Bangladesh University of Business and Technology

Cover Page

PROPOSAL SUMMARY

1. Proposal Title: Human activity recognition in deep learning and machine learning architecture using Vision dataset.
2. Supervisor: M. M. Fazle Rabbi
3. Affiliation: Bangladesh University of Business and Technology
4. Department: Department of Computer Science and Engineering
5. Area of Specialization: Artificial Intelligence
6. Team Members: 1. Habibullah (18192103080) **2. Pallab Majumdar (18192103050)** **3. Mafuja Akter Mitu (18192103068)** **4. Joy Adhikary (18192103062)** **5. Al Ahad Sufian (18192103056)**

Abstract:

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

Bangladesh University of Business and Technology

2. Introduction

Human action recognition is an important task for deep learning. The development of numerous significant applications, including intelligent surveillance systems, human-computer interfaces, health care, security, and military applications, is made possible by the capacity to identify, comprehend, and predict complex human actions. The computer vision industry has focused particularly on deep learning in recent years. This article provides a summary of the state-of-the-art in action recognition using deep learning and video analysis. The most significant deep learning models for human action recognition are presented, and their strengths and weaknesses are highlighted in order to show the current state of deep learning algorithms used to address human action recognition issues in realistic movies. There will a large number of data where data set will be predicted by video classification. On this research work, both machine learning and deep learning algorithm will be implemented. Data set will be different for different model of deep learning and machine learning. We will classify video and predict by the video by using deep learning. On the other hand we will use machine learning model for classify the activity.

3. Literature Review

Human activity recognition," or "HAR," is when a machine can figure out what a person is doing. Kaggle had the challenge to classify six different human activities based on the inertial signals from the phones of 30 volunteers. Data was shown using t-distributed stochastic neighborhood embedding (t-SNE) and machine learning techniques like logistic regression, linear SVC, kernel SVM, and decision trees. This made it easier to put the six different human activities into groups. The results showed that the linear support vector classifier and the gated recurrent unit in deep learning were more accurate at recognizing human activities than other classifiers. The Kaggle platform received a dataset from UCI with the name UCI HAR Dataset that may be used to derive insightful information and spot significant trends.[1]

T-SNE is a dimensionality reduction approach that reduces the tendency for points to cluster in the map's center, which helps to visualize data. The principal component analysis isn't as good as newer methods, especially when reducing a data point from a higher dimension to a lower dimension. To study human activity recognition, the right machine learning and data mining techniques need to be developed. Zameer Gulzar et al. say that threshold-based algorithms are easier and faster, but the machine algorithm gives a reliable answer. Tahmina Zebin and her coworkers used CNN to automate the process of learning features from data from multiple channels over time. They then focused their study on how deep learning algorithms can be used to recognize human activities.[2]

Seung Min Oh et al. came up with a method in which deep learning models are trained and tested using raw time series data and machine learning models are trained and tested using expert-generated features. Logical regression is used to find the best hyperplanes that clearly and linearly separate six activities from each other. This is done for classification problems in machine learning. Finding margin-maximizing hyperplanes that more accurately categorize six different human activities makes use of linear SVM. [7] The sigmoidal activation function is used in the squashing approach to convert greater signed distance values to a value between 0 and 1. Recurrent neural networks like RBF Kernel SVM and Bidirectional Long Short Term Memory (LSTM) are capable of learning and remembering lengthy input data sequences. Recurrent neural networks use a gating technique called GRU

that has just two gates, the reset and update gates, respectively. Raw time series data was used to train recurrent neural networks, which needed 1,542 parameters, while the gated recurrent unit model needed 4,230 parameters and was 92.60% accurate. This study found that machine learning models were better than deep learning models at classifying the six different things people do. These discoveries could be used to make smartwatches and other devices that track what a user does and let them know what their daily activity record is. The article discusses the challenges of visually analyzing seizure data for epilepsy diagnosis and localization of seizure onset zones and the limitations of the current computer vision approach for automated seizure classification. The authors propose a novel deep learning-based approach incorporating spatiotemporal aspects of seizure semiology and using infrared and depth video data for a 3-class (FLE, TLE, non-epileptic) seizure classification pipeline. The approach is based on the analysis of specific movements of interest (MOI) and their dynamics and biomechanical characteristics. The study utilizes the largest 3D-video-EEG database in the world and achieves promising results for cross-subject validation of near-real-time seizure detection and classification, outperforming previously published methods. [3]

