ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

Streamlined Deep Learning Model For Video-Based Human Action Recognition

[1]Mr. K. Lakshmi Narayana, [2]Mediboyina Sekhar,

[1]Assistant Professor, Department of Master of Computer Applications,
Rajamahendri Institute of Engineering & Technology.
Bhoopalapatnam, Near Pidimgoyyi,Rajahmundry,E.G.Dist.A.P. 533107.

[2]Student, Department of Master of Computer Applications,
Rajamahendri Institute of Engineering & Technology.
Bhoopalapatnam, Near Pidimgoyyi,Rajahmundry,E.G.Dist.A.P. 533107.

Abstract—

Many researchers in the field of computer vision have lately focused on Human Action Recognition (HAR) from a visual stream. Because it has so many potential uses some examples of which are health monitoring, home automation, and teleimmersion. Nevertheless, it continues to encounter challenges such as human variation, occlusion, lighting variations, and complex backdrops. The features gathering technique and proper execution of learning data are crucial to the assessment criteria. Neural networks are only one of many remarkable products of Deep Learning (DL).

However, in order for a good classifier to assign a label, a strong features vector is necessary. Data sets cannot be complete without features. The computational cost and performance of the method may be impacted by feature extraction. Using the SoftMax layer, we extracted features from the picture sequence using the pre-trained deep learning models VGG19, Dense Net, and Efficient Net, and we categorized each action. f1-score, AUC, precision, and recall are used to assess performance on the UCF50 action dataset, which consists of 50 parts. Results from the tests showed that VGG19-90.11, DenseNet-92.57, and EfiicientNet-94.25 were the most accurate models.

Keywords—Transfer Learning, CNN, VGG19, UCF50

INTRODUCTION

Whether it's the naked eye or a sophisticated sensor, in HAR an action is anything that can be observed. As a matter of fact, walking demands continuous

paid attention to an individual situated inside the optical field. It is possible to classify actions into four groups based on the parts of the body that are required to carry them out. [1]. Expression on the face is the foundation of gesture. Requires neither physical nor verbal means of expression.

Human activity included walking, playing, and punching.

Interaction includes not just human-object interactions but also human-to-human interactions such as embracing and handshakes. When more than two actions are taking place, such as a mix of gestures and interaction, it is referred to as group activity. When an action is performed, it usually involves two or more actors. For computer vision researchers, HAR has been an indispensable tool in the last 20 years. A person or people's actions may be detected and identified using HAR, which is built on a database of observations. This may be done for a variety of individuals. There was now an urgent need to advance human-computer interaction because of this. The vast variety of potential applications for this area of study attracts scholars from all around the world. It has several notable uses, including environmental modeling, health monitoring, automation, surveillance video, and image categorization and retrieval. [1]. There is an inherent hierarchical structure to human actions, and this structure indicates the numerous levels. These levels may be classified into three main groups. To begin, at the most fundamental level, there is an atomic element; these action primitives stand in for the more complex human acts. After the action primitive level comes the actions/activities level. Complex interactions reflect the highest degree of human activity classification.

Due to the breadth of each of these groups, they merit independent research. The main reason for this is because human actions in real life are often unpredictable and ambiguous. HAR encounters a number of challenges. Some examples include gender bias, interactions involving more than one topic, and differences in inter-class engagement. A four-step method is involved in human activity recognition from videos. We begin by extracting

features from provided picture sequences. Various handcrafted methods can be employed in the feature extraction process, such as SIFT (scale-invariant feature transform), SURF (speed up robust feature), shape-based, pose-based, optical flow, and many more [1]. With this strategy, the model learns all the features from the picture sequences automatically, making feature extraction a breeze. Pose and gesture patterns may be extracted from video sequences and frames that show people engaging in various tasks. Size variations, bad lighting, wrong perspectives, and background clutter are some of the hurdles that make this a tough assignment. Learning and recognition of actions based on extracted characteristics is the next level. An integral aspect of action learning and recognition is learning new models that are instructed by extracted features. Determining which features are pertinent to which action classes and evaluating those features using classifiers are other crucial steps in the process. Among the many prominent ways to address the HAR problem are the ML methodology and the DL method. In the first, more traditional version of AI, the user is still involved in the process of designing, dictating, and honing the extracted attributes and action characterization. We expect the deep neural network (DNN) to perform better using the second approach. The second method relies on the expectation that the DNN can mimic human intelligence and solve all of the qualities automatically [1][2].

