity and the activity performed at the current time instance t_i.

4. **Suggestion Generation:** The proposed framework generates time-sensitive suggestions after analyzing the scale of physical inactivity of the patient.

5. **Information Transmission:** The professionals can also retrieve the information related to the performed physical activities of the patient from the cloud to carry out an accurate medical diagnosis.

Deep learning-assisted motor movement recognition

In motor movement recognition, the nature of the data is sequential and highly dependent on the time factor (LeCun and Bengio, 1995; Plötz, Hammerla, and Olivier, 2011). Deep learning methods are considered as one of the best solutions to address the time factor. Therefore, deep learning-based hybrid approach is proposed for recognizing physical movements and for sequential modelling. The convolutional neural network (CNN) considered the input data as independent from the output and recognizes the type of movement from the given data sample. It can also be observed that the semantics of an activity can be perceived better if the complete sequence of an activity is processed before generating the final outcome. For dynamic feature modelling, Gated Recurrent Unit (GRU) network is proposed to deal with sequential movement dependency. The architecture of the proposed Multi-stage Convo-GRU model for movement recognition is presented in Fig. **??**.

Movement representation (Convolutional Neural Network (CNN)): Identifying the type of movement is considered as an initial component of motor movement representation. The CNN architecture is considered as the most suitable solution for extracting local dependency from 1D time-series based

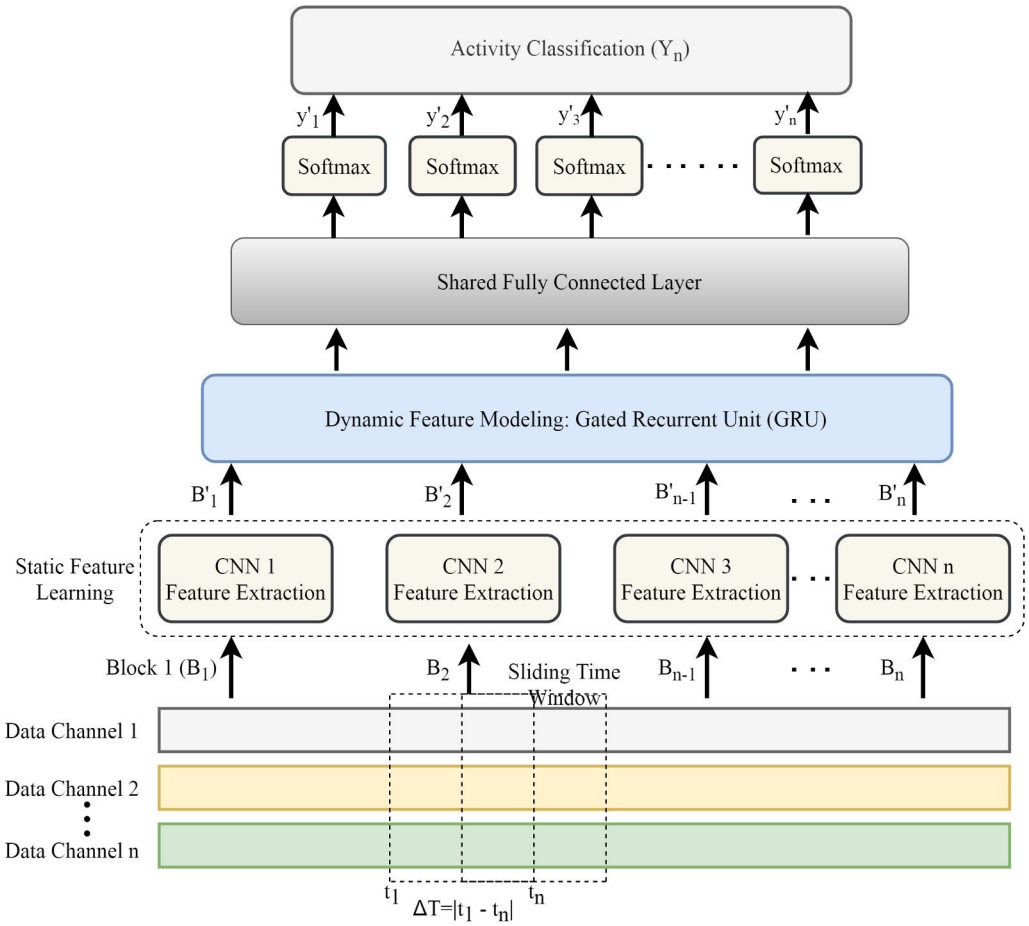

FIGURE 5.3: The proposed architecture of Deep Convo-GRU model for dynamic feature modelling.

movement signals. Inspired by the hierarchical feature extraction technique, CNN consists of convolutional layers with pooling layers to learn data features (Goodfellow, Bengio, and Courville, 2016). The process of feature modelling provides a new feature space that analyzes the raw signals over time. Filters of convolutional layers are used to extract local patterns from the data and pooling layer condenses the pattern representation. Earlier designed solutions demonstrate the effectiveness of CNN models that can separate temporal patterns from single channel (Yang et al., 2015) or between multiple channels (Chen and Xue, 2015). But long-term temporal patterns such as cycling, jogging, and climbing stairs might not be recognized effectively

utilizing CNN architecture. To overcome the constraint of CNN sequential modelling, Gated Recurrent Unit (GRU) model based combined approach is proposed.

Dynamic feature modelling (Gated Recurrent Unit (GRU)): As discussed previosuly, CNN assumed the activity samples independent on time. Hence, this assumption became invalid in the process of sequential modelling. It is expected that representing time dependency would enhance the performance for movement recognition. To achieve the proposed objective, Recurrent Neural Networks (RNNs) can be used to model the long-term patterns (Hammerla, Halloran, and Plötz, 2016). The GRU (Cho et al., 2014) with the CNN model provides more stability in the case of a lesser amount of training data. Therefore, we proposed a GRU network to calculate sequential relationships between activity samples. The reason for implementing a GRU-based model instead of Long Short-Term Memory (LSTM) network is that GRU-based network needs fewer parameters to train over LSTM and has a very less chance of overfitting (Cho et al., 2014).

The achieve the goal of sequential feature modelling for movement recognition, the GRU network takes the feature matrix X = $[x_1, x_2, ..., x_n]$ generated by CNN model as input, where feature matrix x_i represents the current activity state. On the other hand, x_n represent to the state of an activity which is used to provide input to the initial cell of GRU network at a particular time instance t_i of the current time module ΔT. The GRU cell uses two gates (Reset gate, Update gate) to decide the flow of the data for final output generation. A reset gate r_t is responsible to decide the sensitivity of the information.

$$r_t = \sigma(W_r x_n + U_r h_{t-1} + b_r) \tag{5.1}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{5.2}$$

where W_r and U_r are trainable weights of the GRU network. An alternate cell state c_t is generated by combining $r_t \circ h_{t-1}$ with the current input state value x_t. The symbol $\circ$ represents elementwise multiplication.

$$c_t = tanh(Wx_t + (U(r_t \circ h_{t-1}) + b) \tag{5.3}$$

$$tanh = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{5.4}$$

If the value of r_t gate is close to 0, most of the previous cell state h_{t-1} is forgotten by the alternative cell state c_t. After forgetting the value of alternate state c_t, update gate u_t determines the formation of an actual state h_t by combining the state of alternative cell c_t with the previous cell state h_{t-1}.

$$u_t = \sigma(W_u x_t + U_u h_{t-1} + b_u) \tag{5.5}$$

$$h_t = u_t \circ h_{t-1} + (1 - u_t) \circ c_t \tag{5.6}$$

The final state h_t is transformed with a fully-connected layer. The dimension of the hidden layer matches with the number of classes of the physical movements. To generate final output (O_t), Softmax function is used.

$$O_t = softmax(W_o h_t) \tag{5.7}$$

where W_o represents the weights of the final full-connected layer. The cross-entropy error calculation function is adapted to measure the loss between calculated result O_t and actual result P_t for time ΔT as follows.

$$L < O_t, P_t >= -\sum_{i=1}^{m} O_t \log(P_t) \tag{5.8}$$

After calculating the error loss, an optimizer named ADAM stochastic optimizer is utilized to optimize the weights of the hidden nodes by determining the learning rate during the process of backpropagation (Kingma and

Ba, 2014). BackPropagation Through Time algorithm (BPTT) (Werbos, 1990) is used to update the weights of the network by adjusting the learning rate.

5.3.3 Cloud Analytics:- Data Management, Archive

The Edge layer analyzes sensor data to determine the type of motor movement. Due to the limited storage space, edge nodes are not competent enough to store all the performed and recognized events for long-term. The cloud layer provides a suitable platform to store all the recognized results generated by the edge nodes to be used for further medical analysis and diagnosis. It incorporates two components: 1) Data Storage and 2) Data Retrieval. The cloud layer stores the overall information about the individual under observation including medical history e.g., age, past illnesses, prescribed medications etc. The sensor data and the calculated movement scores are sent to cloud database by utilizing either Message Queue Telemetry Transport (MQTT) or Hyper Text Transfer Protocol (HTTP). Based on the request made by the doctor or caretaker, the cloud stage retrieves the patient information related to the requested time module and transmits it for further examination.

Data Management

The primary goal of data management layer is to store and managed all the recognized results in a common format that can be further utilized to provide health precautions. The detailed procedure of information storage on the cloud is mentioned in Algorithm 1 which is further diagrammatically explained in Fig. **??**.

Definition 3: (Regular Motor Movements (RMM)) Considering an activity A performed at a particular time instance t_i, then Regular Motor Movements (RMM) are defined as a value calculated by the edge node for activity A at time t_i is represented as: (t_i, A).

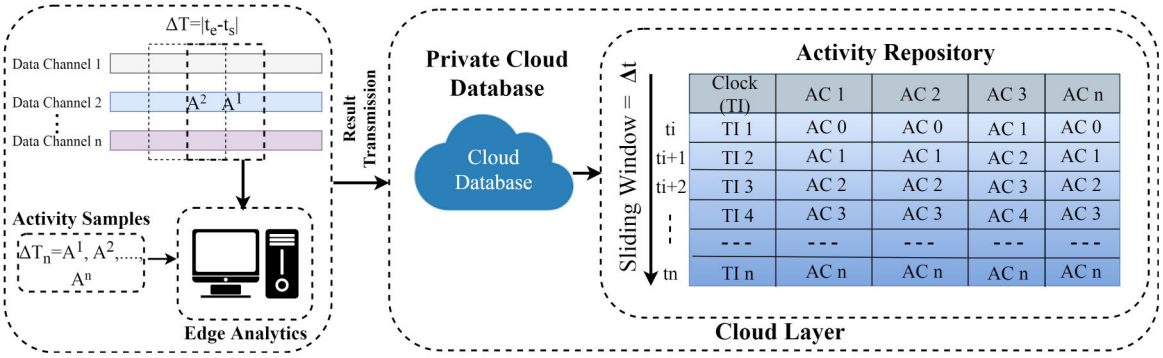

FIGURE 5.4: The process of Cloud-layer based Information Storage.

