International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7222
***Article Received**: 26 April 2023 **Revised:** 12 June 2023 **Accepted:** 30 June 2023*

Exploration of Deep Learning Models for Video Based Multiple Human Activity Recognition

Jitha Janardhanan[1], Dr. S. Umamaheswari[2]
[1]Research Scholar,Department of Computer Science
Dr.G.R.D College Of Science
Coimbatore,Tamil Nadu, India
e-mail: jithajanardhanan@gmail.com
[2]Associate Professor,Department of Computer Science
Dr.G.R.D College Of Science
Coimbatore,,Tamil Nadu, India

Abstract—: Human Activity Recognition (HAR) with Deep Learning is a challenging and a highly demanding classification task. Complexity of the activity detection and the number of subjects are the main issues. Data mining approaches improved decision-making performance. This work presents one such model for Human activity recognition for multiple subjects carrying out multiple activities. Involving real time datasets, the work developed a rapid algorithm for minimizing the problems of neural networks classifier. An optimal feature extraction happens and develops a multi-modal classification technique and predicts solutions with better accuracy when compared to other traditional methods. This paper discussing on HAR prediction in four phases namely (i) Depthwise Separable Convolution with BiLSTM (DSC-BLSTM); (ii) Enhanced Bidirectional Grated Recurrent Unit with Long Short Term Memory (BGRU-LSTM); (iii) Enhanced TimeSformer Model with Multi-Layer Perceptron Neural Networks classification and (iv) Filtering Single Activity Recognition are described.In comparison to previous efforts like the DSC-BLSTM and BGRU-LSTM classifications, the experimental result of the ETMLP classification attained 98.90%, which was more efficient. The end outcome revealed that the new model performed better in terms of accuracy than the other models.

Keywords:: Data mining, Deep Neural network, LSTM, GRU, TimeSformer, Feature extraction.

I. INTRODUCTION

The method of extracting formerly undiscovered knowledge and finding intriguing patterns from a large collection of data is known as data mining. Due to the extensive use of information technology and recent developments in multimedia systems, users now have access to a tremendous amount of multimedia data. As it includes text, image, meta-data, visual, audio, and visual data, a video is an example of multimedia data. It is widely employed in a variety of significant applications, including security and surveillance; entertainment, medicine, educational programmes, and sports. As one of the primary research challenges for the data-mining community aims to identify and explain intriguing patterns from the vast volume of video data.

Despite being a simple process to collect and store video data, it can be difficult to extract information from it. As processing video data using computer vision algorithms necessitates features in a structured manner, one of the most crucial stages is to convert the video data from an unstructured data set into a structured data set [1, 2]. Deep neural network techniques [3-6] are used to mine the video attributes from the video key frame prior to the data-mining algorithms being applied. This removes digitalization noise and illumination changes to prevent false positive detection [7].

Human Activity Recognition (HAR) that occurs in video databases is analyzed and predicted in this work. One of the difficulties in identifying a person's physical activity from their movement pattern in a specific environment is HAR.

In HAR, a variety of computer vision methods, ranging from image processing to machine learning to deep learning, are used to identify human activity. There are numerous hybrid strategies that combine two separate techniques to identify the activity, such as the use of both image processing and machine learning. In many methods, human behavior is detected by combining two or more machine learning techniques. Since the publication explored activity recognition from video, a number of neural network-based deep learning techniques, including MobileNET [8], Long Short Term Memory (LSTM) [9], [10], and Transformers methodologies, are available. These methods significantly increased the effectiveness of detecting the actions. The goal of this work is to combine adaptive deep learning methods to increase the effectiveness of real-time activity recognition.

This work involves an extensive study on the deep learning architectures for the recognition in terms of robustness, scalability and storage effectiveness. The main objective of this work is to extend the deep learning approaches on target classification. In a traditional HAR system [11-12], smart devices collect raw data from sensors. Machine learning

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7222
***Article Received**: 26 April 2023 **Revised:** 12 June 2023 **Accepted:** 30 June 2023*

methodologies are used to extract data associated with it. The raw data from sensors are processed to retrieve features that extract relevant information. The following figure 1 depicts the overview of the video action recognition approach.