Figure 1 depict ML and DL base classification for HAR.

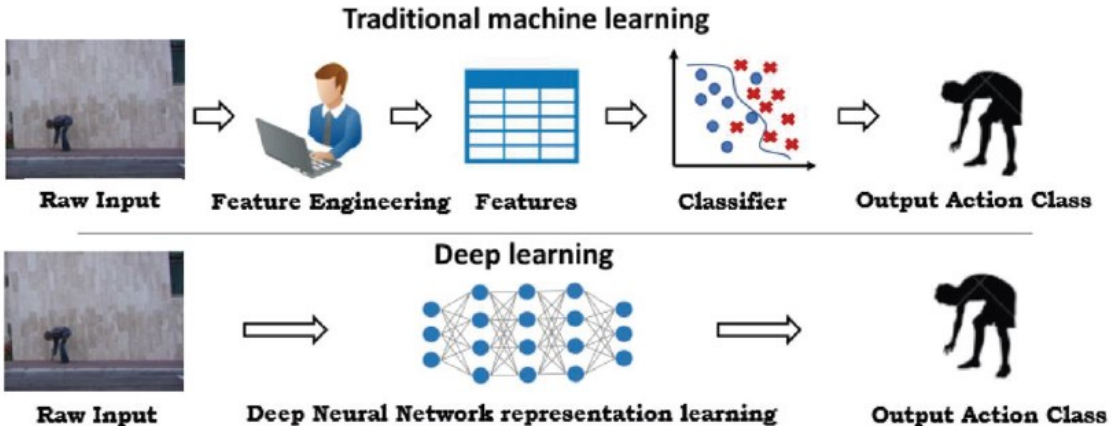


Fig. 1. A graphical representation of the conventional ML methods and the cutting-edge DL methods employed for HAR [2].

Numerous machine learning (ML) foundational approaches, including support vector machines (SVM), Markov models (MM), and random forests (RF), have been used for many years in an effort to

resolve the HAR problems that are related to it, such as the backdrop clutter, noise, and class similarity issues.
Even with very little data and very severe constraints, seasoned ML algorithms have achieved outstanding results. The preprocessing phase with the created feature makes machine learning algorithms time-consuming and requires extra attention; they should perform better.
If the amount of data is enormous. DL has made significant progress in recent years. This is because deep learning research has successfully completed tasks in many different areas of study, including object identification in frames, action recognition, frame categorization, and natural language processing, among many others.

With its structure proven successful for both unsupervised and reinforced learning, DL drastically cuts down on the work needed to identify the right features compared to typical ML algorithms. This is all because to the unmanned features pulled out via several hidden layers. This has led to an increase in the number of HAR frameworks that rely on deep learning.
A brief overview of the research article is as follows:
First, we provide a brief introduction to human action recognition, and then we look at how machine learning and deep learning approaches this problem. After that, in Section 2, we will discuss the kinds of human action recognition methods that have been suggested and how accurate they are. In Section 3, we talk about the methods and the outcomes on the dataset. The current state of computer vision research and its anticipated future directions are reviewed in Section 4.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

RELATED WORK

A lot of work has been done in the topic of human action recognition. Given its adaptability, there is a great deal of room to enhance the forecasting of human behavior. Quite a lot