1. Definition 3 gives an immediate result related to each physical movement performed by the individual.

2. Similarly, different standard motor movements are associated with a particular time module.

Algorithm 1: Cloud-layer based Information Storage
Input: Time Window (TW) ΔT_i based Data Sample
Step 1: For ($t_1 \rightarrow \Delta T$ [Time Window]).
Step 2: Determine the type of movement performed by the patient in a specific Time instance (tn).
Step 3: If (Motor Movement = Network_class(i))
Step 4: Information Storage: **Step 4.1:** Generate checkpoint to save recognized results at a time window δt. **Step 4.2:** Release cloud space by deleting previous logs **Step 4.3:** Result storage in private cloud by following sliding window approach
Step 5: Exit

Data Archive

The process of retrieving a specific requested time module-based patterns from the cloud is known as information mining. It becomes necessary to extract information in a sequential time format for effective health analysis. Time series based data abstraction technique is performed by Temporal Mining to formulate Temporal Activity Logs (TAL). The transient time definition-based mining process is absolutely dependent on temporal mining technique (Sacchi et al., 2007).

Temporal Activity Logs (TAL): TALs formulation becomes imperative to analyze the day-to-day physical performance of the patient in real-time. In a requested time module (ΔT) for data abstraction, TALs represents the subportion of recognized movement results which are considered for the assessment of the patient. In simple words, the temporal activity log comprises of activities which are further considered by doctors or caretakers for medical supervision. The process of TAL formation is represented in Fig.**??**.

Definition 4 (Temporal Activity Logs Granule (TALG)): Given TALG function, the Temporal Activity Logs Granule (TALG) is a subset of recognized movement results from an activity space A in a requested time window ΔT. TAGL is represented in the form of a tuple as follows: $[< T_s >< TAGL >< \Delta T >< m_{n-1} >, < m_n >, < m_{n+1>} < T_e >]$

In addition, the measurement of the time window (ΔT) can be changed relying on the application domain. Mathematically, Temporal Activity Logs (TAL) are represented as TAL $\{(m_1, m_2, ..., m_n)\ \Delta T \subset$ Activity Space$\}$. The process of result storage in private cloud database is demonstrated in Fig. **??**.

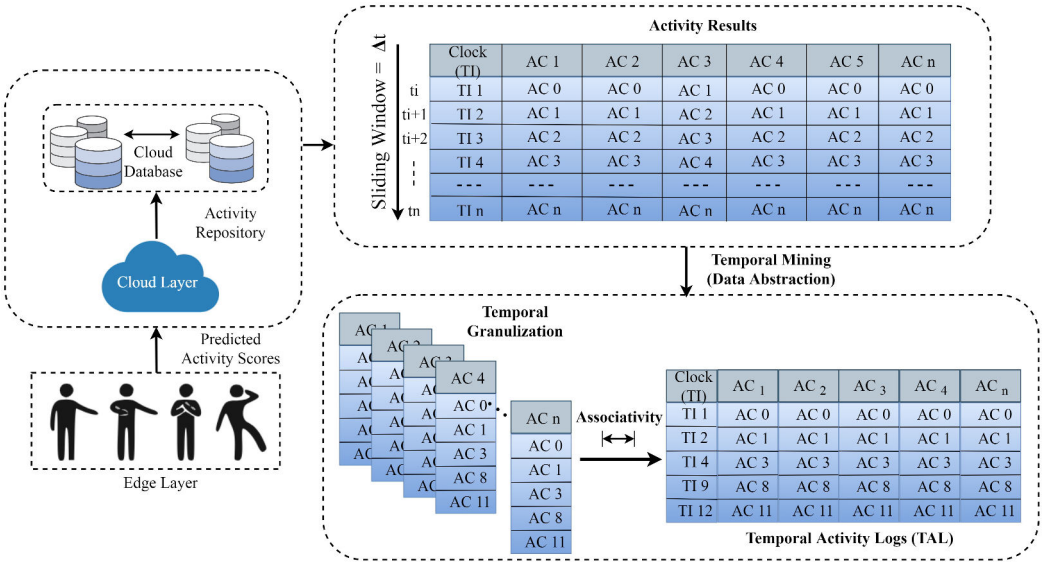

FIGURE 5.5: Temporal Mining based Temporal Activity Log (TAL) formulation.

5.3.4 Real-time Physical Inactivity-based Suggestion Generation: Primary Healthcare

In smart healthcare, real-time notifications have a crucial role for medical representatives and caretakers to notify about the present status of the patient. Any interruption in the generation of notification may cause serious consequence for patients. Compared to cloud-assisted frameworks, the edge-assisted framework has limited media and communication channels to generate notifications. The main advantage of Edge-based notifications system is, it acts independently over the local network or other mobile services like GSM to generate and deliver notifications. It enhances the reliability of the framework for smart healthcare domain. In addition, data processing at the network edge can also limit the traffic among remote servers and cloud which helps to reduce the delay factor in decision making and latency in notification delivery. The proposed mechanism of suggestion-based notification generation follows two phases:

1. In the initial phase of the notification generation, multi-scale Convo-GRU Model-based physical inactivity calculation.

2. In the second phase, suggestion-based notification generation along with the logs deliverance.

The proposed two-stage mechanism of auto-suggestion generation empowers specialists to analyze data from a more extensive point of view. In addition, the transmission of activity logs in real-time also enhances the efficacy of decision making. The proposed method of suggestion generation is referenced in Algorithm 2.

Algorithm 2: Edge Analytics based suggestion generation with Temporal Activity Log (TAL) deliverance
Input: Current Time Window (TW) ΔT based Activity Sample.
Step 1: Extract Temporal Activity Log (TAL) related to the requested time module ΔT.
Step 2: Do; **Step 2.1:** TAL storage in local memory of the edge node
Step 3: Calculate aggregated value of the performed activity related to the current stored TAL.
Step 4: If (Aggregated_Activity_Score $\leq$ Threshold)
Step 5: Do; **Step 5.1:** Edge node generates the suggestion based notification to the patient, caretaker and relative doctor in real-time. **Step 5.2:** Transfer Temporal Activity Log (TAL) based on the requested time instance ΔT_n to a respective medical representative.
Step 6: Loop Step 1 after each distinct time interval.
Step 7: Exit

In algorithm 2, the framework continuously analyze the the scale of physical inactivity for random time module. A suitable estimation of threshold has been determined by the well-being expert to investigate the required

scale of physical activeness of the patient. The edge-based computational devices also stored the calculated value of physical inactiveness for further transmission.

5.4 System evaluation and performance analysis

iFogSim (Gupta et al., 2017) simulator is utilized for simulating the proposed environment to evaluate the feasibility of Edge-based movement prediction solution. For experiments, 3D data is captured by installing wearable sensors on the subject's left ankle, chest, and right wrist. Total 10 activities such as standing, sitting, walking, lying, waist bends forward, climbing, knees movement, front arms elevation, jogging, jumping, cycling, and running are considered for monitoring purpose. The activities are performed in and out of the lab environment. The captured dataset is tested on different window sizes to analyze every possible physical situation. The recognition performance of the proposed framework is also validated on MHEALTH Dataset (Banos et al., 2014; Banos et al., 2015) to justify the capability of movement recognition. IoT-based application development platform named Ubidots (Karumbaya and Satheesh, 2015) is utilized to generate and transmit notifications to doctors and caretakers at the edge layer. The labelled dataset related to each raw acceleration signal is recorded on the 50Hz sampling rate. Several experimental scenarios are performed to determine computational and movement recognition performance of the framework as follows:

1. Quality-of-Services based quantification
2. Movement prediction performance analysis
3. Movement recognition performance validation on publicly available dataset

5.4.1 Quality-of-Services quantification

The performance of the proposed framework is justified in term of energy usage, network delay, and network bandwidth cost. Furthermore, the performance of the proposed framework is validated by comparing with the Cloud-based solutions. For general applicability, the distribution of services in term of Edge-Cloud integration has also been exhibited under an alternate number of sensors.

- Determination of interoperability
- Overall QoS quantification

Interoperability determination

It became imperative to determine the distribution of the service among Edge and Cloud to achieve Edge-Cloud based interoperability in the healthcare domain. The incorporation of Edge-Cloud in service-based load distribution has been discussed in this sub-section.

Number of sensors-based resource distribution: Fig. **??** shows the service distribution based on the number of sensors in Edge-Cloud assisted data processing environment. By analyzing the plots of Fig. **??**, it can be observed that the edge-based computational resources can handle sufficient number of sensors associated with the application. By expanding the number of sensors, the services in Edge additionally increases. By achieving the considerable number of services with explicit rate of CPU usage of edge servers, the applications are required to send to Cloud for further administration. Therefore, we can observed that Edge-based services become steady while the number of services on cloud start to increase.

Overall QoS quantification

To evaluate the resource-based overall QoS, we have concentrated on the resource utilization rate of the proposed framework in the edge-based data

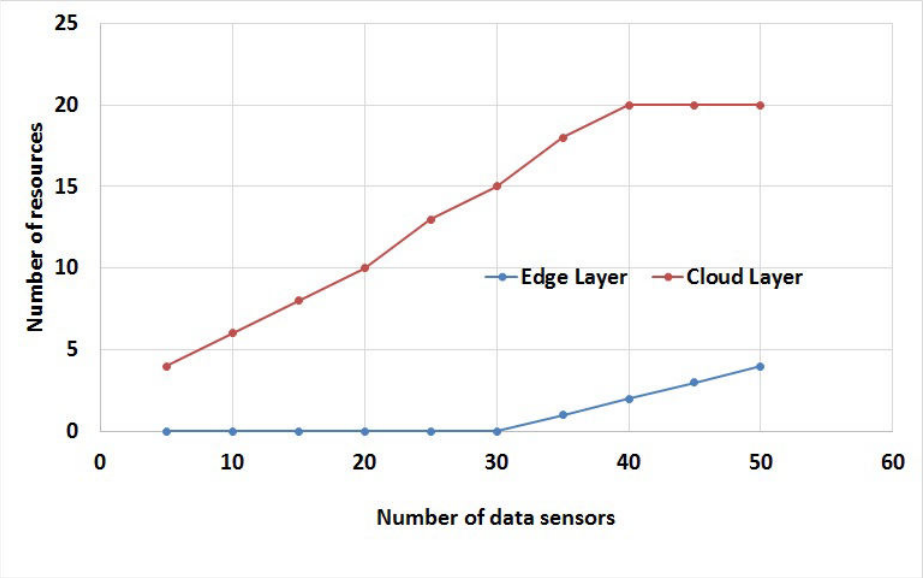

FIGURE 5.6: Integration of Edge-Cloud varying number of sensors.

processing environment. We evaluate the degree of resource optimization as follows:

Delay rate analysis: Plots in Fig. **??**(a) represents that sharing the same communication link by multiple healthcare applications in remote Cloud-based solution creates network congestion, higher data round-trip time, and reduces the bandwidth segment. Thus, an average rate of delay in the network perceived by the applications turns out to be high in the cloud. On the other hand, the average rate of delay for information accessibility is low in Edge-based solutions. The primary reason for less amount of delay in edge-based solutions is the existence of various correspondence communication links between the source and computational data processing servers. In this manner, edge computing-based solutions help to maintain the flow of the data diminish the rate of network delay.

Rate of energy consumption: In Cloud-based solutions, a solitary Virtual Machine (VM) is responsible to execute an application. However, in

Edge-based solutions, various Micro Computation Instances (MCIs) collaboratively execute an application. An MCI is a lightweight data processing component compared to VM and consumed less amount of energy for application execution compared to the cloud. Therefore, Fig. **??**(b) represents that the rate of energy consumption of MCIs while executing a number of applications is not more than VMs.