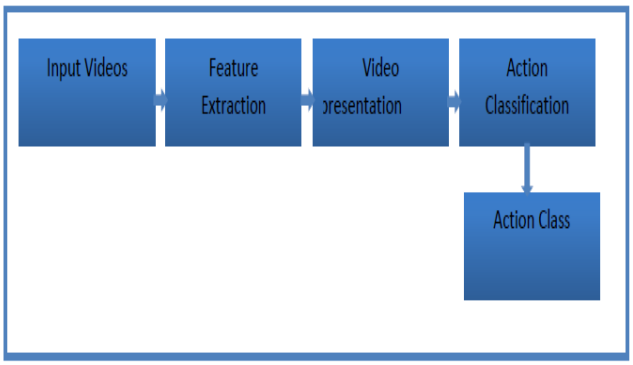
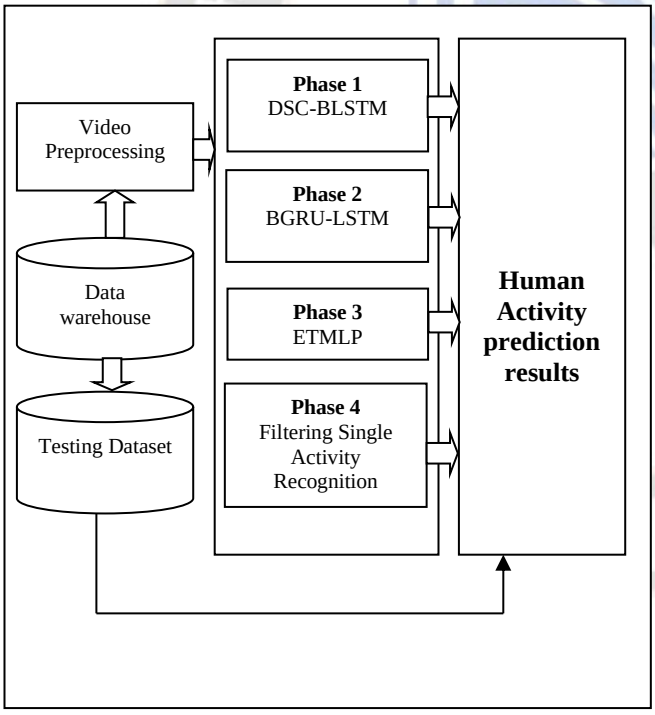


Figure 1. Overview of the video recognition approach

This research seeks to develop an effective data mining method of video preprocessing, feature extraction and classification effectively to recognize the multi-class activity classification among real-time surveillance video dataset. Each stage has a different method described in figure 2.



The remainder of the paper is structured as follows: Sect. 2 provides specifics on Related Work. In Section 3, research approaches for phases 1, 2, and 3 of the categorization process are used to predict HAR, and Section 4 describes the experimental findings. Section 5 has the conclusion.

II. RELATED WORK

Anindya Das et.al, (2019) [13] presented a comprehensive survey on the human based activity recognition along with the challenges based on wearable, environmental and smartphone sensors. The focus factors on the data preprocessing part, segmentation and filtering. A list of sensor devices and applications for use are discussed. Some of the benchmark datasets were analyzed that included information on sensors, attributes and activity classes. The discussion concluded with an investigation of activity recognition methods using some of the benchmark datasets.

Albert Florea et.al (2019) [14] in his work proposed a neural network model that was built using Keras API with Tensorflow Library. The architectures that were investigated were RNN, Residual neural network, Visual Geometry and RCNN. The weights for the neural network base models were initially set using weights from ImageNet. Results from the feature extraction utilizing the base models and ImageNet weights were encouraging. The implementation of LSTM spatio-temporal sequence prediction and Deep Learning with thick layers was effective.

Nidhu Dua et.al (2021) [15] presented a Deep NN model which combines CNN with GRU that executes regular attribute extraction and categorization. Raw data from wearable sensors were used for the training with nominal pre-processing. The local attributes are handled by CNN whereas the long-term dependencies were handled by GRU layers. The model has three head architecture with different convolutional filter dimensions.

Anurag et.al, (2021) [16] presented a novel model that mines spatio-temporal tokens from the key video and is then encoded by series of encoders. Long sequence video frames are handled through the factorization of variants. Usually transformer models are effective for larger datasets. Kinetics 400 and 600, Epic kitchens are the datasets used that outperforms other existing models based on deep 3D. A pure transformer model is proposed for video classification. The main operation performed is the self-attention computed with a sequence of spatiotemporal tokens.

Jitha Janardhanan and S. Umamaheswari (2022) [17] discussed a DSC-BLSTM method which reduces both the number of constraints that may be discovered and the execution duration of the pooled training and testing method, is one of the recommended network system's redeeming qualities. The bidirectional LSTM technique can be used to merge the positive and negative time directions.

Jitha Janardhanan and S. Umamaheswari (2023) [18] talked about how the trend of multiple human activity detection in smart surveillance is fraught with challenges, such as the need for real-time examination of massive quintiles of video data while keeping minimal processing complication. Scalable,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7222
***Article Received**: 26 April 2023 **Revised:** 12 June 2023 **Accepted:** 30 June 2023*

effective object detection and deep skip connections are already present in EfficientNET systems. Using the COCO 2017 picture dataset, the scaled model must predict the items in order to work. this study exclusively uses the compound scaling method. It just recognizes objects; it doesn't analyse various human activities or the interactions between different classes of people.