Over the last ten years, a number of feature-based approaches for visual human activity identification have been developed, including both manually created and machine taught methods. In the past, methods for identifying human activities depended on subjective features that were too concerned with minute atomic processes that don't seem to have any practical utility [3]. While these approaches do produce very accurate models, their main drawback is the amount of data preparation required and the difficulty in generalizing them to real-world scenarios. Many spatiotemporal methods for video activity analysis have been developed since convolutional neural networks (CNNs) became successful in text and visual classification; these algorithms are able to automatically train and classify from raw RGB video [4]. In order to extract spatial and temporal video data for action recognition, Shuiwang Ji et al.[5] presented a 3D convolution approach. Consequently, the proposed architecture uses the video sequence to generate several data channels, each of which is subjected to subsampling and convolution. In order to find one's way about within buildings, Gu et al. presented a DL-based approach to locomotion detection. Instead of manually building the required features, their technique used stacked denoising auto-encoders to learn data properties automatically [6]. As compared to another classifier, the suggested study approach claims to have achieved higher precision. A novel method for identifying an action from RGB (Color model) video was developed by Aubry et al. [7]. First, you'll need to remove the human skeleton from the movie by removing its movements. We used Open Pose [8], a Deep Neural Network (DNN)-based approach for employee identification, to extract a 2-dimensional skeleton with 18 identified joints from each individual's body. An image classifier is used to transform motion patterns into RGB images in the second case.

Motion data is stored in the R, G, and B channels. The result is an RGB image suitable for use in an action scene. One possible use for neural networks in the future is action recognition, building on its current usage for picture classification. A dual-stream model was proposed by Dai et al. [9] that locates action in visual frames using an attention-based long short-term memory (LSTM) structure. The issue of neglecting visual attention was supposedly resolved, they said. The architecture achieved a 96.9% accuracy rate with the UCF11 dataset, a 98.6% accuracy rate with the UCF Sports dataset, and a 76.3% accuracy rate with the j-HMDB dataset. A skeleton-based approach to action recognition using a hierarchical RNN model was developed by Du et al. [10]. In addition, five distinct deep RNN designs were tested against their proposed methods. They employed the HDM05 dataset, the Berkeley MHAD dataset, and the MSR Action-3D dataset throughout their examination. The Correlational Convolutional LSTM was created by Majd and Safabakhsh [11] by adding spatial and motion information to an existing LSTM module and creating temporal links. Their results showed a 92.3% correctness rate and a 61.0% accuracy rate when tested on the popular UCF101 and HMDB51 benchmark datasets, respectively. Qi et al.[12] proposed stag-Net, an alternative method for constructing a semantic RNN, with the aim of identifying both group and individual activities. Using a structural RNN, they expanded their semantic network model to include time as a fourth dimension.

Team efforts accounted for 90.5% of the Volleyball dataset, whereas individual efforts accounted for 8.5%.

Huang et al.[13] used a 3D convolutional neural network (ConvNet) to extract features based on posture by merging information on the subject's 3-dimensional stance, 2-dimensional appearance, and motion. Because we anticipate that the color joint features obtained by 3-D CNNs would be computationally demanding, we apply convolution to each of the 15 channels of the heatmap in order to lower the noise. In both Inception and Batch Normalization, Wang et al. used the (BN-inception) network design [14]. Similar to twostream networks, the previously described method uses RGB variation frames to mimic visible change and optical flow fields in addition to RGB and optical flow frames to prevent background motion. In [15], the author used a graph pooling network and a GCN with a channel attention method for the joints.

Last but not least, the SGP architecture enhanced convolution by including a human skeletal network. Specific information about the human body is retrieved by use of kernel receptive areas. While reducing calculation costs, the proposed SGP method has the ability to greatly enhance GCNs' ability to collect depending on motion characteristics.