The difficulties of visually interpreting seizure data for epilepsy diagnosis and seizure onset zone localization, as well as the shortcomings of the current computer vision approach for automated seizure categorization, are discussed in this article. The authors suggest a brand-new deep learning-based pipeline for a 3-class (FLE, TLE, non-epileptic) seizure classification pipeline that incorporates spatiotemporal elements of seizure semiology. For the 2 classes (FLE vs. TLE) and the 3 classes (0.763 vs. 0.083), the system obtained a promising cross-subject validation f1-score of 0.833 0.061 and 0.763 0.083, respectively. The outcomes beat all previously published techniques, demonstrating the viability of the deep learning strategy to allow 24/7 epilepsy monitoring. A morphological dilation approach was used to fill in the missing pixel values that cause pepper noise in the Kinect v2 depth signal. Using batches of 500 and 1000 samples for the 2- and 3-class scenarios, respectively, the models were trained for 2000 iterations. The design with the highest validation f1 score was finally selected, using early stopping. This essay assesses a deep-learning seizure classification pipeline that classifies seizures using I3D, LSTM, and Extended LSTM. The efficiency of I3D features collected from three datasets was assessed using depth cropping and temporal slicing, two techniques. Macro averages of the 5-fold cross-validation metrics were used to assess the classifiers' performance. With an F1 score of 0.833 0.061, the LSTM classifier with 2D cropping had the best performance. For the 3D cropped dataset B, the I3D classifier scored best in the 2-class case. By visually verifying the movies, the outcomes of the two cropping techniques were assessed, and in the majority of cases, the region of interest was correctly identified by the developed algorithm. There are many challenges involved in building the deep learning seizure classification pipeline, including managing the enormous amounts of data needed for the classifier's training and testing, making sure the system can accurately classify seizures in real-world scenarios, and generalizing the system to new data effectively. The 3-class architecture in particular demonstrated a high potential to be used for near-real-time classification and alert of seizures in hospitals in the future for online event-detection applications in epilepsy monitoring units. [4]

The importance of human action recognition in computer vision research and its applications in systems for patient monitoring and surveillance are discussed in this essay. It examines sequential and space-time volume encoding strategies for actions, as well as hierarchical action recognition methods for uncommon action states, including statistics-based, syntactic, and description-based approaches. It also looks at single-layer recognition algorithms and representation and combination techniques-based approaches. The approach used in this paper includes a methodical and thorough investigation of the existing literature, a comparison and analysis of various machine learning algorithms for identifying human behaviors, and recommendations for further study on the topic. The accuracy for the general-purpose class in the KTH dataset, which covers six acts carried out by 25 participants in

four different settings, was 97.6. The 93 episodes in the Weizmann dataset depict nine individuals performing ten typical tasks. There are 13 activities in the INRIA XMAS Motion Acquisition Sequences (IXMAS) dataset that 11 persons performed three times each throughout the course of a day. 25 people walked on a treadmill in the CMU 3D room, and four different movements of each person are captured in the CMU Motion of Body (MoBo) dataset. In the HOHA-I dataset, video tests for eight activities from 32 movies are present. The test set is begun with 20 movies that weren't utilized in the training with 211 instances, and it attained an accuracy of 56.8% (Gilbert et al.) on the movie class. The two training sets begin with 12 movies and 219 examples. The HOHA-II dataset consists of video tests for 69 different films, with 12 different activity kinds and 10 different setting types. While the test set begins with 36 movies and 884 samples, the training set starts with 823 examples and 33 films. Three shading films and four dim-scale video groupings make up the human Eva-I dataset. [4] Two people who were executing a combination of walking and running with an accuracy of 84.3% are included in the Human Eva II dataset. The six categories based on the CMU Mocap dataset contain 23 subgroups. Motion capture is done using 12 Vicon infrared MX-40 cameras, each with a 120-megapixel resolution and 100% accuracy. Using a range of space-time methodologies are several datasets (UCF Sports Action, UCF YouTube Action, and i3DPost Multi-View). This document gives a summary of various studies that examine how open datasets might be used to evaluate recognition algorithms for human movement recognition. Multi-view datasets from KTH, Weizmann, IXMAS, CMU MoBo, HOHA, HOHA-2, and Human Eva are included in the dataset. With the aid of some empirical machine learning methods, the prediction model is created, trained, and tested. Based on a statistical category, the hierarchical approaches of Wang's 11th Approach, Yin's 2010 Approach based on the Statistical Category, Accuracy of KTH (82%), Zeng's 2010 Approach based on the Statistical Category, Accuracy of KTH (92.1%), and WZMN (100%) are compared. Researchers and professionals who are developing or utilizing machine learning algorithms to understand human actions may find the study to be helpful.

