Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

An Enhanced Approach for Graph Neural Network Human Activity Recognition Using Deep Learning Technique

Velantina V[1], Dr. V. Manikandan[2], Dr. P. Manikandan[3]

[1]*Research scholar, Department of Computer science and engineering, Jain University, Karnataka, India,*

[2] Assistant Professor, Department of CSE, Jain University, Karnataka, India,

[3]Associate Professor, Department of CSE, Jain University, Karnataka, India,

ARTICLE INFO

Received: 18 Nov 2024

Revised: 22 Dec 2024

Accepted: 28 Dec 2024

ABSTRACT

Human activity recognition (HAR) has emerged as an essential area of study in video analysis. It has evolved substantial attention as a result of its applications in a diverse range of disciplines, such as healthcare, surveillance, and human-computer interaction. But achieving real-time robustness remains a challenge due to environmental factors such as occlusions and shifts in motion dynamics. This study proposes the Temporal Graph Neural Network (TGNN) for the recognition of human activity in order to integrate multimodal feature extraction and temporal graph adaptive fusion techniques. For the purpose of structured representation learning, the temporal graph neural network is implemented to encapsulate the spatial and temporal dependencies of human activities, Contrastive learning is used to which refines feature discrimination, thereby improving generalization across a variety of conditions. An EfficientNetB0-based whale optimization algorithm is further integrated for optimal hyperparameter tuning, which reduces the computational overhead Extensive experiments were carried out on HMDB51 datasets with an accuracy of 90.06%. This method is superior to existing HAR baseline models. It offers a real-time and scalable HAR solution that is appropriate for deployment in various applications.

Keywords: Human activity recognition, adaptive fusion techniques, temporal graph neural networks, human-computer interaction, efficientnetB0, whale optimization algorithm, deep learning.

1. INTRODUCTION

Human activity recognition (HAR) is an area of active investigation because it is used in fields including sports analytics, smart surveillance technologies, healthcare, and human-computer interface. Humans perform basic daily tasks like driving, cleaning, gaming, and so forth, many other fundamental movements, including standing, sitting, bending, and sprinting. Any system that involves human-computer interaction must be designed with an understanding of the human process in mind. The eligible features will be extracted from the provided cues by human activity recognition algorithms [1]. Wearable computing, computer vision, machine learning, and pattern recognition are among the fields where HAR finds extensive use. This HAR system uses raw data to identify the various actions of the person. It has uses in collaboration between humans and robots, recreational sports, video surveillance, gaming, digital marketplace, defense, rehabilitation, and gesture language recognition [2–3].

In the subject of HAR, video cameras are frequently used external devices. A range of visual sources, including RGB, RGB-D, and infrared cameras, are used by camera-based HAR systems to extract actions from the video data, depending on the activity that the camera recorded. Inferring actions based on human engagement with the environment is made easier by environmental and auditory data. HAR techniques do not record events in all three dimensions, even though the majority of them are vision-based and focus on identifying actions from monocular RGB movies. A quick development of cost-effective three-dimensional data collection in 3D activity detection [4].

The ability of graph neural network to accurately model spatial-temporal patterns is essential for accurately identifying complicated human activities. To improve comprehension of the dynamics of motion, graph structures, for example, can be used to model the interactions between various body joints [5]. As it relates to the different tasks being carried out, graph-based HAR systems have actually been demonstrated to surpass traditional CNN-centric models in terms of accuracy and resilience [6]. When there are overlapping activities, such as when several actions overlap in time or smoothly switch from one action to another in a brief amount of time, HAR systems have difficulty correctly identifying those scenarios [7]. Because of its limited processing power and battery life, any working design must be small, resource-efficient, and maintain high accuracy metrics. so, the potential approaches under investigation today have shown shortcomings in multi-modal analysis with respect to image, video, and sensor data. These system's overall performance is decreased across datasets and contexts due to poor feature alignment between modalities, which causes weaknesses rather than strengths [8].