Instance cost analysis: Fig. **??**(c) represents the difference in between the cost of instance processing on edge layer and cloud layer. The plots represents that the cost of instance processing in Edge-based solutions are very less as compared to Cloud-based solutions. Resource provisioning is also possible in Edge-based solutions by paying according to the context of the module. In Edge-based solutions, the instances of the application are processed by MCIs. On the other hand, in Cloud-based solutions, VMs are predefined to process either a complete application or a single instance of an application and required to pay whole for services.

Bandwidth utilization: By analyzing the plots of Fig. **??**(d), it can be observed that the edge technologies contribute to save the network bandwidth consumption by bringing the local services close to the customer and helps to build a sustainable IoT-based healthcare infrastructure.

5.4.2 Movement prediction performance analysis

The performance of the proposed movement prediction methodology is validated as follows:

- The best hyperparameters for the system performance,
- Overall efficiency of motor movement recognition,
- Overall system throughput time on the network edge,
- Decision-making efficiency based on temporal granule formulation with deliverance to end user.

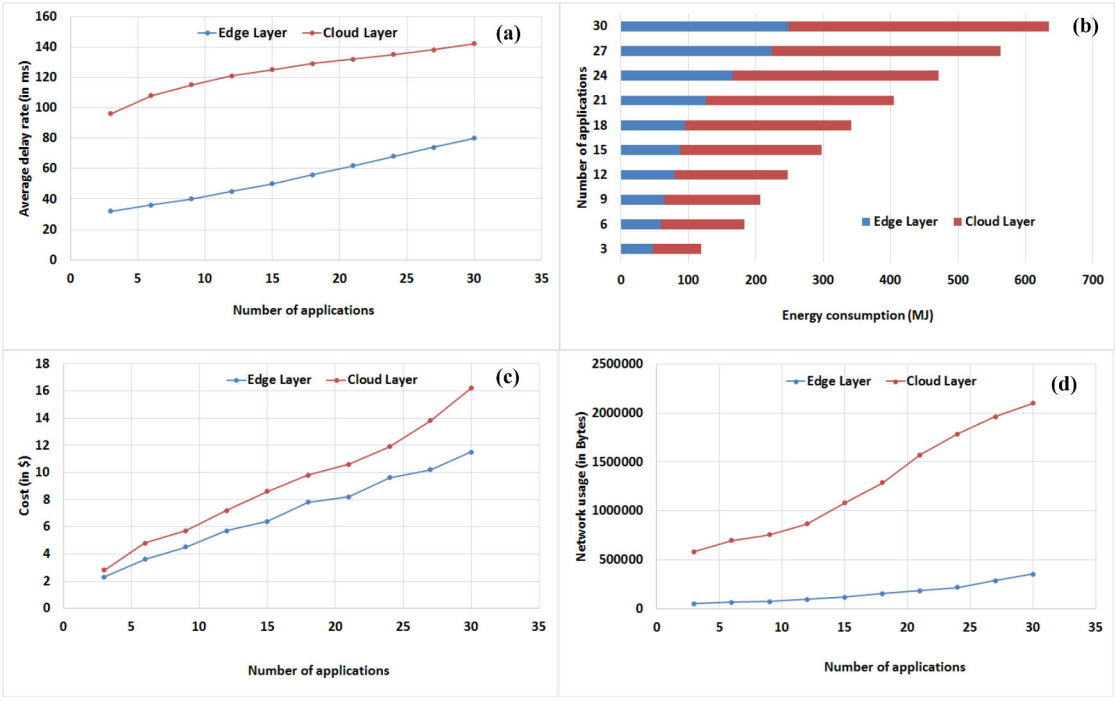

FIGURE 5.7: Overall QoS quantification; (a) Delay rate analysis, (b) Rate of energy consumption, (c) Instance cost analysis, (d) Bandwidth utilization.

Hyperparameter selection of Multi-stage Convo-GRU model

In this phase, we have incorporated a Greedy Tuning approach to train the hyperparameters of the proposed approach for selecting the most suitable configuration for movement recognition. The efficiency of the feature extraction is calculated by adjusting convolution layers, filter size, pooling size, the number of feature maps. The best hyperparameters of the CNN model is selected by adjusting 4 layers with the incorporation of the learning rate, max-pooling, and padding to convolve the data to extract features. Total 250 epochs with 0.001 learning rate are set to train the system by following the early stopping criterion after 150 epochs to halt the process of training after analyzing the low error rate. The model with the low error rate is saved during the validation for the testing phase.

Number of layers: The best-selected hyperparameters for the proposed

deep learning model are described in Fig. **??**. A considerable steady increment in the performance of the CNN with layer 1, layer 2 and layer 3 has been observed during the validation stage of the system. The increment in the accuracy of validation data is much smaller in layer 3 and layer 4. Furthermore, we observed that after adding layer 4, the model starts to decrease its performance compared to layer 3. So, we selected three layer CNN architecture for feature extraction.

Number of feature maps and size: For the configuration of layer 1, layer 2, and layer 3, the feature maps (F) are set to 90, 110, 130, and 150 and achieve the accuracy of 83.88%, 88.97%, 90.67%, and 90.44% respectively as shown in Fig **??**(b). After analyzing the graph **??**(b), it can be concluded that the accuracy of the model does not increase after applying 130 feature maps. Furthermore, the calculated results should be generally consistent for each input data by adding each layer. After finalizing the number of layers with feature maps, we determined the system performance by changing the size of the feature maps. Fig. **??**(c) shows the performance of the model by testing on 1×5 to 1×10. As we can see, filter size from 1×8 to 1×10 provide the best recognition accuracy on the test dataset.

Size of pooling layers: Fig. **??**(d) demonstrates the impact of pooling size on CNN model. Dissimilar to the filter size, pooling size does not affect the performance of the CNN model. Over the multiple runs, pooling size of 1×3 is selected as sufficient to perform pooling operation.

Finalized hyperparameters: After finalizing the size of pooling layer, the present best hyperparameters for CNN model is as follows: Layers = 3, Feature maps = 130, Filter size = 9, and Pooling size = 3, and the model produce the accuracy of 90.68% on the test set. By adding 2 fully-connected layers of 1024-nodes with the selected CNN model improves the accuracy on the test dataset by 1.03%.

Number of GRU units: After the completion of CNN training process,

the fully connected layers are replaced with a proposed network of GRU units for dynamic feature modelling. Moreover, Fig. **??**(e) represents the overall accuracy of movement recognition by varying number of GRU units within the GRU cell. We observe that the GRU model with N = 64 number of units in the GRU cell improves the movement recognition performance of the system.

Number of units in hidden-layer: After finalizing most suitable hyper-parameters of the proposed methodology, we evaluated the influence of the last fully connected layer on the movement recognition performance. To get the best recognition precision, an alternate number of nodes are selected and the recognition performance is represented in Fig. **??**. By adding 700 nodes in hidden-layer, the efficiency of the 91.51% has been achieved for the F-measure measurement, which is considerably better as compared to existing state-of-the-art techniques. Using 2000 hidden nodes the accuracy reaches up to 91.81%. The outcomes demonstrate that the recognition performance may be additionally improved by increasing the hidden nodes. The system has been trained for 150 epochs. As we got the best motor movement recognition rate between 115 to 148 epochs, we conclude that it is optimal to stop the model training process at 150 epochs.

The explanation behind getting comparative precision for the selected size of kernels is directly proportional to the type of the dataset. The number of test samples contained various type of exercises. As we know, the scale of movements is dynamic and also varies with person. Therefore, the selection of a particular size of a time window became imperative to aggregate the type of movement with data segment. It can also be observed that the precision based on the size of kernels are also varied with activity. By considering the prespective of resource utilization with maximum accuracy, the tuning of the size of kernal is possible. Moreover, the quantization strategy has been applied to migrate the proposed multi-scale Convo-GRU model to IoT-based edge servers (Lai, Suda, and Chandra, 2018). Through these conducted experiments, the CPU/GPU-based data processing load, execution time limits

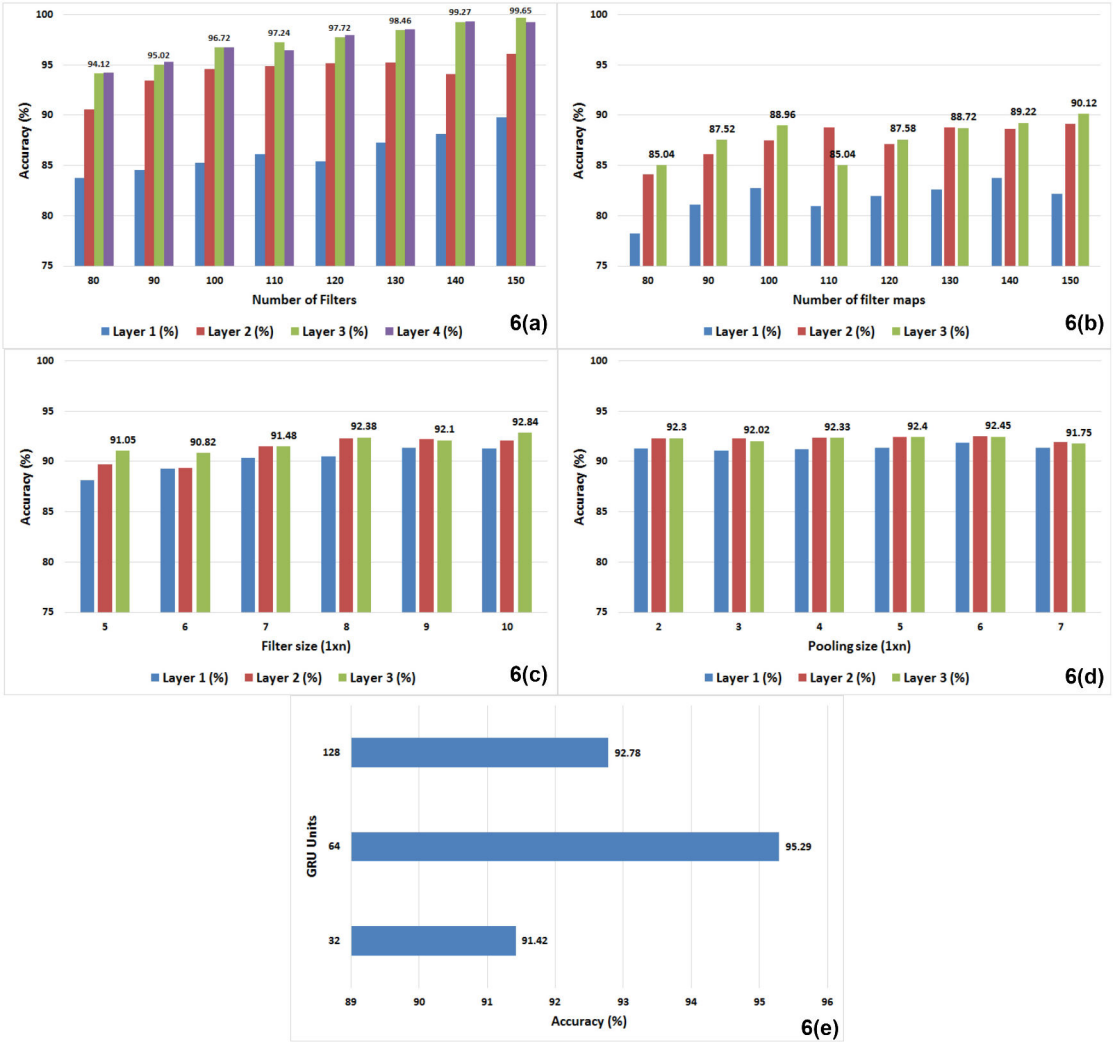

FIGURE 5.8: Deep Convo-GRU-based best hyperparameter selection: (a) number of CNN layer, (b) number of filter map in convolution layer, (c) size of filter maps, (d) size of pooling layer, and (e) number of GRU units.

has been recorded. These calculated results validate the principles of optimal load distribution.