4. Problem Statement

On this research work, we will find out the activity or human action recognition by using deep learning and machine learning. Both types of algorithms will be used here to predict the accuracy of activity. One of the biggest issues with smart video security is activity detection. To identify human activity in surveillance footage is a basic challenge for computer vision. [5] These applications require real-time detection functionality, but it typically takes a long period to find the actual activity.

5. Research Objective

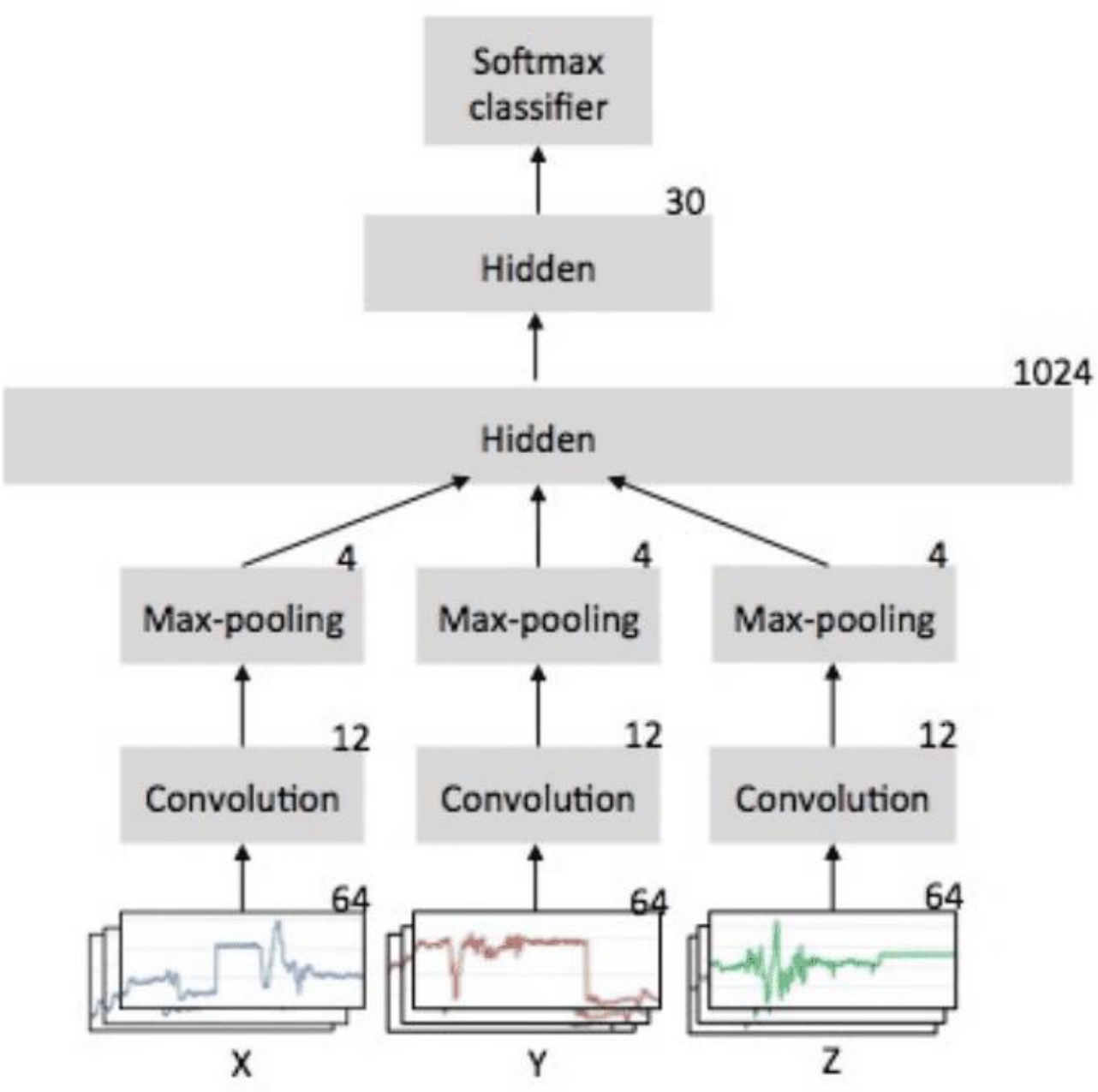
In this research work, Deep learning models like neural networks and CNN algorithms will be used for video classification. Also, some frameworks like TensorFlow, and OpenCV can be used to predict the activity. There is a large number of datasets to train the model which we will use. Deep learning as well as machine learning are both used in this study project. For machine learning and deep learning, there are different types of data. The model is able to forecast activity level thanks to deep learning techniques. We will examine the accuracy of algorithms and how they operate. Algorithms for classifiers are used with machine learning models. Classifier Techniques Logistic Regression, Naive Bayes, SVM, and Decision Tree are some of the model algorithms used. To help with data visualization and a better understanding of the data set, we will evaluate algorithmic accuracy.