Jitha Janardhanan and S. Umamaheswari (2022) [19] illustrated action recognition based on videos as one of the most popular techniques for comprehending human behaviour. Comparing videos to image-based action recognition, videos offer substantially more information. Over the past ten years, reducing action ambiguity, various dataset-focused studies, novel models, and learning techniques have improved video action recognition. The authors developed a HAR to introduce a classification algorithm that uses self-attention over patches and human regions called ETMLP Neural Networks Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts if possible. True-Type 1 or Open Type fonts are preferred. Please embed symbol fonts, as well, for math, etc.

III. RESEARCH METHODOLOGY

The research methodology performs the HAR process divided into four phases. The first phase involves DSC-BLSTM in real-world kinetics 400 video datasets [20]. The second phase involves an enhanced BGRU-LSTM in real-time Bengaluru home apartment datasets. The third phase involves an Enhanced ETMLP in real-time Bengaluru home apartment datasets. The fourth phases provides an enchancement to the third one by filtering into single activity.The Human activity recognition process considers the overall process architecture diagram as described in figure 3.

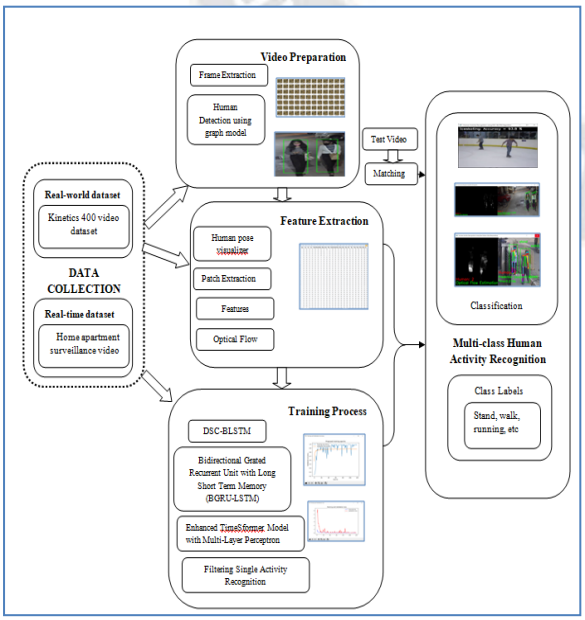


Figure.3: Proposed overall architecture Diagram

Phase 1: DSC-BLSTM

In phase I work attempts to classify the human activity recognition based on Feature extraction and Deep neural network classification. This technique is an improved version of the two-layer, MobileNetV2 convolution network method for feature extraction. DC (Depthwise Convolutions) and PC (Point wise Convolutions) are two examples. The proposed MobileNetV2 uses three-by-three DSC, which requires nine times less computing than conventional convolutions. Figure 4 gives an explanation of the training model.

To achieve the optimal features, best features are implemented based on DSC and saved in trained datasets. Features are combined in the third phase, and results are run through a classification learner in the last step. The prediction activity result is then obtained by using the Depthwise Separable Convolution with Bidirectional-LSTM (DSCBLSTM) classifier to determine the test frame's similarity matrix value to the trained model.

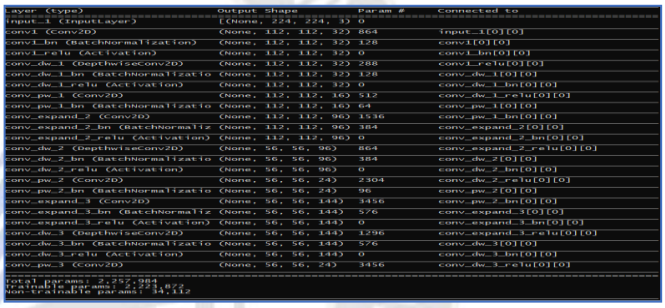


Figure.4: DSC Model result

Phase 2: BGRU-LSTM

Using enhanced human posture estimation and a range of human activities, a real-time home apartment surveillance video dataset is used in phase II to choose the relevant features. This is done by combining the strength of the EfficientNET feature extraction with BGRU-LSTM classification. The EfficientNet model extracts 36 features of human pose. Categorical and Real value attributes make up the data extraction features format. The feature extraction outcomes are shown in Figure 5.