The proposed deep learning-based method combines novel preprocessing components with advanced feature extraction and hyperparameter optimization strategies to resolve existing issues. The processing pipeline uses Gradient-based Joint Histogram Equalization to update image contrast alongside noise reduction features for better resolution of visible characteristics. VGG19 functions as the extractive feature technique because it uses its deep CNN capabilities to detect multi-scale patterns which provide detailed information in the data. Utilizing the bio-inspired optimization technique WOA helps tune learning rate and dropout hyperparameters in order to run an efficient model training process. The classification task relies on EfficientNetB0 as a model which strikes an ideal balance between computational capacity and accurate predictions. The pipeline achieves better multi-modal data interconnection through systematic synchronization techniques which enhance recognition of accuracy during complex usage scenarios. The main objectives of this study consist of:

- To design a dynamic temporal graph network that integrates RGB, depth, and skeleton data for comprehensive human action recognition.
- To enhance multimodal feature formation with better discrimination, constructing positive and negative pairs across modalities and time steps to train the model to differentiate human activities that are similar from activities that are dissimilar.
- To Use the WOA to tune the model's optimization parameters for performance and stability.
- Evaluating model performance through high accuracy and solid evaluation metrics.

2. RELATED WORK

Human action recognitionacquired significant attention for its ability to comprehensively detect activity recognition. The framework was proposed as a dual-stage CNN with cross-validation. An F1-score of 84.6% was achieved. Despite the fact that numerous researchers have developed precise algorithms for identifying human behaviors, there is still a significant amount of scope for improvement. The hyper-parameter optimization strategy has not been successful in accurately identifying human actions in previous studies [9]. Despite GCN's exceptional performance has been primarily driven by techniques such as CNN, LSTM, and others. In order to resolve this discrepancy, a deep learning model that employs graph convolution networks (GCN) is implemented, a GCN model was employed to identify actions. In order to mitigate the constraints imposed by the scarcity of publicly accessible action datasets, the model's performance was enhanced through feature extraction, fine-tuning, and curriculum learning. Accuracy increased by 20% to 30% as a result of features and fine-tuning, eventually achieving 82.24%.[10].

A sensor employing an in-embedded neural network in real-time, an innovative approach to human activity recognition (HAR). The low-cost a inertial measurement unit (IMU) worn by the subject on the chest to capture their motion. Employing a convolutional neural network (CNN) on the microcontroller to recognize and predict the action, the sensor eliminates the requirement for extra processing hardware. The paper provides a complete description of the sensor and method employed to predict real-time human actions. Experimental results confirm the better inference capability and real-time accuracy of the proposed method for real-time embedded activity recognition [11].
Semantic2Graph was proposed to address the high energy costs and low precision issues associated with previous video action segmentation methods. In order to enhance performance and decrease processing time, the model integrates semantic edge connections with multi-modal attribute features into its graph-based structure. The method's complexity presents prospective challenges, but it enables the identification of dependencies over a longer time with fewer resources [12]. MMCL is a inter modal collaborative-learning framework that enhances the efficacy of human pose based action recognition. Despite the fact that skeletons remain efficient during inference phases, users of this approach resolve skeletal system limitations by incorporating multimodal large language models (LLMs) throughout the training process. The model provides superior results in conjunction with generalization capabilities; however, this is achieved at the expense of multimodal integration challenges [13].
Hyper-MV serves as a framework for multi-view event-based action recognition, which integrates data from multiple perspectives to overcome the constraints of single-view recognition. The process of feature integration is enhanced through the development of a hypergraph neural network. The system's efficacy is enhanced by its high accuracy; however, the complexity of the system, which is a result of the integration of multiple views, is a limitation [14]. A novel approach to the recognition of autonomous human activity in domain-generalized contexts, which is derived from their investigation of the challenges associated with body motion camera footage. This methodology accomplishes data training independence by integrating a deep neural network concept that integrates vision transformers and residual networks into its three-stream model. The technique provides enhanced data robustness, but it necessitates minimal target domain examples. However, its requirement for multivariate data processing may present processing challenges [15]. A technique that utilizes limited educational samples to enable the segmentation and recognition of joint activities, supervised by timestamps. This method is capable of utilizing unlabelled data to enhance execution by combining class-activation maps and optimal transport theory. The method has fewer annotation requirements; however, the ongoing scarcity of data continues to cause difficulties [16].