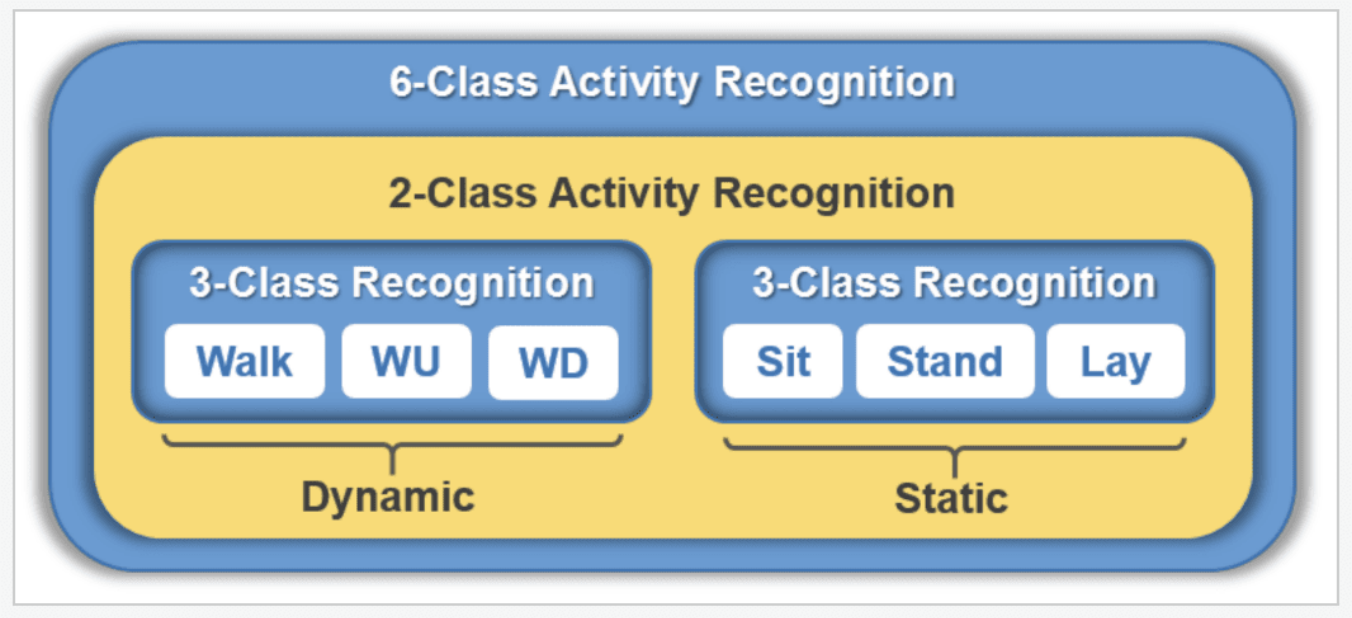
6. Research Methodology :

We need to discuss data preparation before we delve into the particular neural networks that can be applied to recognize human activity. Both varieties of neural networks that are appropriate for time series classification require special data preparation in order to create a model. That is, in a "supervised learning" manner that enables the algorithm to link signal data to a category of activities. A simple data preparation method includes dividing the input signal data into windows of signals, where a given window may have one to several seconds of observation data. This method was used for both neural networks and classical machine learning methods on hand-crafted features. Many people refer to this as a "sliding window. "Human activity recognition aims to infer the actions of one or more persons from a set of observations captured by sensors. Usually, this is performed by following a fixed length sliding window approach for the features extraction where two parameters have to be fixed: the size of the window and the shift "[6].