Overall accuracy of motor movement recognition

The performance of the proposed model is validated by comparing it with other state-of-the-art machine learning and deep learning models, such as

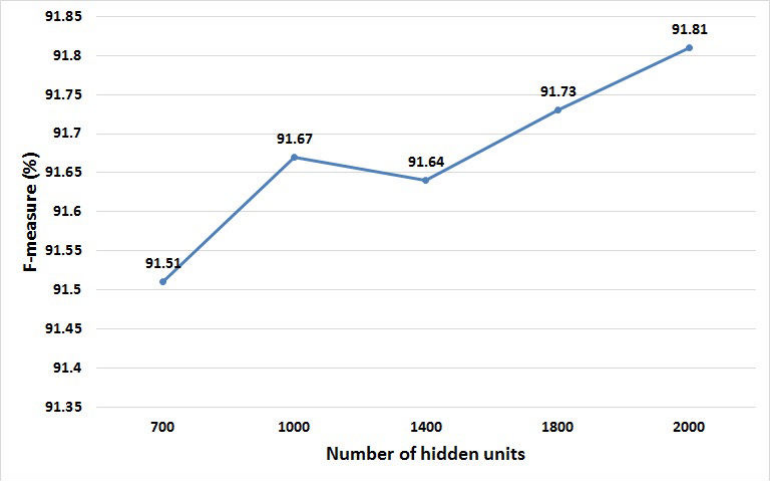

FIGURE 5.9: Multi-scale Convo-GRU model performance with different numbers of hidden nodes.

EMSA, HMM and DBN. The recognition of each classifier is assessed by calculating Precision (exactness), Recall (review or completeness), and F-measure. Following equations are used to calculate the efficiency of each class.

$$Precision = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)} \tag{5.9}$$

$$Recall = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)} \tag{5.10}$$

The average combination of both recall and precision can be estimated by utilizing F-measure parameter.

$$F - measure = \frac{2 * (Precision * Recall)}{Precision + Recall} \times 100\% \tag{5.11}$$

Table **??**, **??**, **??** and Plots in Fig. **??** illustrate the average performance of state-of-the-art classifiers and proposed recognition model across 10 activities. Based on the calculated outcomes, it can be concluded that the proposed movement recognition solution has performed better over other state-of-the-art solutions. The proposed deep learning approach is able to accomplish better precision by numerating to nearly 94.24% compared to EMSA with

87.18%, HMM with 81.07%, and DBN with 83.56%. In the case of recall, the proposed model has also acquired 89.58% which is comparatively better to EMSA with the accuracy of 82.52%, HMM with 76.28%, and 78.48% for DBN model. At last, the value of F-measure (91.85%) is also achieved by the proposed model is better comparing to EMSA, HMM and DBN. Subsequently, we can conclude that the proposed deep learning approach is highly efficient to recognize dynamic motor movements.

TABLE 5.3: Precision classification scores for each activity.

Activities	**EMSA (%)**	**HMM (%)**	**DBN (%)**	**Proposed (%)**
Sitting and relaxing	92.24	88.34	89.62	99.92
Standing still	75.25	68.24	70.02	88.24
Walking	75.72	71.23	73.58	85.26
Lying down	94.5	85.49	87.23	98.74
Waist bends forward	87.12	83.65	80.24	87.42
Knees Movement	95.23	87.21	89.62	96.84
Frontal arms elevation	94.02	87.82	91.12	98.82
Jogging	88.42	80.12	84.28	88.57
Cycling	88.62	83.24	88.23	98.26
Running	80.02	75.92	77.24	100
MEAN	**87.18**	**81.07**	**83.56**	**94.24**

The confusion matrix (Table **??**) of the proposed method described the overall accuracy of the system in motor movement recognition for the test data. It can be observed that one of the significant misclassifications is concerned with 'jogging' and 'running'. This misclassification can be easily explained by the fact that the feet of the individual is moving in a pattern similar to running while jogging. However, these misclassifications are not so critical. In future, we intend to improve the proposed movement recognition model for eliminating such misclassifications.

System throughput time on edge node

We have summarized a complete classification throughput of the proposed technique in Table **??**. When the recognition time of the proposed model is

TABLE 5.4: Recall classification scores for each activity.

Activities	**EMSA (%)**	**HMM (%)**	**DBN (%)**	**Proposed (%)**
Sitting and relaxing	90.27	83.24	82.36	92.23
Standing still	70.42	62.36	65.24	86.54
Walking	73.74	68.24	71.24	83.88
Lying down	90.25	81.24	83.25	93.54
Waist bends forward	87.12	80.45	78.26	85.24
Knees Movement	87.25	81.68	84.57	92.38
Frontal arms elevation	83.54	78.46	81.06	95.67
Jogging	82.36	77.85	82.51	82.57
Cycling	83.72	80.23	84.02	92.65
Running	74.52	70.69	75.67	89.52
MEAN	**82.52**	**76.28**	**78.48**	**89.58**

TABLE 5.5: F-measure classification scores for each activity.

Activities	**EMSA (%)**	**HMM (%)**	**DBN (%)**	**Proposed (%)**
Sitting and relaxing	93.52	81.24	82.42	96.49
Standing still	65.87	60.28	64.72	83.54
Walking	61.02	66.58	68.29	85.26
Lying down	92.57	85.94	83.78	98.70
Waist bends forward	88.62	88.71	84.26	89.28
Knees Movement	92.54	79.25	88.29	96.56
Frontal arms elevation	98.24	85.58	87.85	98.72
Jogging	86.78	88.84	85.28	88.41
Cycling	87.52	78.24	84.36	90.65
Running	80.25	68.26	78.26	87.54
MEAN	**84.75**	**78.52**	**80.94**	**91.85**

TABLE 5.6: Processing time (in seconds) per inference.

Methods	**Feature Extraction + Classification**	**Total (seconds)**
EMSA	8.104 + 0.148	8.252
HMM	11.057 + 0.084	11.141
DBN	9.027 + 0.045	9.072
Proposed System	4.06 + 0.10	4.16

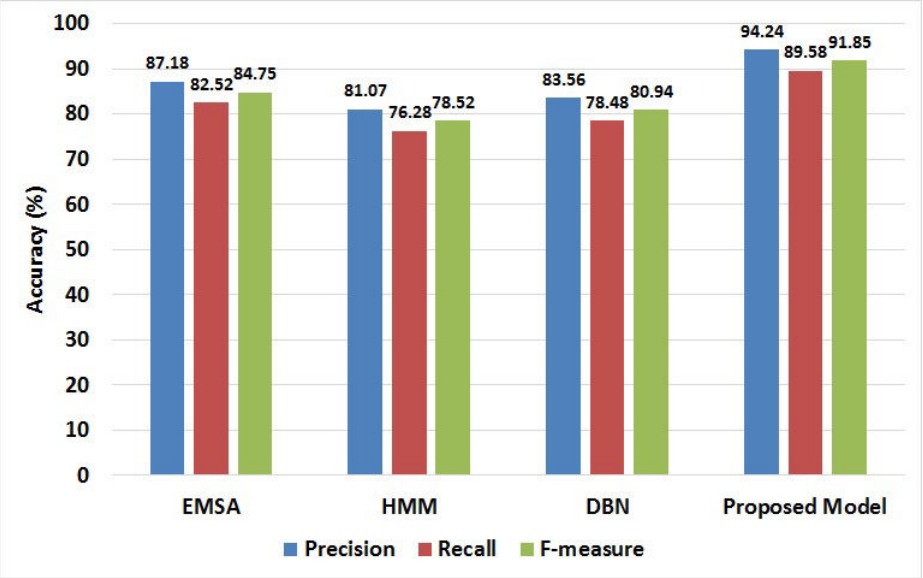

FIGURE 5.10: Best model selection based on the performance measurements.

compared with other state-of-the-art, the proposed strategy is faster around 6.981 seconds. The main reason for the large gap in recognition time is, the comparative techniques need to pre-process every activity sample by applying feature extraction operation before movement recognition.

Evaluation of temporal granule formulation with end-to-end decision-making efficiency

Multi-scale Convo-GRU method is proposed to recognize and classify the motor signals into different classes, which are further transferred to the cloud database for long-term storage. Temporal mining technique is used to retrieve the activity records from the private cloud storage to formulate time specific activity logs. These logs are further transferred to doctors and caretakers for diagnosis and analysis. Temporal granule efficiency depends on the time taken by the proposed model to formulate Temporal Activity Log (TAL) and the deliverance of TALs to respective end user by performing information abstraction operation on the cloud database. Fig. **??** described the

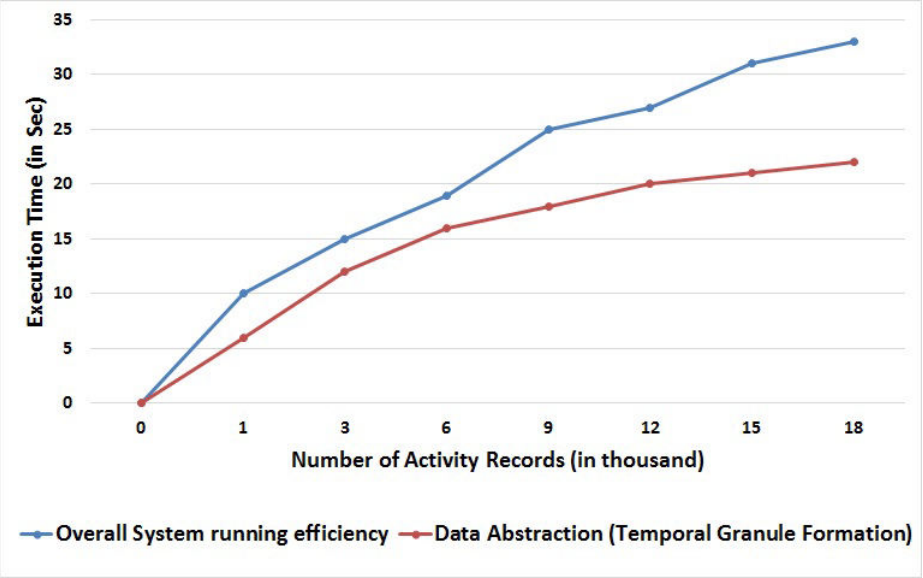

FIGURE 5.11: Temporal efficiency of the system temporal activity log formulation based decision making.

time taken by the proposed system for information abstraction is 22.8 seconds, which is profoundly adequate, particularly when a substantial number of records are included. The proposed framework took 33.7 seconds as the overall running time of TAL formation and deliverance which is highly acceptable

5.4.3 Movement recognition performance validation on publicly available dataset

MHEALTH Dataset includes fundamental signs and recordings of body movement. Total 12 type of physical activities have been captured from 10 Individuals with the assorted physical profile. The type of activities is listed in table **??**. The individual was focused on data collection and activity prediction in the data.