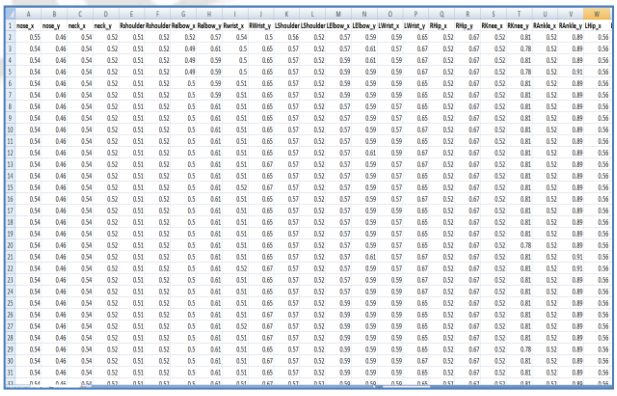


Figure.5: Feature extraction result

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7222
***Article Received**: 26 April 2023 **Revised:** 12 June 2023 **Accepted:** 30 June 2023*

The bidirectional GRU-based LSTM deep learning method to identify various human actions using real-time surveillance video datasets. The proposed method was combined with two different GRU with LSTM models to produce Bidirectional GRU with LSTM. Since a bidirectional model was chosen, both forward and backward cells were employed. The input data are sent to the GRU's input sequence, and after passing through the LSTM, they are then created. Data from the GRU's output layer is fed into the LSTM's input layers for convolution. Before delivering them to the fully linked layer, the ReLU layer operates them after that. The classification outcome is then output by SoftMax. In Figure 6 shows an improved EfficientNet with Tensor flow pose feature extraction method with BGRC-LSTM flow diagram.

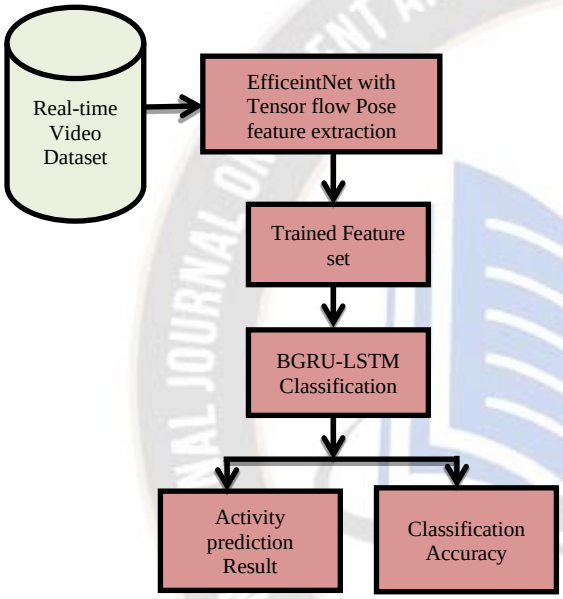


Figure .5: BGRU-LSTM Algorithm Flow

Phase 3: ETMLP

Using an Improved EfficientNet Feature extraction and ETMLP classification method, Phase III comprises offering numerous solutions with the optimal feature extraction in a real-time home apartment surveillance video dataset. The EfficientNet transfer learning technique initially handles the estimation of the human position. EfficientNet is unique in that it can achieve good accuracy with a limited set of parameters. Second, an improved variant of the BGRU-LSTM convolution network classification model is the ETMLP technique. The DSC- BLSTM model improves recognition accuracy but does not concentrate on multi-human activities that contain annotations. The present condition of multiple-human activity recognition in different real-time CCTV home apartment movies is provided in this research along with the issues that require further consideration. The classification algorithm technique of ETMLP Neural Networks learns trustworthy spatial and temporal representations and deploys them immediately to a range of downstream tasks. In a range of real-time video datasets, the results produced are extremely accurate. Figure 6 shows the ETMLP process flow.

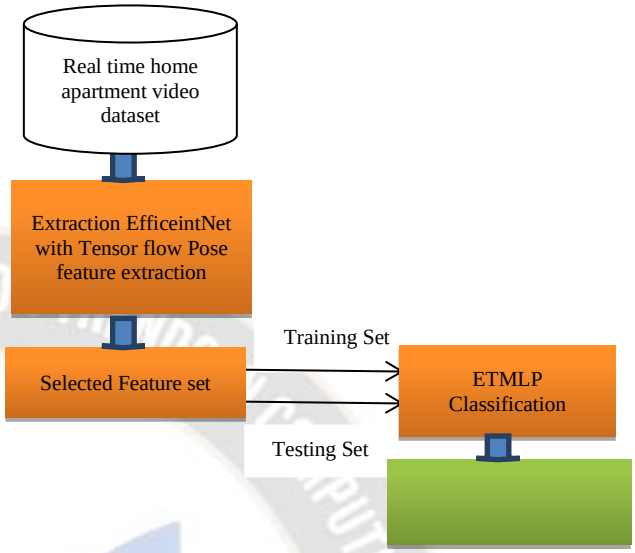


Figure 6: ETMLP Prediction Classification

Phase 4: Filtering Single Activity Recognition

A Filtering Single Activity Recognition using User Constraint Prediction in phase 4 aims to provide filter a particular human activity prediction results in real-time CCTV Home apartment video datasets. This approach is an improved version of the ETMLP model. With the help of contextualized human activity data and a user-specified specific activity in a video, this method may predict the posture of a human activity by using visual regions for pose estimation. In figure 7 describes the process of filtering singe activity recognition.