3. METHODOLOGY

This study specifically used to develop a robust and efficient HAR system by leveraging deep learning techniques, advanced pre-processing, and optimization algorithms to address existing challenges in the field. The primary goal is to refine the accuracy, scalability, and computational effectiveness of action recognitionmodels, particularly when deployed in real-world conditions with diverse datasets and multi-modal inputs.

The input data is collected exclusively from HMDB51 dataset videos, and key frames are extracted from the video clips. These reference frame work are then render as image, which serve as the primary input for the model. The pre-processing step employs Gradient-based Joint Histogram Equalization, which enhances image contrast and normalizes the data, improving visibility of critical features while reducing noise that could interfere with the recognition process. This ensures that subsequent stages of the model can effectively extract relevant patterns from the input images. Next, VGG19, a well-established deep CNN, is used for feature extraction. VGG19 captures hierarchical features from the images, enabling the system to recognize complex patterns and nuances in human motion. These extracted features are then passed to EfficientNetB0, which serves as the primary classifier for the HAR system. To optimize the performance of EfficientNetB0, the whale optimization algorithm (WOA) is employed. WOA fine-tunes critical hyperparameters of EfficientNetB0. Finally, synchronization techniques are incorporated to align features across different modalities (in this case, video frames and images), improving multi-modal integration and overall recognition performance. As shown in Figure 1 represents the block diagram of HAR system model.

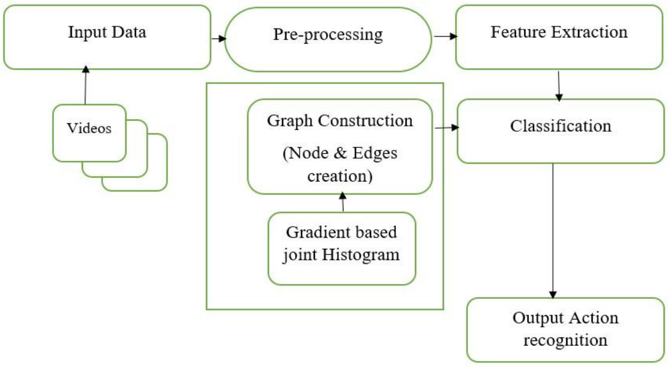


Figure 1.Architecture diagram of HAR

3.1 Input data and key frame selection

Video sequences serve as the key component of data for the HAR system. Multiple video sources are selected since they provide detailed temporal and spatial information that help to identify human activities precisely. Courtesy of video sequence extraction techniques key frames are retrieved. Video selection for key frames follows an information-centric approach as it chooses moments which contain the most relevant content from the overall video data to reduce processing time.

A key frame selection method properly removes repetitive video segments allowing for more efficient processingas per Eq. (1).

Let the video sequence be represented as,

$$V = \{F_1, F_2, ...F_N\} \tag{1}$$

Where, F_i denotes the individual frames and N denotes the total no. frame count

The key frame extraction process can be expressed as per Eq. (2),

$$K = \{F_{k_1}, F_{k_2}, ...F_{k_m}\} \tag{2}$$

Where K represents the set of selected frames, m denotes the no. of k frames and k_i indicates the indices of the extracted frame such that $1 \le k_1 < k_2 < ...K_m \le N$. These key frames are converted into images,as per Eq. (3)

$$I = \{I_1, I_2, ...I_m\} \tag{3}$$

Where, I_j corresponds to the image representation of the key frame F_{kj}. These images I form input to the subsequent pre-processing and feature extraction stage.