Table **??** presents the average activity recognition results of each individual. The proposed multi-scale Convo-GRU model achieved 98.16% mean

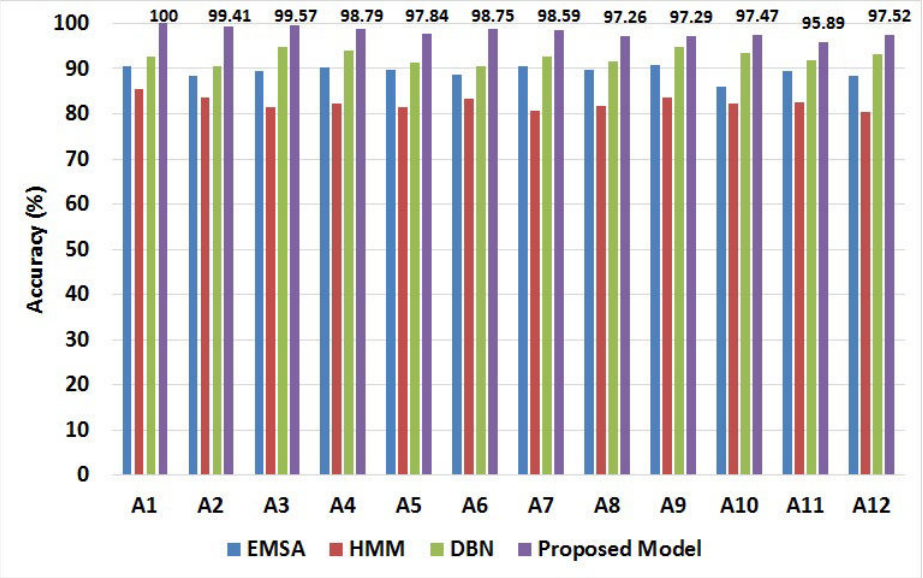

FIGURE 5.12: Comparative analysis of four different approaches on MHEALTH dataset.

prediction rate, which is adequately satisfactory.

The performance of the proposed methodology is validated on MHEALTH dataset and compared with the previously considered machine learning and deep learning approaches. The superiority of deep learning approaches over state-of-the-art machine learning approaches can be analyzed in Fig. **??**. In machine learning approaches, the experimental outcomes generated by EMSA is comparatively higher by achieving the 89.38% rate of precision as compared to HMM with the mean precision rate of 82.79%. On the other hand, in deep learning-based approaches, experiments showed better performance of DBN as compared to machine learning approaches by achieving 92.56% mean prediction performance. However, the recognition performance of DBN is still lower as compared to the proposed Multi-stage Convo-GRU-based movement recognition approach. Therefore, we can conclude that the capability of sequential motor movement recognition is highly satisfactory.

5.5 Conclusion

In the presented study, we have proposed edge analytics assisted deep learning based motor movement recognition framework to analyze the scale of physical inactivity of the patient under medical supervision. The purpose of using edge processing is to enhance the recognition efficiency with minimal delay for real-time decision making. The proposed deep learning approach not only exploits the time-series-based local dependency from 1D signals but also provide stability to the system for detecting temporal dynamics from unbalanced data in the real-world scenario. The CNN helps to extract features from captured activity signals without performing any normalization. The temporal dependencies are further captured from the extracted features by utilizing the sequential data learning principles of GRU methodology. The last fully connected layer also help to increase the generalization of the system with accurate activity classification. Temporal activity logs are formulated and delivered to medical specialists for improving the basic decision-making capability related to the physical performance of the individual. The network load efficiency for efficient network utilization is also maintained by utilizing a data abstraction technique. The above-calculated outcomes validate the efficiency of the motor movements recognition with timely deliverance of records on demand. In this manner, we can conclude that the proposed system is highly suitable for providing an effective healthcare-based monitoring environment for patients in every category.

TABLE 5.7: Confusion matrix based accuracy measurement.

Activities	**I (%)**	**II (%)**	**III (%)**	**IV (%)**	**V (%)**	**VI (%)**	**VII (%)**	**VIII (%)**	**IX (%)**	**X (%)**
I (Sitting and relaxing)	**100**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
II (Standing still)	0.0	**91.24**	0.0	0.0	0.0	0.0	0.0	2.8	0.0	0.0
III (Walking)	0.0	0.0	**90.26**	0.0	0.0	0.0	0.0	0.0	0.0	0.0
IV(Lying down)	0.0	0.0	0.0	**98.74**	0.0	1.05	0.0	0.0	0.0	0.0
V (Waist bends forward)	0.0	0.0	0.0	1.18	**88.42**	0.0	0.0	0.0	0.0	0.0
VI (Knees bending (crouching))	0.0	0.0	3.52	0.0	0.0	**95.84**	0.0	0.0	0.0	0.0
VII (Frontal elevation of arms)	0.0	0.0	0.0	0.0	0.0	4.58	**97.82**	0.0	0.0	0.0
VIII (Jogging)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**90.57**	0.0	0.0
IX (Cycling)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**100**	0.0
X (Running)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.8	0.0	**100**

TABLE 5.8: List of activities of MHEALTH dataset

Sr.no	Physical activity	Activity ID
1.	Standing still	I_1
2.	Sitting and relaxing	I_2
3.	Lying down	I_3
4.	Walking	I_4
5.	Climbing stairs	I_5
6.	Waist bends forward	I_6
7.	Frontal elevation of arms	I_7
8.	Knees bending	I_8
9.	Cycling	I_9
10.	Jogging	I_{10}
11.	Running	I_{11}
12.	Jump front and back	I_{12}

TABLE 5.9: Individual-based mean accuracy of activity recognition

	Individual	Activity prediction accuracy (%)
Mean accuracy	I_1	98.47
	I_2	97.71
	I_3	98.78
	I_4	98.29
	I_5	98.09
	I_6	98.19
	I_7	98.59
	I_8	97.19
	I_9	97.89
	I_{10}	98.19
Mean prediction		98.16

Chapter 6

Conclusion and Future Work

6.1 Thesis Summary

This thesis explored the utilization of innovative technologies, such as Computer Vision, Edge Analytics and Deep Learning in different application areas of the healthcare domain. The advanced principles of computer vision technology enabled with deep learning were shown to be useful for characterizing regular and irregular behavior of an individual based on the requirement of the application. The sensors provide quantitative information about the individual in a continuous manner. By utilizing a well-organized environment, indications based on physical and psychological behavior were illustrated and evaluated. This dissertation proposed four application-oriented real-time monitoring solutions to achieve an intelligent assistive environment for smart healthcare domain. However, human-computer interaction based novel strategies are proposed, they can be modified for different domains as well.

Findings:

- In chapter 2, we presented a novel computer vision-assisted activity monitoring framework to analyze the physical activities of an autistic child for irregular behavior prediction. The primary motive behind this study is to improve the safety measures for the child in an indoor environment by generating time-sensitive alerts for caretakers and doctors. In this study, a deep 3D CNN and LSTM based activity prediction

methodology are proposed to recognize physical irregularities. The 3D CNN model extracts the spatiotemporal features from the video templates for stance prediction with the subject position. The LSTM model calculates the temporal relationship in the feature maps to analyze the scale of irregularity. Moreover, in order to deal with such irregularities, a time-sensitive alert-based decision process is proposed in the present work to generate early warnings to the doctor and caretaker. The proficiency of the system is increased by storing the performed activity scores in the local database of the system which can be further utilized to provide medicinal or therapeutic assistance.

- By extending the complexity of the previous proposed study, in chapter 3, a novel computer vision assisted deep learning based posture monitoring system is presented. The monitoring framework is responsible to predict physical abnormalities of an individual to analyze the scale of Generalized Anxiety Disorder (GAD). We used deep learning-assisted 3D Convolutional Neural Network (CNN) technology for spatiotemporal feature extraction and Gated Recurrent Unit (GRU) model to exploit the extracted temporal dynamics for adversity scale determination. To validate the prediction performance of the proposed system, extensive experiments are conducted on three challenging datasets, NTU RGB+D, UTD-MHAD and HMDB51. The proposed methodology achieved comparable performance by obtaining the mean accuracy of 91.88%, 94.28%, and 70.33%, respectively. Furthermore, the average prediction time taken by the proposed methodology is approximately 1.13 seconds which demonstrates the real-time monitoring efficiency of the system.

- To extend the utility of the above-presented studies and to increase the applicability of the proposed studies in the smart healthcare domain, an edge analytic-assisted monitoring framework is discussed in chapter 4. The proposed monitoring solution is applicable to predict

health afflictions from physical activities of an individual in their ambient environment in real-time. Several advanced services and techniques such as cloud storage service, data mining technique, and real-time alert-based services are utilized in the proposed framework to fulfil the requirements of healthcare solutions. Performed activity based records are maintained by transmitting the predicted activity results on cloud and are stored using the sliding window approach which can be further utilized to analyze the physical condition of an individual for long-term medication. The proposed edge-based alert generation mechanism takes less time for decision making to enhance the personal and medical satisfaction. Experimental outcomes justify the superiority of the proposed framework over the conventional methodologies with higher prediction accuracy and less latency rate in alert-based decision making.

- The above-presented studies are completely dependent on visual sensors which have several limitations such as, region coverage constraint, privacy constraint, and many others. To overcome the above-mentioned constraints, we presented a wearable sensors based activity monitoring framework in chapter 5. The main objective of the proposed study is to calculate the scale of the physical inactivity of the patient to make real-time health suggestions. Graphical Processing Unit (GPU) enabled edge nodes are utilized for efficient data processing. An application scenario is proposed to validate the ideology of the proposed system in the healthcare environment. The performance is compared with both machine learning and deep learning-based approaches to justify the proposed system. iFogSim simulator is utilized to simulate the proposed scenario and the performance is validated based on the computation of movement recognition efficiency, network bandwidth efficiency, interoperability, Edge-based data processing reliability and alert generation-based patient security.

6.2 Future direction

Due to ever-increasing attention in the advancement of healthcare domain, several health monitoring applications and products are available in the market for users. By considering the sensitive nature of healthcare, the future direction of the research is divided into two domains:

- **Combined approach for different data modalities:** The future directions of this research can be based on the utilization of computer vision and wearable sensor-based data modalities in a single framework to monitor the physical as well as psychological well being of individuals more effectively.
- **Security, privacy and reliability-oriented research dimensions:** Patient privacy is also of extreme importance due to the involvement of computer vision based techniques. By considering the continuous advancement in smart healthcare-based monitoring solutions, there is a strong need to be focused on the development of reliable and secure architectures to handle the most sensitive information of the patient.

In the smart healthcare system, it is important to maintain the trade-off in data transmission, data processing and an acceptable value for the response time. The future research of the research proposed in this dissertation should be in that direction.

Bibliography

Abu-El-Haija, Sami et al. (2016). "Youtube-8m: A large-scale video classification benchmark". In: *arXiv preprint arXiv:1609.08675*.

Ahmed, Ejaz et al. (2017). "Bringing computation closer toward the user network: Is edge computing the solution?" In: *IEEE Communications Magazine* 55.11, pp. 138–144.

Akula, Aparna, Anuj K Shah, and Ripul Ghosh (2018). "Deep learning approach for human action recognition in infrared images". In: *Cognitive Systems Research* 50, pp. 146–154.