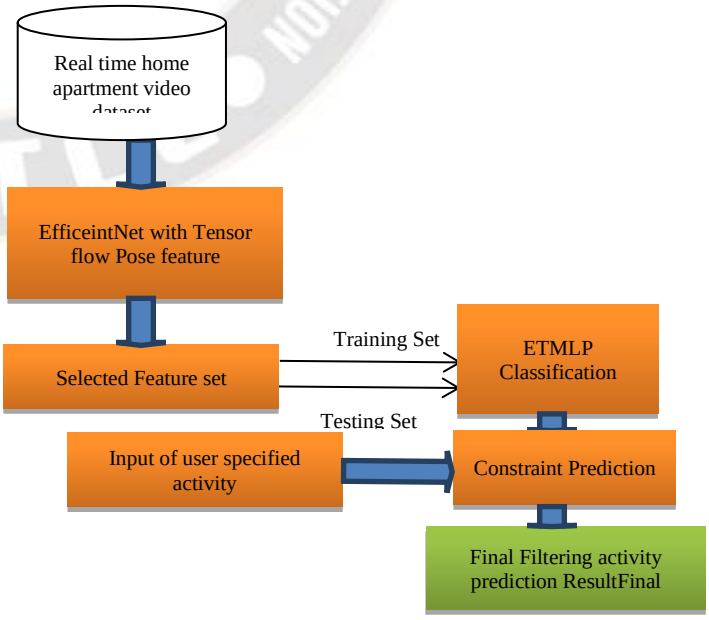


Figure .6: Filtering Singe Activity Recognition

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7222
***Article Received**: 26 April 2023 **Revised:** 12 June 2023 **Accepted:** 30 June 2023*

IV. RESULT AND DISCUSSION

The experimental results evaluate the overall performance of DSC-BLSTM, BGRU-LSTM and ETMLP Neural Networks classification algorithms. The three types of video datasets collected are one of real-world and remaining two of real-time CCTV Home apartment datasets. The experimental findings are run on a Windows 10 computer running Python 3.8 simulations, an Intel I7 series 3.40 GHz four-core processor, and 16GB of main memory. Figures 7 & 8 shows the proposed ETMLP Classification model predicts the multi-class human activity prediction of Training and validation accuracy; Training and validation loss. Figure 9 shows the Performance measure execution result. Meanwhile, the precision and recall measures of existing Bidirectional Long Short Term Memory (Bidir-LSTM) [21] method with all three proposed phases DSC-BLSTM, BGRU-LSTM and ETMLP shown is table 1 and figures 10 and 11. The classification accuracy and f1-score measures of existing Bidir-LSTM [21] method with all three phases DSC-BLSTM, BGRU-LSTM and ETMLP shown are table 3 and figures 12 and 13.