3.2 Pre-processing using Gradient-based Joint Histogram Equalization (GBJHE)

A sophisticated **Gradient-based Joint Histogram Equalization**pre-processing algorithm is used to enhance the quality and readiness of input images. The implementation of GBJHE serves two purposes by enhancing image contrast alongside normalization of intensity values so the system can effectively interpret patterns and motion characteristics despite noise disturbances.

Each fully linked layer was followed by L2 regularization to minimize overfitting during the fine-tuned model's implementation.The mathematical representation of feature extraction using VGG-19 layers are discussed in the below section.

a) Convolutional layer: The input images are passed through the convolutional layers of VGG-19. Each convolutional layer applies, extracts spatial features using 3×3 kernels. The convolutional is expressed,as per Eq. (4)

$$F_l(x, y) = \sum_{i=1}^{k} \sum_{j=1}^{k} I(x+i, y+j) \cdot K_l(i, j) + b_l \tag{4}$$

Where, $F_l(x, y)$ denotes the feature map at position (x, y) in layer l, $I(x+i, y+j)$ denotes the input pixel values in the respective field, $K_l(i, j)$ denotes weights of the convolutional kernel and bias term denoted b_l.

b) Activation layer: This layer adds non-linearity to the feature maps. The mathematical equations of the activation layer as per Eq. (5)

$$A_l(x, y) = \mathrm{Re}\,LU(F_i(x, y)) = \max(0, F_l(x, y)) \tag{5}$$

c) Pooling layer:Following the groups of convolutional layers, max-pooling layers further decrease the dimensions, but keep the main elements. It downsamples the feature maps to reduce dimensionality while retaining important features. The equation of the pooling layer as per Eq. (6)

$$P_l(x, y) = \max_{i, j \in k} A_l(x+i, y+j) \tag{6}$$

Where, the pooling window size is denoted as k.

3.3 Feature extraction using VGG-19

The network employs VGG19 as its deep CNN model to extract features from pre-processed images because of its well-known performance in image recognition operations. The process of extracting features stands as a fundamental essential step for developing accurate modelling classifications. Pre-processing with GBJHE enhances image contrast alongside edge retention to generate a complex and detailed set of features that supports upcoming tasks including detection and segmentation. The VGG-19 model extracts complex physical and temporal connection information from priority images allowing it to recognize detailed human motion sequences. The image features extracted from VGG-19 originate from the model's pre-training experience on ImageNet giving a solid foundation for different image processing functions and applications even with small datasets.

3.4 Classification

The extracted features are subsequently fed into EfficientNetB0, which is the main classifier of the HAR system. To fine-tune the performance of EfficientNetB0, the WOA is utilized. WOA optimizes key hyperparameters of EfficientNetB0, including learning rate, dropout, and layer-specific settings, to achieve an optimal trade-off between computational efficiency and recognition accuracy. This optimization improves the model's training efficiency, generalization, and overall performance.

Compound scaling is a method utilized by CNN architectures such as Efficient Net to enhance accuracy.

4. RESULTS AND DISCUSSION

This section primarily addresses the implementation with the details of evaluation. Subsequently, the comparators' performance is compared with that of the proposed model. Later, the efficiency of the model is verified with more testing.

Dataset description: In this paper, HMDB51 dataset is employed which is a large collection of realistic video footage from a variety of sources. The dataset contains 6,766 video clips of 51 action classes (e.g., "laugh," "jump," etc), with a minimum of 101 clips per class. Three distinct training/testing divides are used in the original assessment approach. Each action class has 30 test clips and 70 training clips in each split. The final performance is calculated as the average efficiency over these three splits. Only 5734 video clips were selected from this dataset for our study.