Albanese, Massimiliano et al. (2008). "A constrained probabilistic petri net framework for human activity detection in video". In: *IEEE Transactions on Multimedia* 10.8, pp. 1429–1443.

Alshurafa, Nabil et al. (2014). "Designing a robust activity recognition framework for health and exergaming using wearable sensors". In: *IEEE Journal of Biomedical and Health Informatics* 18.5, pp. 1636–1646.

Amor, Boulbaba Ben, Jingyong Su, and Anuj Srivastava (2016). "Action recognition using rate-invariant analysis of skeletal shape trajectories". In: *IEEE transactions on pattern analysis and machine intelligence* 38.1, pp. 1–13.

Anderson, Charles H, David C Van Essen, and Bruno A Olshausen (2005). "Directed visual attention and the dynamic control of information flow". In: *Neurobiology of attention*. Elsevier, pp. 11–17.

Arif, Muhammad and Ahmed Kattan (2015). "Physical activities monitoring using wearable acceleration sensors attached to the body". In: *PloS one* 10.7, e0130851.

Association, American Psychiatric et al. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Atallah, Louis et al. (2011). "Sensor positioning for activity recognition using wearable accelerometers". In: *IEEE transactions on biomedical circuits and systems* 5.4, pp. 320–329.

Baccouche, Moez et al. (2011). "Sequential deep learning for human action recognition". In: *International workshop on human behavior understanding*. Springer, pp. 29–39.

Baek, Jonghun et al. (2004). "Accelerometer signal processing for user activity detection". In: *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, pp. 610–617.

Baio, Jon et al. (2018). "Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2014". In: *MMWR Surveillance Summaries* 67.6, p. 1.

Banos, Oresti et al. (2014). "mHealthDroid: a novel framework for agile development of mobile health applications". In: *International workshop on ambient assisted living*. Springer, pp. 91–98.

Banos, Oresti et al. (2015). "Design, implementation and validation of a novel open framework for agile development of mobile health applications". In: *Biomedical engineering online* 14.2, S6.

Bao, Ling and Stephen S Intille (2004). "Activity recognition from user-annotated acceleration data". In: *International conference on pervasive computing*. Springer, pp. 1–17.

Barlow, Horace B (1989). "Unsupervised learning". In: *Neural computation* 1.3, pp. 295–311.

Berndt, Donald J and James Clifford (1994). "Using dynamic time warping to find patterns in time series." In: *KDD workshop*. Vol. 10. 16. Seattle, WA, pp. 359–370.

Bernstein, Gail A et al. (2017). "Use of computer vision tools to identify behavioral markers of pediatric obsessive–compulsive disorder: A pilot study". In: *Journal of child and adolescent psychopharmacology* 27.2, pp. 140–147.

Buch, Norbert, Sergio A Velastin, and James Orwell (2011). "A review of computer vision techniques for the analysis of urban traffic". In: *IEEE Transactions on Intelligent Transportation Systems* 12.3, pp. 920–939.

Cai, Linqin et al. (2018a). "Human action recognition using improved sparse Gaussian process latent variable model and hidden conditional random filed". In: *IEEE Access* 6, pp. 20047–20057.

Cai, Linqin et al. (2018b). "Robust human action recognition based on depth motion maps and improved convolutional neural network". In: *Journal of Electronic Imaging* 27.5, p. 051218.

Cao, Yang et al. (2016). "Share communication and computation resources on mobile devices: A social awareness perspective". In: *IEEE Wireless Communications* 23.4, pp. 52–59.

Carreira, Joao and Andrew Zisserman (2017). "Quo vadis, action recognition? a new model and the kinetics dataset". In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308.

Chaaraoui, Alexandros Andre, Pau Climent-Pérez, and Francisco Flórez-Revuelta (2013). "Silhouette-based human action recognition using sequences of key poses". In: *Pattern Recognition Letters* 34.15, pp. 1799–1807.

Chatfield, Ken et al. (2014). "Return of the devil in the details: Delving deep into convolutional nets". In: *arXiv preprint arXiv:1405.3531*.

Chen, Chen, Roozbeh Jafari, and Nasser Kehtarnavaz (2015). "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor". In: *2015 IEEE International conference on image processing (ICIP)*. IEEE, pp. 168–172.

Chen, Guang, Yuexian Zou, and Can Zhang (2019). "STMP: spatial temporal multi-level proposal network for activity detection". In: *International Conference on Multimedia Modeling*. Springer, pp. 29–41.

Chen, Xiang et al. (2017). "A quality-of-content-based joint source and channel coding for human detections in a mobile surveillance cloud". In: *IEEE Transactions on Circuits and Systems for Video Technology* 27.1, pp. 19–31.

Chen, Yuqing and Yang Xue (2015). "A deep learning approach to human activity recognition based on single accelerometer". In: *2015 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, pp. 1488–1492.

Chéron, Guilhem, Ivan Laptev, and Cordelia Schmid (2015). "P-cnn: Pose-based cnn features for action recognition". In: *Proceedings of the IEEE international conference on computer vision*, pp. 3218–3226.

Cho, Kyunghyun et al. (2014). "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *arXiv preprint arXiv:1406.1078*.

Chu, Wen-Sheng, Yale Song, and Alejandro Jaimes (2015). "Video co-summarization: Video summarization by visual co-occurrence". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3584–3592.

Cippitelli, Enea et al. (2016). "A human activity recognition system using skeleton data from RGBD sensors". In: *Computational intelligence and neuroscience* 2016, p. 21.

Ciptadi, Arridhana, Matthew S Goodwin, and James M Rehg (2014). "Movement pattern histogram for action recognition and retrieval". In: *European conference on computer vision*. Springer, pp. 695–710.

Cleland, Ian et al. (2013). "Optimal placement of accelerometers for the detection of everyday activities". In: *Sensors* 13.7, pp. 9183–9200.

Cook, Albert M and Janice Miller Polgar (2014). *Assistive Technologies-E-Book: Principles and Practice*. Elsevier Health Sciences.

Das, Srijan et al. (Feb. 2018). "A Fusion of Appearance based CNNs and Temporal evolution of Skeleton with LSTM for Daily Living Action Recognition". In:

Dawn, Debapratim Das and Soharab Hossain Shaikh (2016). "A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector". In: *The Visual Computer* 32.3, pp. 289–306.

Dean, J and S Ghemawat (2004). "Simplified data processing on large clusters, Sixth Symp". In: *Oper. Syst. Des. Implement* 51.1, pp. 107–113.

Deboeverie, Francis et al. (2016). "Human gesture classification by brute-force machine learning for exergaming in physiotherapy". In: *2016 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, pp. 1–7.

Donahue, Jeffrey et al. (2015). "Long-term recurrent convolutional networks for visual recognition and description". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2625–2634.

Du, Yong, Wei Wang, and Liang Wang (2015). "Hierarchical recurrent neural network for skeleton based action recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1110–1118.

Dunlap, Glen, Kathleen Dyer, and Robert L Koegel (1983). "Autistic self-stimulation and intertrial interval duration." In: *American journal of mental deficiency*.

El-Sayed, Hesham et al. (2018). "Edge of things: The big picture on the integration of edge, IoT and the cloud in a distributed computing environment". In: *IEEE Access* 6, pp. 1706–1717.

Ezzahout, Abderrahmane and Rachid Oulad Haj Thami (2013). "Conception and development of a video surveillance system for detecting, tracking and profile analysis of a person". In: *2013 3rd international symposium ISKO-Maghreb*. IEEE, pp. 1–5.

Fasching, Joshua et al. (2012). "Detecting risk-markers in children in a preschool classroom". In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, pp. 1010–1016.

Fasching, Joshua et al. (2015). "Classification of motor stereotypies in video". In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 4894–4900.

Fasching, Joshua et al. (2016). "Automated coding of activity videos from an OCD study". In: *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 5638–5643.

Feichtenhofer, Christoph, Axel Pinz, and Richard P Wildes (2017). "Spatiotemporal multiplier networks for video action recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4768–4777.

Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman (2016). "Convolutional two-stream network fusion for video action recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1933–1941.

Fulceri, Francesca et al. (2015). "LOCOMOTION AND GRASPING IMPAIRMENT IN PRESCHOOLERS WITH AUTISM SPECTRUM DISORDER." In: *Clinical Neuropsychiatry* 4.

Gaidon, Adrien, Zaid Harchaoui, and Cordelia Schmid (2013). "Temporal localization of actions with actoms". In: *IEEE transactions on pattern analysis and machine intelligence* 35.11, pp. 2782–2795.

Ganz, Michael L (2007). "The lifetime distribution of the incremental societal costs of autism". In: *Archives of pediatrics & adolescent medicine* 161.4, pp. 343–349.

Gao, Junfeng et al. (2018). "Computer Vision in Healthcare Applications". In: *Journal of healthcare engineering* 2018.

Garcia Lopez, Pedro et al. (2015). "Edge-centric computing: Vision and challenges". In: *ACM SIGCOMM Computer Communication Review* 45.5, pp. 37–42.

Gers, Felix A, Nicol N Schraudolph, and Jürgen Schmidhuber (2002). "Learning precise timing with LSTM recurrent networks". In: *Journal of machine learning research* 3.Aug, pp. 115–143.

Ghasemzadeh, Hassan and Roozbeh Jafari (2011). "Physical movement monitoring using body sensor networks: A phonological approach to construct spatial decision trees". In: *IEEE Transactions on Industrial Informatics* 7.1, pp. 66–77.

Ghasemzadeh, Hassan, Roozbeh Jafari, and Balakrishnan Prabhakaran (2010). "A body sensor network with electromyogram and inertial sensors: Multimodal interpretation of muscular activities". In: *IEEE transactions on information technology in biomedicine* 14.2, pp. 198–206.

Ghasemzadeh, Hassan, Vitali Loseu, and Roozbeh Jafari (2010). "Structural action recognition in body sensor networks: Distributed classification based on string matching". In: *IEEE Transactions on Information Technology in Biomedicine* 14.2, pp. 425–435.

Giansanti, Daniele, Velio Macellari, and Giovanni Maccioni (2008). "New neural network classifier of fall-risk based on the Mahalanobis distance and kinematic parameters assessed by a wearable device". In: *Physiological measurement* 29.3, N11.

Gjoreski, Hristijan, Mitja Lustrek, and Matjaz Gams (2011). "Accelerometer placement for posture recognition and fall detection". In: *2011 Seventh International Conference on Intelligent Environments*. IEEE, pp. 47–54.

Gjoreski, Martin et al. (2016). "How accurately can your wrist device recognize daily activities and detect falls?" In: *Sensors* 16.6, p. 800.

Gkioxari, Georgia and Jitendra Malik (2015). "Finding action tubes". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 759–768.

Goldman, Sylvie et al. (2009). "Motor stereotypies in children with autism and other developmental disorders". In: *Developmental Medicine & Child Neurology* 51.1, pp. 30–38.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press.

Grabill, Kristen et al. (2008). "Assessment of obsessive–compulsive disorder: a review". In: *Journal of anxiety disorders* 22.1, pp. 1–17.

Graves, Alex (2013). "Generating sequences with recurrent neural networks". In: *arXiv preprint arXiv:1308.0850*.

Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton (2013). "Speech recognition with deep recurrent neural networks". In: *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, pp. 6645–6649.