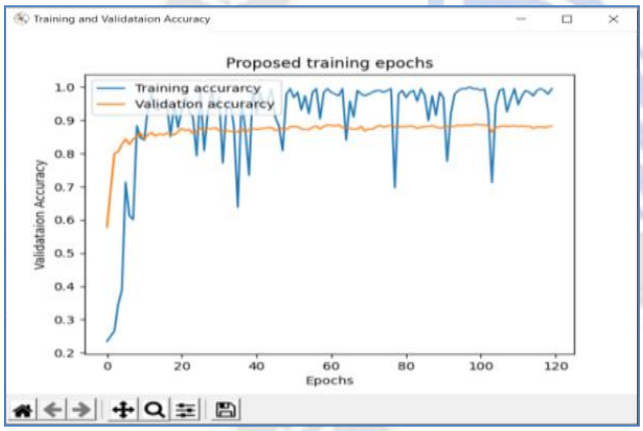


Figure.7: Overall ETMLP training validation Accuracy plot result

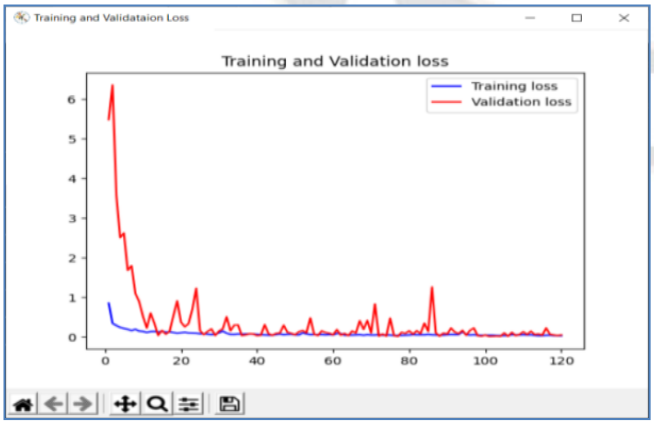


Fig.ure8: Validation of the whole ETMLP training Loss plot

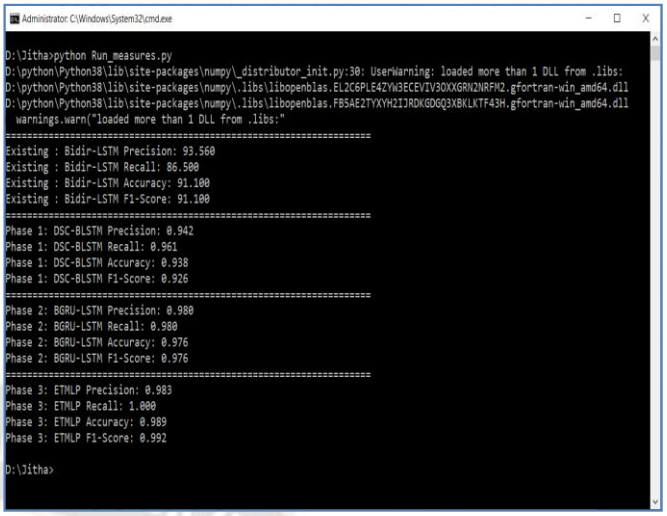


Figure.9: Performance measure execution result conclusion

TABLE 1: COMPARISON OF PRECISION AND RECALL MEASURES WITH PROPOSED DSC-BLSTM, BGRU-LSTM AND ETMLP ALGORITHM OF REAL-TIME CCTV VIDEO DATASET

Methods	Bidir-LSTM	DSC-BLSTM	BGRU-LSTM	ETMLP
Precision	93.56	94.2	98.0	98.3
Recall	86.50	96.1	98.0	100

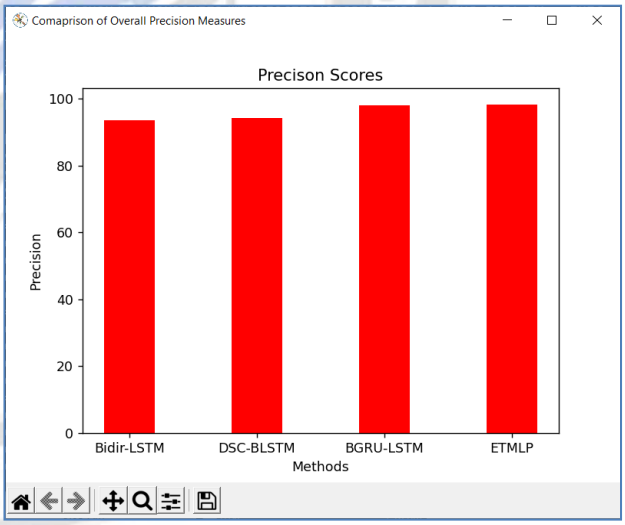


Figure.10: Comparison of Overall Precision scores of existing and proposed methods

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7222
***Article Received**: 26 April 2023 **Revised:** 12 June 2023 **Accepted:** 30 June 2023*

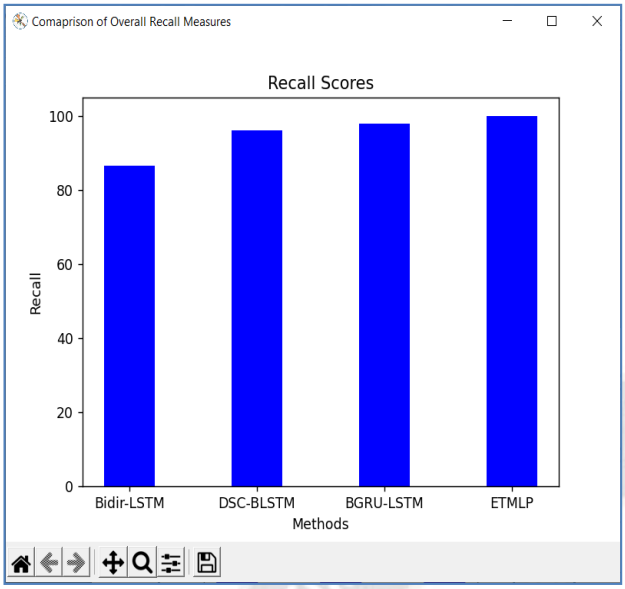


Figure.11: Comparison of Overall Recall scores of existing and proposed methods

TABLE 2: COMPARISON OF THE PROPOSED DSC-BLSTM, BGRU-LSTM, AND ETMLP ALGORITHMS OF CLASSIFICATION ACCURACY AND F1-SCORE MEASURES