A. **Experimental setup:**In this study, the classifier was trained, tested, and validated using the HMDB51. Ninety percent of the data to train and ten percent of data for testing the model. For the purpose of the study, a system with a 2.4 GHz Intel(R) processor, 16 GB of RAM, and an Nvidia P100 GPU was utilized.

B. **Comparative study with various Baseline models:** A comparison with the most advanced methods in the field can be made to accurately evaluate the methodology's effectiveness. The CNN, DNN, GNN and proposed model were all examined as shown in Table 1.

Table 1. Performance analysis with various methods

TP 90				
Model	**Accuracy**	**Precision**	**recall**	**F1-score**
GNN	79.89	75.67	78.34	73.45
LSTM	76.56	77.78	73.12	75.34
CNN	79.78	72.45	75.89	70.78
DNN	72.56	74.34	77.12	74.56
KNN	75.56	76.23	72.45	76.45
Proposed model with efficientNetB0 based WOA	90.06	89.11	88.82	88.81

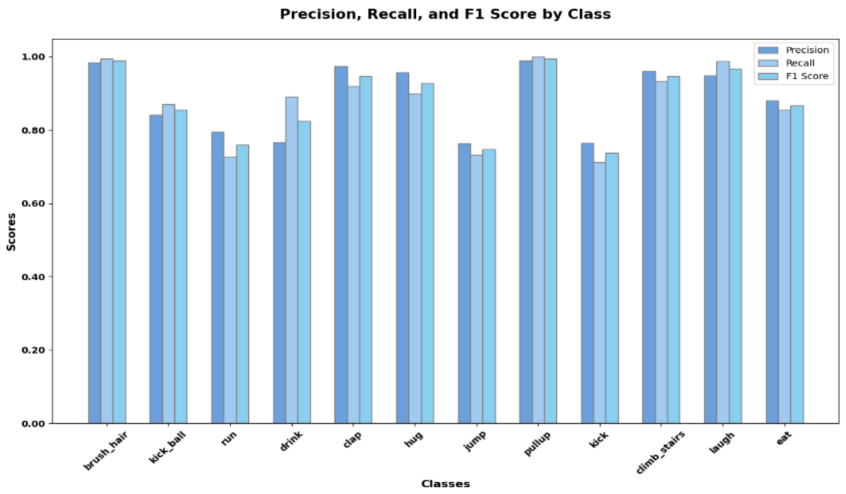


Figure 2.Precision, Recall, F1 score graph representation HAR model

The above Figure 2 describes the precision, it measures how many positive instances were predicted correctly. Recall, however, examines how many instances were predicted correctly as actual instances. The F1-Score is the harmonic mean between precision and recall, and it serves to balance out false negatives and false positives. In the discussion, there is great precision, recall, and F1-scores across most classes, indicating great classification performance. However, some actions, like running and kicking, have slightly lower scores, which could be due to similarities between classes or motion blur in the videos. Overall, the model boasts an accuracy of 90.06%, though there are some variations in performance across different classes.

C. **Evaluation metrics:**The general effectiveness of classification was evaluated using a range of measuring metrics. Several measuring criteria were evaluated to examine the encompassing recall, accuracy, and precision, as well as the F1 measure. Evaluation of the effectiveness of the system was done using the below parameters

- **Accuracy:**The model's general prediction accuracy is measured by this score. The percentage of samples that are correctly classified is calculated based on the total number of samples in the data set.

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

- **Precision:** The ratio of true positives to the true positive rate compares all true positives with false positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- ***Recall:***The measure that measures its ability to identify each positive attitude in the data. The ratio represents the ratio of true positives to the total number of true positives and true negatives.