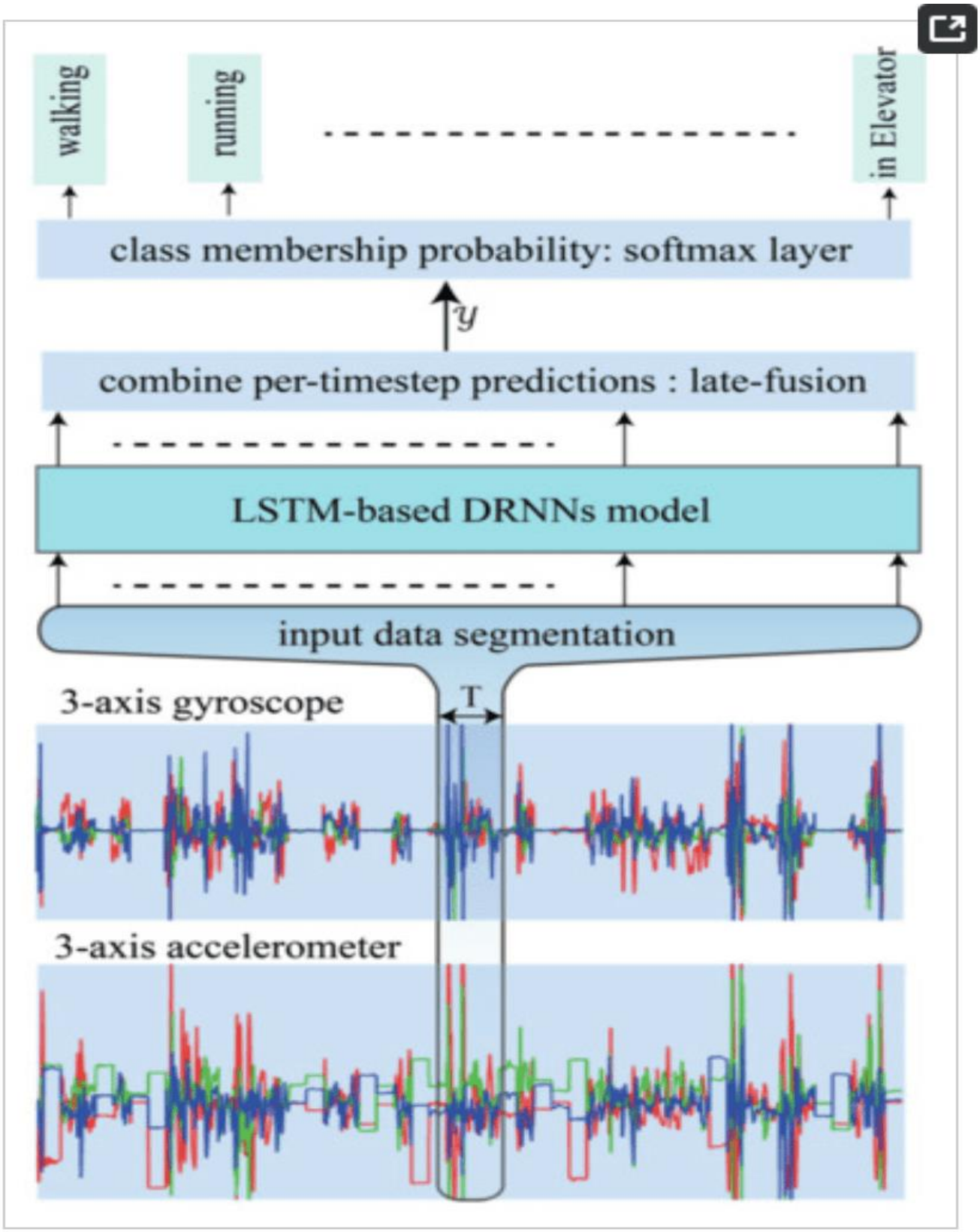
Traditionally, the collected sensor data was analyzed and condensed using techniques from the area of signal processing. These techniques involved feature engineering, which involved developing domain-specific, sensor- or signal-processing-specific features and perspectives of the initial data. On the cleaned-up data, statistical and machine learning algorithms were subsequently trained. The signal processing and domain knowledge needed to analyze the raw data and design the features needed to fit a model are a limitation of this method. Each novel dataset or sensor modality would call for this expertise. Basically, it is costly and not scalable. "However, in most daily HAR tasks, those methods may heavily rely on heuristic handcrafted feature extraction, which is usually limited by human domain knowledge. Furthermore, only shallow features can be learned by those approaches, leading to undermined performance for unsupervised and incremental tasks. Due to those limitations, the performances of conventional [pattern recognition] methods are restricted regarding classification accuracy and model generalization."