Methods	**Bidir-LSTM**	DSC-BLSTM	BGRU-LSTM	ETMLP
Accuracy	91.1	93.8	97.6	98.9
F1-Score	91.1	92.6	97.6	99.2

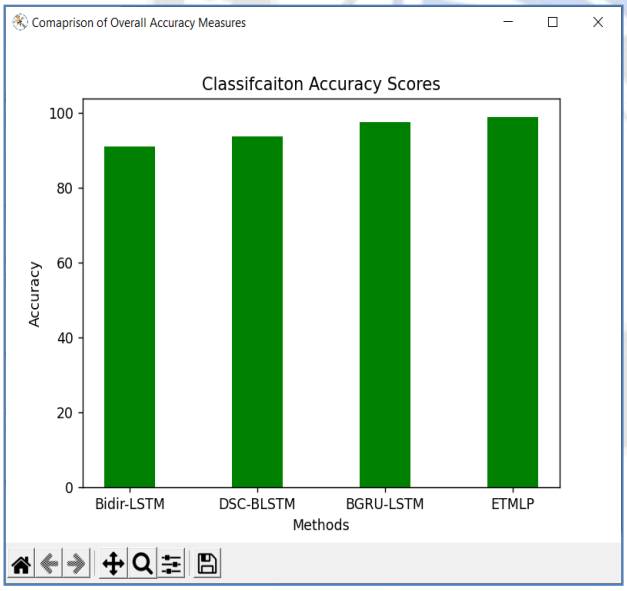


Figure.12: Comparison of Overall classification Accuracy scores of existing and proposed methods

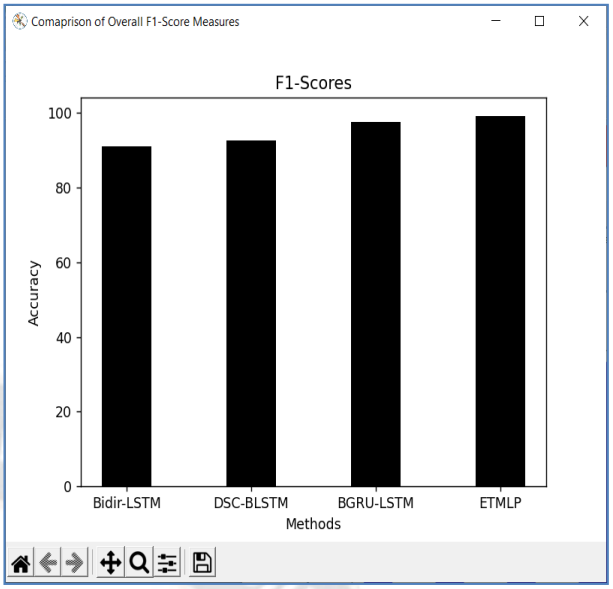


Figure.13: Comparison of Overall F1-score of existing and proposed methods

V. CONCLUSION

The best classification technique in data mining has been examined and developed in this study for the accurate prediction of human behavior in real-world and real-time video datasets. The HAR prediction in video datasets took many categories classes including stand, walk, running, etc.. The DSC-BLSTM technique was used in phase 1 of the work to choose the suitable characteristics for the classification method. This approach uses DSC and BLSTM convolution, which helps to lower the overall computing cost of the training and testing technique as well as the amount of parameters that may be learned. In phase 2 36 features were selected by improved EfficientNet model and BGRU-LSTM classification was used to predict the multiple HAR prediction in different condition. By supplementing uniform categorization tokens with contextualized human activity data and interacting with them in this manner. In phase 3 finally introduces the ETMLP Neural Network classification algorithm, which applies self-attention over the patches and human regions to estimate human pose. In the end, this research's classification accuracy, which was 98.90%, outperformed classifications from earlier studies using the DSC-BLSTM and BGRU-LSTM methods.

REFERENCES

[1] Anguita .D, Ghio .O, Oneto.L, Parra .X, Reyes-Ortiz .L, "Energy Efficient Smartphone-Based Activity Recognition Using Fixed-Point Arithmetic," Journal of Universal Computer Science, vol. 19, no. 9, pp. 1295–1314, 2013.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lu˘ci´c, and Cordelia Schmid, "Vivit: A video vision transformer". ICCV, 2021.