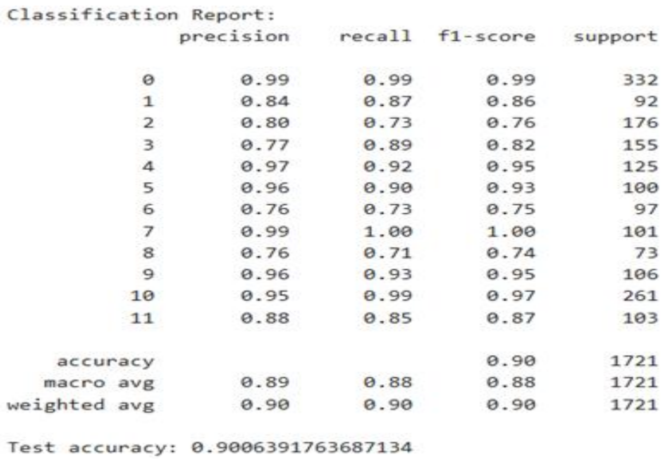
$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score:** It is a unique measure that combines precision and recall, thus both features are captured. It is particularly useful when dealing with an unbalanced dataset where one class is much more frequent than the other.

$$\text{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The Figure 2. illustrates the comparison measures of different performance measures. The outcome was such that it revealed that this method had a productive result as compared to other baseline models.

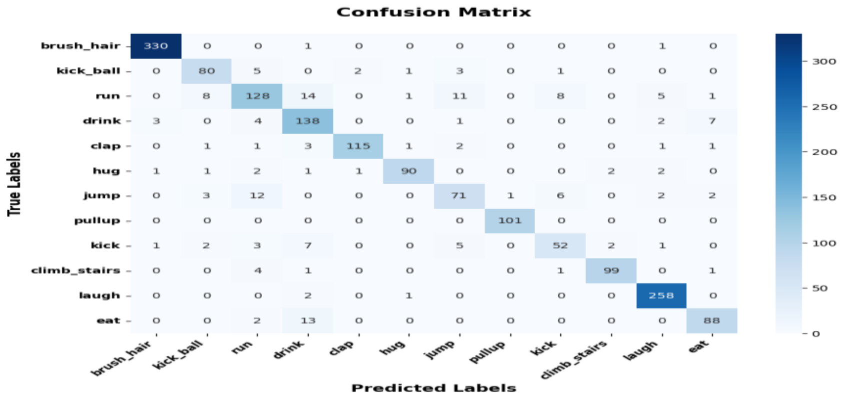
The following Figure 3. is the Classification report for different evaluation measures comprising precisions, recall, F1 score which reveals that this method had a successful result and achieved 90.06% accuracy.



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

           0       0.99      0.99      0.99       332
           1       0.84      0.87      0.86        92
           2       0.80      0.73      0.76       176
           3       0.77      0.89      0.82       155
           4       0.97      0.92      0.95       125
           5       0.96      0.90      0.93       100
           6       0.76      0.73      0.75        97
           7       0.99      1.00      1.00       101
           8       0.76      0.71      0.74        73
           9       0.96      0.93      0.95       106
          10       0.95      0.99      0.97       261
          11       0.88      0.85      0.87       103

    accuracy                           0.90      1721
   macro avg       0.89      0.88      0.88      1721
weighted avg       0.90      0.90      0.90      1721