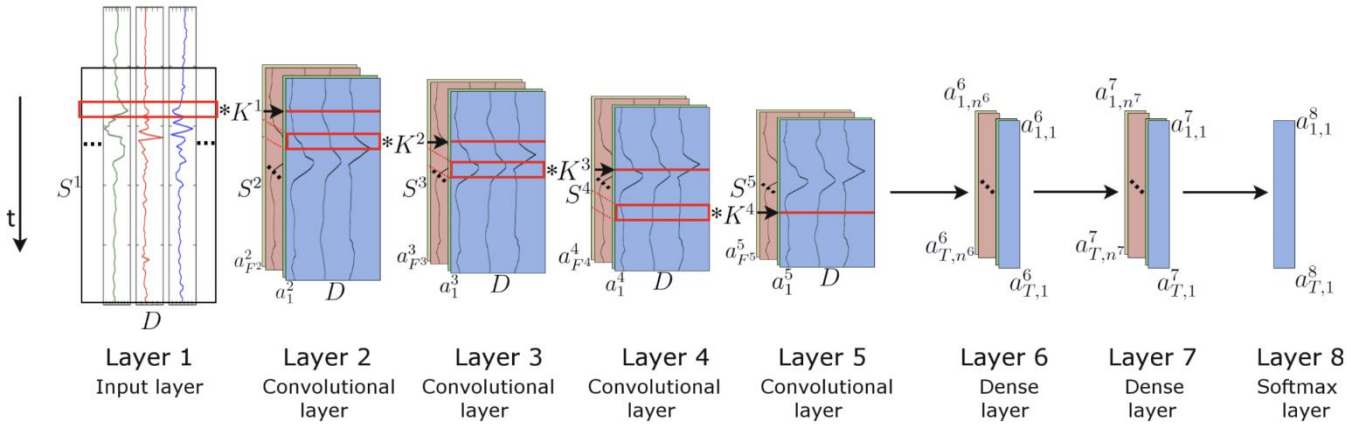
A kernel that reads the input in small segments at a time and steps across the full input field is used by convolutional layers to read an input, such as a 2D image or a 1D signal. Each read produces an input that is cast onto a filter map and serves as a representation of how the input was interpreted internally. Feature map projections are reduced to their fundamental components using signal averaging or signal maximizing techniques in pooling layers. The convolution and pooling layers can be repeated in depth to provide the input signals with numerous layers of abstraction. One or more completely connected layers that interpret what has been read and map this internal representation to a class are frequently the output of these networks. A kernel used by convolutional layers reads an input, such as a 2D image or a 1D signal, in small segments and steps across the full input field. Input from each read is projected onto a filter map, representing an internal understanding of the input. By using a signal averaging or signal maximizing method, pooling layers reduce the feature map projections to their bare minimum. Repeating the convolution and pooling layers in depth allows for numerous layers of input signal abstraction. These networks typically produce one or more fully connected layers that interpret the information received and map it to a class.[8]


Softmax classifier
30
Hidden
1024
Hidden
4
4
4
Max-pooling
Max-pooling
Max-pooling
12
12
12
Convolution
Convolution
Convolution
64
64
64
X
Y
Z