[3] J. Schmidhuber, "Deep Learning In Neural Networks: An Overview", Neural Networks, vol. 61, pp. 85-117, 2015.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7222
***Article Received**: 26 April 2023 **Revised:** 12 June 2023 **Accepted:** 30 June 2023*

[4] N. Y. Hammerla, S. Halloran, and T. Ploetz, "Deep, Convolutional, And Recurrent Models For Human Activity Recognition Using Wearables," arXiv preprint arXiv:1604.08880, 2016

[5] Unnam, A. K. ., & Rao, B. S. . (2023). An Extended Clusters Assessment Method with the Multi-Viewpoints for Effective Visualization of Data Partitions. International Journal of Intelligent Systems and Applications in Engineering, 11(2s), 81–87. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2511

[6] S. M¨unzner, P. Schmidt, A. Reiss, M. Hanselmann, R. Stiefelhagen, and R. D¨urichen, "Cnn-Based Sensor Fusion Techniques For Multimodal Human Activity Recognition," in Proceedings of the 2017ACM International Symposium on Wearable Computers, ser. ISWC '17. New York, NY, USA: ACM, 2017, pp. 158–165.

[7] A. Jain and V. Kanhangad, "Human Activity Classification in Smartphone's Using Accelerometer and Gyroscope Sensors," IEEE Sensors Journal, vol. 18, no. 3, pp. 1169-1177, 1 Feb.1, 2018

[8] Li, S., Seybold, B., Vorobyov, A., Lei, X., Jay Kuo, C.C.: Unsupervised video object segmentation with motion-based bilateral networks. In: ECCV. pp. 207-223, 2018

[9] Sandler .M, Howard .A, Zhu .M, Zhmoginov .A, and Chen .L. C, "Mobilenetv2: Inverted Residuals And Linear bottlenecks." IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp 4510–4520.

[10] Francisco Javier Ordóñez, Daniel Roggen, "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition", Wearable Technologies, Sensor Technology Research Centre, Sensors, 2016.

[11] N Srivastava, E Mansimov, R Salakhudinov," Unsupervised Learning Of Video Representation Using LSTM", International conference on machine learning, pp 843–852, 2015.

[12] Aggarwal, J. K., and Ryoo, M. S., "Human activity analysis: a review". ACM Computing Survey, 2011, pp.1–43.

[13] A. Jain and V. Kanhangad, "Human Activity Classification in Smartphone's Using Accelerometer and Gyroscope Sensors," IEEE Sensors Journal, vol. 18, no. 3, pp. 1169-1177, 1 Feb.1, 2018.

[14] Das Antar, Anindya & Ahmed, Masud & Ahad, Md Atiqur Rahman. (2019). Challenges in Sensor-based Human Activity Recognition and a Comparative Analysis of Benchmark Datasets: A Review. 134-139. 10.1109/ICIEV.2019.8858508.

[15] AlbertFlorea and FilipWeilid, "Deep Learning Models for Human Activity Recognition", Book series, ComputerEngineering, 2019.

[16] Nidhi, Dua., Shiva, Nand, Singh., Vijay, Bhaskar, Semwal. (2021). Multi-input CNN-GRU based human activity recognition using wearable sensors. Computing, 103(7):1461-1478. doi: 10.1007/S00607-021-00928-8

[17] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lu˘ci´c, and Cordelia Schmid, "Vivit: A video vision transformer". ICCV, 2021.

[18] Jitha Janardhanan and S. Umamaheswari, "Vision based Human Activity Recognition using Deep Neural Network Framework" International Journal of Advanced Computer Science and Applications(IJACSA), 13(6), 2022.

[19] Jitha Janardhanan and S. Umamaheswari, "Recognizing Multiple Human Activities Using Deep Learning Framework", International Inforamtion and Engineering Technology Association (IIETA), Vol. 36, Iss. 5, (Oct 2022): 791-799. DOI:10.18280/ria.360518.

[20] Jitha Janardhanan and S. Umamaheswari, "Multi-Class Human Activity Prediction using Deep Learning Algorithm", Int J Intell Syst Appl Eng, vol. 10, no. 4, pp. 480–486, Dec. 2022.

[21] Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, "The Kinetics Human Action Video Dataset", 2017.

[22] Yu Zhao, Rennong Yang, Guillaume Chevalier, Ximeng Xu, and Zhenxing Zhang, "Deep Residual Bidir-LSTM for Human Activity Recognition Using Wearable Sensors", Mathematical Problems in Engineering, ,Volume 2018