Test accuracy: 0.9006391763687134
```

Figure 3.Classification report for HAR.

A. Confusion matrix Heatmap: It demonstrates the confusion matrix, which accurately categorizes and identifies each class label, as illustrated in Figure 4. The diagonal cells (which represent true positives) become darker as the number of correct predictions increases



	brush_hair	kick_ball	run	drink	clap	hug	jump	pullup	kick	climb_stairs	laugh	eat
brush_hair	330	0	0	1	0	0	0	0	0	0	1	0
kick_ball	0	80	5	0	2	1	3	0	1	0	0	0
run	0	8	128	14	0	1	11	0	8	0	5	1
drink	3	0	4	138	0	0	1	0	0	0	2	7
clap	0	1	1	3	115	1	2	0	0	0	1	1
hug	1	1	2	1	1	90	0	0	0	2	2	0
jump	0	3	12	0	0	0	71	1	6	0	2	2
pullup	0	0	0	0	0	0	0	101	0	0	0	0
kick	1	2	3	7	0	0	5	0	52	2	1	0
climb_stairs	0	0	4	1	0	0	0	0	1	99	0	1
laugh	0	0	0	2	0	1	0	0	0	0	258	0
eat	0	0	2	13	0	0	0	0	0	0	0	88

Figure 4. illustrates the confusion matrix heatmap

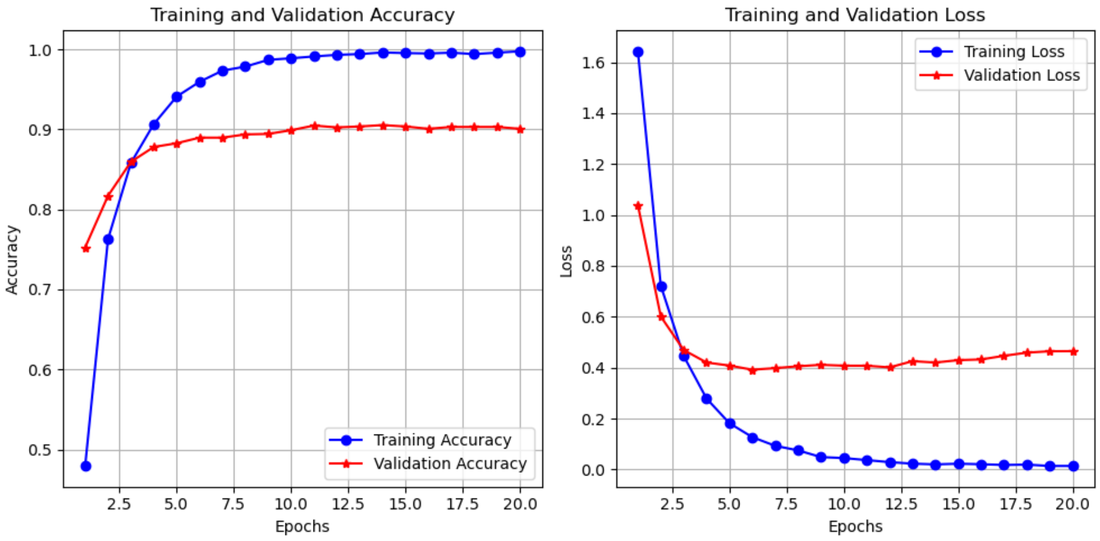


Figure 5.Training and testing dataset Accuracy and Loss of HAR model.

The training accuracy increases consistently, nearly reaching 100%, as illustrated in Figure 5, whereas the validation accuracy levels at approximately 90%. This implies that the model is performing satisfactorily in its generalization. The loss curve indicates that the model is well-trained and does not suffer from significant overfitting, as the training loss has experienced a substantial decrease and the validation loss has stabilized.

5.CONCLUSION

The major emphasis of this research is the design and development of the effective HAR model through the integration of state-of-the-art tuning methods. The new pipeline uses Gradient-based Joint Histogram Equalization to improve image quality and normalize input data, enhancing essential feature visibility. VGG19 is used for efficient feature extraction, and EfficientNetB0, a compact yet effective classifier, optimizes recognition accuracy versus computational efficiency.The WOA also fine-tunes EfficientNetB0's hyperparameters, increasing training efficiency and total systemperformance. The system was thoroughly tested on the HMDB51 dataset and exhibited excellent performanceon metrics including accuracy, precision, recall, and F1 score. Multi-epoch analysis and k-fold validation validatedthe model's robustness and scalability. Comparative analyses with state-of-the-art methods, including GNN,CNN, DNN, methods, establish the superiority of the proposed method in terms of recognitionachieved an accuracy of 90.06% and computational efficiency. The current study introduces a thorough HARframework, appropriate for real-world use cases with various datasets and constrained environments.The future work can involve the use of more data modalities, enhance real-time processing, and investigate applicationsin healthcare and sports analysis fields. This work substantially contributes to the development of HAR systems,establishing a foundation for efficient and scalable solutions.

REFERENCES