Griffith, Gemma M et al. (2010). "Using matched groups to explore child behavior problems and maternal well-being in children with Down syndrome and autism". In: *Journal of Autism and Developmental Disorders* 40.5, pp. 610–619.

Gupta, Ankur et al. (2014). "3D pose from motion for cross-view action recognition via non-linear circulant temporal encoding". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2601–2608.

Gupta, Harshit et al. (2017). "iFogSim: A toolkit for modeling and simulation of resource management techniques in the Internet of Things, Edge and Fog computing environments". In: *Software: Practice and Experience* 47.9, pp. 1275–1296.

Hammerla, Nils Y, Shane Halloran, and Thomas Plötz (2016). "Deep, convolutional, and recurrent models for human activity recognition using wearables". In: *arXiv preprint arXiv:1604.08880*.

Han, Yamin et al. (2018). "Going deeper with two-stream ConvNets for action recognition in video surveillance". In: *Pattern Recognition Letters* 107, pp. 83–90.

Hashemi, Jordan et al. (2012). "A computer vision approach for the assessment of autism-related behavioral markers". In: *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*. IEEE, pp. 1–7.

Hashemi, Jordan et al. (2014). "Computer vision tools for low-cost and non-invasive measurement of autism-related behaviors in infants". In: *Autism research and treatment* 2014.

Health, U.S. Department of and Human Services (2017). *Autism spectrum disorder (ASD)*.

Hinton, Geoffrey E and Ruslan R Salakhutdinov (2006). "Reducing the dimensionality of data with neural networks". In: *science* 313.5786, pp. 504–507.

Hutt, C, SJ Forrest, and J Richer (1975). "Cardiac arrhythmia and behaviour in autistic children". In: *Acta Psychiatrica Scandinavica* 51.5, pp. 361–372.

Hutt, Corinne and Christopher Ounsted (1966). "The biological significance of gaze aversion with particular reference to the syndrome of infantile autism". In: *Behavioral science* 11.5, pp. 346–356.

Ijjina, Earnest Paul and Krishna Mohan Chalavadi (2017). "Human action recognition in RGB-D videos using motion sequence information and deep learning". In: *Pattern Recognition* 72, pp. 504–516.

Jefferis, Barbara J et al. (2012). "Longitudinal associations between changes in physical activity and onset of type 2 diabetes in older British men: the influence of adiposity". In: *Diabetes Care* 35.9, pp. 1876–1883.

Ji, Shuiwang et al. (2013). "3D convolutional neural networks for human action recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 35.1, pp. 221–231.

Karpathy, Andrej et al. (2014). "Large-scale video classification with convolutional neural networks". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732.

Karumbaya, Athena and Gowri Satheesh (2015). "Iot empowered real time environment monitoring system". In: *International Journal of Computer Applications* 129.5, pp. 30–32.

Kazemi, Vahid and Josephine Sullivan (2014). "One millisecond face alignment with an ensemble of regression trees". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1867–1874.

Kern, Nicky, Bernt Schiele, and Albrecht Schmidt (2003). "Multi-sensor activity context detection for wearable computing". In: *European Symposium on Ambient Intelligence*. Springer, pp. 220–232.

Khan, Adil Mehmood et al. (2010). "A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer". In: *IEEE transactions on information technology in biomedicine* 14.5, pp. 1166–1172.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kiranyaz, Serkan, Turker Ince, and Moncef Gabbouj (2015). "Real-time patient-specific ECG classification by 1-D convolutional neural networks". In: *IEEE Transactions on Biomedical Engineering* 63.3, pp. 664–675.

Kolda, Tamara G and Brett W Bader (2009). "Tensor decompositions and applications". In: *SIAM review* 51.3, pp. 455–500.

Krishna, Ranjay et al. (2017). "Dense-captioning events in videos". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 706–715.

Krishnan, Narayanan C and Diane J Cook (2014). "Activity recognition on streaming sensor data". In: *Pervasive and mobile computing* 10, pp. 138–154.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*, pp. 1097–1105.

Kuehne, Hildegard et al. (2011). "HMDB: a large video database for human motion recognition". In: *2011 International Conference on Computer Vision*. IEEE, pp. 2556–2563.

Lai, Liangzhen, Naveen Suda, and Vikas Chandra (2018). "Cmsis-nn: Efficient neural network kernels for arm cortex-m cpus". In: *arXiv preprint arXiv:1801.06601*.

Laptev, Ivan (2005). "On space-time interest points". In: *International journal of computer vision* 64.2-3, pp. 107–123.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *nature* 521.7553, p. 436.

LeCun, Yann, Yoshua Bengio, et al. (1995). "Convolutional networks for images, speech, and time series". In: *The handbook of brain theory and neural networks* 3361.10, p. 1995.

LeCun, Yann et al. (1988). "A theoretical framework for back-propagation". In: *Proceedings of the 1988 connectionist models summer school*. Vol. 1. CMU, Pittsburgh, Pa: Morgan Kaufmann, pp. 21–28.

Lee, Tao-Yi et al. "Elastic Motif Segmentation and Alignment of Time Series for Encoding and Classification". In: ().

Li, Binlong, Octavia I Camps, and Mario Sznaier (2012). "Cross-view activity recognition using hankelets". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1362–1369.

Li, Ruonan and Todd Zickler (2012). "Discriminative virtual views for cross-view action recognition". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 2855–2862.

Li, Yanghao et al. (2016). "Online human action detection using joint classification-regression recurrent neural networks". In: *European Conference on Computer Vision*. Springer, pp. 203–220.

Lin, Liang et al. (2016). "A deep structured model with radius–margin bound for 3D human activity recognition". In: *International Journal of Computer Vision* 118.2, pp. 256–273.

Lindemann, Ulrich et al. (2005). "Evaluation of a fall detector based on accelerometers: A pilot study". In: *Medical and Biological engineering and computing* 43.5, pp. 548–551.

Liu, Jun et al. (2016). "Spatio-temporal lstm with trust gates for 3d human action recognition". In: *European Conference on Computer Vision*. Springer, pp. 816–833.

Lord, Catherine, Fred Volkmar, and Paul J Lombroso (2002). "Genetics of childhood disorders: XLII. Autism, part 1: Diagnosis and assessment in autistic spectrum disorders". In: *Journal of the American Academy of Child & Adolescent Psychiatry* 41.9, pp. 1134–1136.

Lord, Catherine et al. (2000). "The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism". In: *Journal of autism and developmental disorders* 30.3, pp. 205–223.

Luo, Xiaochun et al. (2018). "Towards efficient and objective work sampling: Recognizing workers' activities in site surveillance videos with two-stream convolutional networks". In: *Automation in Construction* 94, pp. 360–370.

Luyster, Rhiannon et al. (2009). "The Autism Diagnostic Observation Schedule—Toddler Module: A new module of a standardized diagnostic measure for autism spectrum disorders". In: *Journal of autism and developmental disorders* 39.9, pp. 1305–1320.

Mabrouk, Amira Ben and Ezzeddine Zagrouba (2018). "Abnormal behavior recognition for intelligent video surveillance systems: A review". In: *Expert Systems with Applications* 91, pp. 480–491.

Man, D and A Vision (1982). *A computational investigation into the human representation and processing of visual information*.

Mannini, Andrea et al. (2013). "Activity recognition using a single accelerometer placed at the wrist or ankle". In: *Medicine and science in sports and exercise* 45.11, p. 2193.

Marschollek, Michael et al. (2008). "Assessing elderly persons' fall risk using spectral analysis on accelerometric data-a clinical evaluation study". In: *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, pp. 3682–3685.

Merler, Michele et al. (2012). "Semantic model vectors for complex video event recognition". In: *IEEE Transactions on Multimedia* 14.1, pp. 88–101.

Mikolov, Tomáš et al. (2010). "Recurrent neural network based language model". In: *Eleventh annual conference of the international speech communication association*.

Minhas, Rashid, Abdul Adeel Mohammed, and QM Jonathan Wu (2012). "Incremental learning in human action recognition based on snippets". In: *IEEE Transactions on Circuits and Systems for Video Technology* 22.11, pp. 1529–1541.

Minnen, David et al. (2005). "Recognizing and discovering human actions from on-body sensor data". In: *2005 IEEE International Conference on Multimedia and Expo*. IEEE, pp. 1545–1548.

Moon, Taesup et al. (2015). "Rnndrop: A novel dropout for rnns in asr". In: *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, pp. 65–70.

Mubashir, Muhammad, Ling Shao, and Luke Seed (2013). "A survey on fall detection: Principles and approaches". In: *Neurocomputing* 100, pp. 144–152.

Mullen, Eileen M et al. (1995). *Mullen scales of early learning*. AGS Circle Pines, MN.

Munaro, Matteo and Emanuele Menegatti (2014). "Fast RGB-D people tracking for service robots". In: *Autonomous Robots* 37.3, pp. 227–242.

Myers, Barbara J, Virginia H Mackintosh, and Robin P Goin-Kochel (2009). ""My greatest joy and my greatest heart ache:" Parents' own words on how having a child in the autism spectrum has affected their lives and their families' lives". In: *Research in Autism Spectrum Disorders* 3.3, pp. 670–684.

Narayanan, Michael R et al. (2008). "A wearable triaxial accelerometry system for longitudinal assessment of falls risk". In: *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, pp. 2840–2843.

Nayak, Nandita M, Yingying Zhu, and Amit K Roy-Chowdhury (2013). "Exploiting spatio-temporal scene structure for wide-area activity analysis in unconstrained environments". In: *IEEE Transactions on Information Forensics and Security* 8.10, pp. 1610–1619.

Neverova, Natalia et al. (2016a). "Learning human identity from motion patterns". In: *IEEE Access* 4, pp. 1810–1820.

Neverova, Natalia et al. (2016b). "Moddrop: adaptive multi-modal gesture recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.8, pp. 1692–1706.

Ng, Samantha et al. (2009). "Towards a mobility diagnostic tool: Tracking rollator users' leg pose with a monocular vision system". In: *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, pp. 1220–1225.

Noury, N et al. (2008). "A proposal for the classification and evaluation of fall detectors". In: *Irbm* 29.6, pp. 340–349.

Noury, Norbert (2002). "A smart sensor for the remote follow up of activity and fall detection of the elderly". In: *2nd Annual International IEEE-EMBS Special Topic Conference on Microtechnologies in Medicine and Biology. Proceedings (Cat. No. 02EX578)*. IEEE, pp. 314–317.

Nyan, MN et al. (2006). "Distinguishing fall activities from normal activities by angular rate characteristics and high-speed camera characterization". In: *Medical engineering & physics* 28.8, pp. 842–849.

Office, International Labour (2015). *World employment and social outlook: trends 2015*. International Labour Organization Geneva.

Olguın, Daniel Olguın and Alex Sandy Pentland (2006). "Human activity recognition: Accuracy across common locations for wearable sensors". In: *Proceedings of 2006 10th IEEE international symposium on wearable computers, Montreux, Switzerland*. Citeseer, pp. 11–14.

Pan, Yingwei et al. (2016). "Jointly modeling embedding and translation to bridge video and language". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4594–4602.

Papert, Seymour A (1966). "The summer vision project". In:

Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio (2013). "On the difficulty of training recurrent neural networks". In: *International conference on machine learning*, pp. 1310–1318.

Pehlivan, Selen and Pınar Duygulu (2011). "A new pose-based representation for recognizing actions from multiple cameras". In: *Computer Vision and Image Understanding* 115.2, pp. 140–151.