Recurrent neural networks, also known as RNNs, are a subset of neural networks that were created with the purpose of learning from sequences of data, such as a series of observations made over time or a series of words in a phrase. The long short-term memory network, or LSTM for short, is one particular form of RNN that is perhaps the most popular because of its careful design, which gets around the common challenges associated with training a stable RNN on sequence data. Recurrent neural networks, or RNNs, are a subset of neural networks designed to learn from sequences of data, such as a series of observations made over time or a series of syllables in a phrase. Because of its thoughtful design, which avoids the typical difficulties connected with training a stable RNN on sequence data, the long short-term memory network, or LSTM for short, is one specific type of RNN that is possibly the most well-known. Recurrent neural networks, or RNNs, are a subset of neural networks designed to learn from sequences of data, such as a series of observations made over time or a series of syllables in a phrase. Because of its thoughtful design, which avoids the typical difficulties connected with training a stable RNN on sequence data, the long short-term memory network, or LSTM for short, is one specific type of RNN that is possibly the most well-known. [10]


walking
running
in Elevator
class membership probability: softmax layer
y
combine per-timestep predictions : late-fusion
LSTM-based DRNNs model
input data segmentation
3-axis gyroscope
T
3-axis accelerometer



7. Utilization:

The proposed study has the potential to push existing systems to their limits. In this research our study can be useful in recognising human activity from video footage using cutting-edge AI algorithms. Data preprocessing, feature engineering, and various machine learning algorithms like convolutional and recurrent neural networks are all part of the process. In order to prepare the data, the data from the input signal is split into windows. Each window can contain observations from one second or more than one second. After this, feature engineering is used to create sensor- or signal-processing-specific domain-specific features. This information is used to improve the performance of machine learning algorithms by providing new perspectives on the original data. [9] In this paper, we compare the performance of common deep learning architectures, such as long short-term memory networks and convolutional neural networks, in recognising human actions. Our findings demonstrate the effectiveness of these strategies in identifying human behaviour. Its applications include medicine, athletics, and perhaps security.

Overall, our study demonstrates how the use of cutting-edge machine learning and deep learning techniques can improve human activity recognition from video data. Impacts in areas as diverse as surveillance, athletics, and medical diagnostics are possible.

9. References

[1] Uday, S.S., Pavani, S.T., Lakshmi, T.J. and Chivukula, R., 2022. Classifying Human Activities using Machine Learning and Deep Learning Techniques. arXiv preprint arXiv:2205.10325.

[2] Karácsony, T., Loesch-Biffar, A.M., Vollmar, C., Rémi, J., Noachtar, S. and Cunha, J.P.S., 2022. Novel 3D video action recognition deep learning approach for near real time epileptic seizure classification. Scientific Reports, 12(1), p.19571. American cancer Society, "facts spring 2014| Leukemia Lymphoma Society: Fighting Blood Cancer, Revised April 2014.

[3] Zhang, H.B., Zhang, Y.X., Zhong, B., Lei, Q., Yang, L., Du, J.X. and Chen, D.S., 2019. A comprehensive survey of vision-based human action recognition methods. Sensors, 19(5), p.1005. *Cancers of the haematopoietic system TV* Ajithkumar, HM Hatcher, in Specialist Training in Oncology.

[4] Ke, S.R., Thuc, H.L.U., Lee, Y.J., Hwang, J.N., Yoo, J.H. and Choi, K.H., 2013. A review on video-based human activity recognition. Computers, 2(2), pp.88-131.

[5] Beddiar, D.R., Nini, B., Sabokrou, M. and Hadid, A., 2020. Vision-based human activity recognition: a survey. Multimedia Tools and Applications, 79, pp.30509-30555.

[6] Ramasamy Ramamurthy, S. and Roy, N., 2018. Recent trends in machine learning for human activity recognition—A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), p.e1254.

[7] Hussain, Z., Sheng, M. and Zhang, W.E., 2019. Different approaches for human activity recognition: A survey. arXiv preprint arXiv:1906.05074.

[8] Gaglio, S., Re, G.L. and Morana, M., 2014. Human activity recognition process using 3-D posture data. IEEE Transactions on Human-Machine Systems, 45(5), pp.586-597.

[9] Zeng, M., Nguyen, L.T., Yu, B., Mengshoel, O.J., Zhu, J., Wu, P. and Zhang, J., 2014, November. Convolutional neural networks for human activity recognition using mobile sensors. In 6th international conference on mobile computing, applications and services (pp. 197-205). IEEE.

[10] Mutegeki, R. and Han, D.S., 2020, February. A CNN-LSTM approach to human activity recognition. In 2020 international conference on artificial intelligence in information and communication (ICAIIC) (pp. 362-366). IEEE.