[1] Kulsoom, F., Narejo, S., Mehmood, Z., Chaudhry, H. N., Butt, A., & Bashir, A. K. (2022). A review of machine learning-based human activity recognition for diverse applications. Neural Computing and Applications, 34(21), 18289-18324.

[2] Brenner, M., Reyes, N. H., Susnjak, T., & Barczak, A. L. (2023). RGB-D and thermal sensor fusion: A systematic literature review. IEEE Access, 11, 82410-82442.

[3] Yadav, S. K., Tiwari, K., Pandey, H. M., & Akbar, S. A. (2021). A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. Knowledge-Based Systems, 223, 106970.

[4] Zhang, H. B., Zhang, Y. X., Zhong, B., Lei, Q., Yang, L., Du, J. X., & Chen, D. S. (2019). A comprehensive survey of vision-based human action recognition methods. Sensors, 19(5), 1005.

[5] Feng, L., Zhao, Y., Zhao, W., & Tang, J. (2022). A comparative review of graph convolutional networks for human skeleton-based action recognition. Artificial Intelligence Review, 1-31.

[6] Bsoul, A. A. R. K. (2024). Human Activity Recognition Using Graph Structures and Deep Neural Networks. Computers, 14(1), 9.

[7] Xia, S., Chu, L., Pei, L., Yang, J., Yu, W., & Qiu, R. C. (2024). Timestamp-supervised wearable-based activity segmentation and recognition with contrastive learning and order-preserving optimal transport. IEEE Transactions on Mobile Computing.

[8] Yang, H., Ren, Z., Yuan, H., Xu, Z., & Zhou, J. (2023). Contrastive self-supervised representation learning without negative samples for multimodal human action recognition. Frontiers in Neuroscience, 17, 1225312.

[9] Huang, J., Lin, S., Wang, N., Dai, G., Xie, Y., & Zhou, J. (2019). TSE-CNN: A two-stage end-to-end CNN for human activity recognition. IEEE journal of biomedical and health informatics, 24(1), 292-299.bb

[10]Mohottala, S., Samarasinghe, P., Kasthurirathna, D., &Abhayaratne, C. (2022, August). Graph neural network based child activity recognition. In 2022 IEEE International Conference on Industrial Technology (ICIT) (pp. 1-8). IEEE.

[11] Shakerian, A., Douet, V., ShoarayeNejati, A., & Landry Jr, R. (2023). Real-time sensor-embedded neural network for human activity recognition. Sensors, 23(19), 8127.

[12] Zhang, J., Tsai, P. H., & Tsai, M. H. (2024). Semantic2Graph: graph-based multi-modal feature fusion for action segmentation in videos. Applied Intelligence, 54(2), 2084-2099.

[13] Liu, J., Chen, C., & Liu, M. (2024, October). Multi-modality co-learning for efficient skeleton-based action recognition. In Proceedings of the 32nd ACM International Conference on Multimedia (pp. 4909-4918).

[14] Gao, Y., Lu, J., Li, S., Li, Y., & Du, S. (2024). Hypergraph-based multi-view action recognition using event cameras. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[15] Papadakis, A., & Spyrou, E. (2024). A multi-modal egocentric activity recognition approach towards video domain generalization. Sensors, 24(8), 2491.

[16] Xia, S., Chu, L., Pei, L., Yang, J., Yu, W., & Qiu, R. C. (2024). Timestamp-supervised wearable-based activity segmentation and recognition with contrastive learning and order-preserving optimal transport. IEEE Transactions on Mobile Computing.