Perry, James T et al. (2009). "Survey and evaluation of real-time fall detection approaches". In: *2009 6th International Symposium on High Capacity Optical Networks and Enabling Technologies (HONET)*. IEEE, pp. 158–164.

Plötz, Thomas, Nils Y Hammerla, and Patrick L Olivier (2011). "Feature learning for activity recognition in ubiquitous computing". In: *Twenty-Second International Joint Conference on Artificial Intelligence*.

Quattoni, Ariadna et al. (2007). "Hidden conditional random fields". In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 10, pp. 1848–1852.

Radomsky, Adam S and S Rachman (2004). "Symmetry, ordering and arranging compulsive behaviour". In: *Behaviour Research and Therapy* 42.8, pp. 893–913.

Rajagopalan, Shyam, Abhinav Dhall, and Roland Goecke (2013). "Self-stimulatory behaviours in the wild for autism diagnosis". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 755–761.

Rehg, James et al. (2013). "Decoding children's social behavior". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3414–3421.

Ren, Shaoqing et al. (2015). "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems*, pp. 91–99.

Rimminen, Henry et al. (2010). "Detection of falls among the elderly by a floor sensor using the electric near field". In: *IEEE Transactions on Information Technology in Biomedicine* 14.6, pp. 1475–1476.

Sacchi, Lucia et al. (2007). "Data mining with temporal abstractions: learning rules from time series". In: *Data Mining and Knowledge Discovery* 15.2, pp. 217–247.

Sadanand, Sreemananth and Jason J Corso (2012). "Action bank: A high-level representation of activity in video". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1234–1241.

Scahill, Lawrence et al. (1997). "Children's Yale-Brown obsessive compulsive scale: reliability and validity". In: *Journal of the American Academy of Child & Adolescent Psychiatry* 36.6, pp. 844–852.

Schiffman, Jason et al. (2004). "Childhood videotaped social and neuromotor precursors of schizophrenia: a prospective investigation". In: *American Journal of Psychiatry* 161.11, pp. 2021–2027.

Shahroudy, Amir et al. (2016). "NTU RGB+ D: A large scale dataset for 3D human activity analysis". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1010–1019.

Shi, Guangyi et al. (2009). "Mobile human airbag system for fall protection using MEMS sensors and embedded SVM classifier". In: *IEEE Sensors Journal* 9.5, pp. 495–503.

Shi, Weisong et al. (2016). "Edge computing: Vision and challenges". In: *IEEE Internet of Things Journal* 3.5, pp. 637–646.

Shou, Zheng, Dongang Wang, and Shih-Fu Chang (2016). "Temporal action localization in untrimmed videos via multi-stage cnns". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1049–1058.

Simonyan, Karen and Andrew Zisserman (2014a). "Two-stream convolutional networks for action recognition in videos". In: *Advances in neural information processing systems*, pp. 568–576.

— (2014b). "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556*.

Singhal, Shivam and Vikas Tripathi (2019). "Action recognition framework based on normalized local binary pattern". In: *Progress in Advanced Computing and Intelligent Engineering*. Springer, pp. 247–255.

Sivalingam, Ravishankar et al. (2012). "A multi-sensor visual tracking system for behavior monitoring of at-risk children". In: *2012 IEEE International Conference on Robotics and Automation*. IEEE, pp. 1345–1350.

Song, Sijie et al. (2017). "An end-to-end spatio-temporal attention model for human action recognition from skeleton data". In: *Thirty-first AAAI conference on artificial intelligence*.

Srivastava, Nitish et al. (2014). "Dropout: a simple way to prevent neural networks from overfitting". In: *The Journal of Machine Learning Research* 15.1, pp. 1929–1958.

Sterling, Peter and Simon Laughlin (2015). *Principles of neural design*. MIT Press.

Sun, Lin et al. (2015). "Human action recognition using factorized spatio-temporal convolutional networks". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4597–4605.

Szeliski, Richard (2010). *Computer vision: algorithms and applications*. Springer Science and Business Media.

Thurau, Christian and Václav Hlavác (2008). "Pose primitive based human action recognition in videos or still images". In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1–8.

Tran, Du and Alexander Sorokin (2008). "Human activity recognition with metric learning". In: *European conference on computer vision*. Springer, pp. 548–561.

Tran, Du et al. (2015). "Learning spatiotemporal features with 3d convolutional networks". In: *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497.

Veeriah, Vivek, Naifan Zhuang, and Guo-Jun Qi (2015). "Differential recurrent neural networks for action recognition". In: *Proceedings of the IEEE international conference on computer vision*, pp. 4041–4049.

Viola, Paul and Michael J Jones (2004). "Robust real-time face detection". In: *International journal of computer vision* 57.2, pp. 137–154.

Walczak, Nicholas et al. (2013). "Locating occupants in preschool classrooms using a multiple RGB-D sensor system". In: *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, pp. 2166–2172.

Walker, Elaine F, Tammy Savoie, and Dana Davis (1994). "Neuromotor precursors of schizophrenia". In: *Schizophrenia bulletin* 20.3, pp. 441–451.

Wang, Heng and Cordelia Schmid (2013). "Action recognition with improved trajectories". In: *Proceedings of the IEEE international conference on computer vision*, pp. 3551–3558.

Wang, Heng et al. (2011). "Action recognition by dense trajectories". In: *CVPR 2011-IEEE Conference on Computer Vision & Pattern Recognition*. IEEE, pp. 3169–3176.

Wang, Jiang et al. (2014). "Cross-view action modeling, learning and recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2649–2656.

Wang, Liang et al. (2012). "A hierarchical approach to real-time activity recognition in body sensor networks". In: *Pervasive and Mobile Computing* 8.1, pp. 115–130.

Wang, Liangliang et al. (2017a). "Three-stream CNNs for action recognition". In: *Pattern Recognition Letters* 92, pp. 33–40.

Wang, Limin, Yu Qiao, and Xiaoou Tang (2015). "Action recognition with trajectory-pooled deep-convolutional descriptors". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4305–4314.

Wang, Pichao et al. (2018). "Action recognition based on joint trajectory maps with convolutional neural networks". In: *Knowledge-Based Systems* 158, pp. 43–53.

Wang, Xiaolong, Ali Farhadi, and Abhinav Gupta (2016). "Actions˜ transformations". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2658–2667.

Wang, Yunbo et al. (2017b). "Spatiotemporal pyramid network for video action recognition". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1529–1538.

Weinland, Daniel, Mustafa Özuysal, and Pascal Fua (2010). "Making action recognition robust to occlusions and viewpoint changes". In: *European Conference on Computer Vision*. Springer, pp. 635–648.

Werbos, Paul J et al. (1990). "Backpropagation through time: what it does and how to do it". In: *Proceedings of the IEEE* 78.10, pp. 1550–1560.

Xu, Zhongwen, Yi Yang, and Alex G Hauptmann (2015). "A discriminative CNN video representation for event detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1798–1807.

Yang, Jianbo et al. (2015). "Deep convolutional neural networks on multichannel time series for human activity recognition". In: *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Yang, Xiaodong, Pavlo Molchanov, and Jan Kautz (2016). "Multilayer and multimodal fusion of deep neural networks for video classification". In:

Proceedings of the 24th ACM international conference on multimedia. ACM, pp. 978–987.

Ye, Juan, Graeme Stevenson, and Simon Dobson (2015). "KCAR: A knowledge-driven approach for concurrent activity recognition". In: *Pervasive and Mobile Computing* 19, pp. 47–70.

Ye, Yun et al. (2013). "Wireless video surveillance: A survey". In: *IEEE Access* 1, pp. 646–660.

Yilmaz, Alper, Omar Javed, and Mubarak Shah (2006). "Object tracking: A survey". In: *Acm computing surveys (CSUR)* 38.4, p. 13.

Yin, Jie, Qiang Yang, and Jeffrey Junfeng Pan (2008). "Sensor-based abnormal human-activity detection". In: *IEEE Transactions on Knowledge and Data Engineering* 20.8, pp. 1082–1090.

Yu, Wenlan et al. (2013). "A survey of occupational health hazards among 7,610 female workers in China's electronics industry". In: *Archives of environmental & occupational health* 68.4, pp. 190–195.

Yue-Hei Ng, Joe et al. (2015). "Beyond short snippets: Deep networks for video classification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4694–4702.

Zeiler, Matthew D (2012). "ADADELTA: an adaptive learning rate method". In: *arXiv preprint arXiv:1212.5701*.

Zhang, Jiajia, Kun Shao, and Xing Luo (2018). "Small sample image recognition using improved Convolutional Neural Network". In: *Journal of Visual Communication and Image Representation* 55, pp. 640–647.

Zhang, Zhong et al. (2013). "Cross-view action recognition via a continuous virtual path". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2690–2697.

Zheng, Jingjing and Zhuolin Jiang (2013). "Learning view-invariant sparse representations for cross-view action recognition". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3176–3183.

Zhu, Guangming et al. (2016). "An online continuous human action recognition algorithm based on the Kinect sensor". In: *Sensors* 16.2, p. 161.

Zhu, Hongyi, Hsinchun Chen, and Randall Brown (2018). "A sequence-to-sequence model-based deep learning approach for recognizing activity of

daily living for senior care". In: *Journal of biomedical informatics* 84, pp. 148–158.

Publications

1. Ankush Manocha, Ramandeep Singh, "An intelligent monitoring system for indoor safety of individuals suffering from Autism Spectrum Disorder (ASD)" Journal of Ambient Intelligence and Humanized Computing, (March 2019) 1-16. (SCI/SCIE Indexed with Impact Factor 1.423)

2. Ankush Manocha, Ramandeep Singh, "Computer vision based working environment monitoring to analyze Generalized Anxiety Disorder (GAD)" Multimedia Tools and Applications, (May 2019) 1-28. (SCI/SCIE Indexed with Impact Factor 1.541)

3. Ramandeep Singh, Ankush Manocha, Prabal Verma, "IoT-Fog assisted Sleep Deprivation Prediction Framework for Spinal Cord Injury (SCI) Patients" IEEE Computer. (Accepted, SCI/SCIE Indexed with Impact Factor 1.94)

4. Ankush Manocha, Ramandeep Singh, "Deep learning based an ensembled approach for human activity recognition", 4^{th} International Conference on New Frontiers of Engineering, Management, Social Science and Humanities, 2019.

5. Ankush Manocha, Ramandeep Singh, "Video Analytics and Deep Learning based Intelligent System for Patient Well-being Monitoring", 12^{th} International Conference on Recent Development in Engineering Science, Humanities and Management, 2019.

6. Ankush Manocha, Ramandeep Singh, "A novel edge analytics assisted motor movement recognition framework using Multi-stage Convo-GRU model" Mobile Networks and Applications. (Revision Submitted, SCI/SCIE Indexed with Impact Factor 2.497)

7. Ankush Manocha, Ramandeep Singh, "Edge analytics assisted deep learning based real-time video processing for health affliction prediction" Journal of Real-Time Image Processing. (Under Review, SCI/SCIE Indexed with Impact Factor 1.574)

8. Ankush Manocha, Ramandeep Singh, "Cognitive Intelligence assisted Fog-Cloud Architecture for Generalized Anxiety Disorder (GAD) Prediction" Journal of Medical Systems. (Under Review, SCI/SCIE Indexed with Impact Factor 2.098)