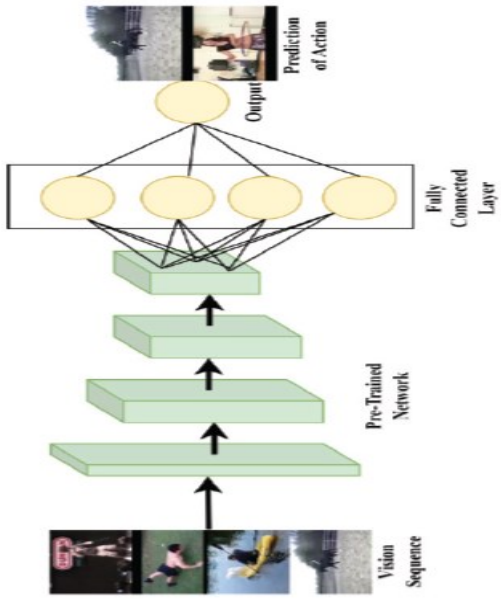
Stream designs used in the study paper [16] include context streams and fovea streams. Although the fovea channel gets the central area at full resolution, the context channel only receives frames at half their original resolution. The research uses each video as a set of short, fixed-length clips to train a model that can distinguish

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

between three distinct pattern classes: Early Fusion, Late Fusion, and Slow Fusion. Through a variety of time-space combinations, CNN is able to generate single-frame animations.
For the purpose of human activity recognition, Singh et al.[17] presented bidirectional-LSTM, a highly connected ConvNet that uses RGB frames as its top layer. All it takes to train the bottom layer of a ConvNet model is one DMI. While the top layers of the pre-trained ConvNet are fine-tuned to extract temporal information from video streams, the ConvNet-Bi-LSTM model is trained from scratch for RGB frames to enhance the features of the pre-trained CNN. The decision layer uses a late fusion strategy that follows the SoftMax layer to merge features in order to acquire a better accuracy result. In order to test how well the proposed model works, four RGB-D (depth) datasets were used, one for each kind of activity (including those involving several people).



METHODOLOGY

The DL model for HAR demonstrates the significant outcome for the categorization of every task. We went over the inner workings of a few deep learning models and how they categorize certain actions.

right on. It takes a lot of processing power to train a deep learning model from the ground up. Learning models are taught differently from transfer learning models. The massive dataset ImageNet is used to train them[18].
More than 1 million photos that are appropriate for training transfer learning models are available in ImageNet. Several different transfer learning models were compared against state-of-the-art methodologies in this study's classification of each activity. Several transfer learning models for action recognition were examined in this study. In Figure 2, we can see the Human Action Recognition model combined with the deep learning model that was trained beforehand.

Methodologies based on transfer learning (TL) are evaluated using Dense Nets [19]. The novel methods that Dense Net neural networks use to deal with disappearing or Increasing gradients and their unique architecture enable feature reuse by letting one layer learn from the feature maps of previous layers. The very deep architecture of VGG[20] is achieved by using small (33) filters, and this is achieved via a transfer learning-based HAR method. Gradient explosions are common in VGG models because of their intricacy. In order to tackle this issue, we used VGG models that included batch normalization layers to manage the gradients. A framework's performance may also be assessed using the Efficient Net[21] technique.
Section A. Dense Net
Dense Net is the name given to a kind of Convolution neural network that employs a feed-forward technique to link each successive network layer. A high-filter-size Conv2D layer is the first stop for the data's path, followed by a dense block that forms dense connections with every subsequent layer. Every Dense Net layer takes data from every layer below it and sends its feature maps to every layer above it.
The B. VGG
We further included VGG [20], a CNN architecture, into the TL-based method for action recognition. Images with a certain ratio, i.e. 512 × 512 pixels (224, 224, 3), are fed into VGG for training. The pictures via a series of convolutional layers equipped with 3-by-3-pixel filters. Spatial pooling is performed via five max-pooling layers following particular conv2D layers. After a series of convolutional layers, thick layers with complete connectivity and a SoftMax prediction layer are added.

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

The VGG19 architecture is shown in Figure 3, where the variables "conv," "pool," and "FC" are associated with the various layers.

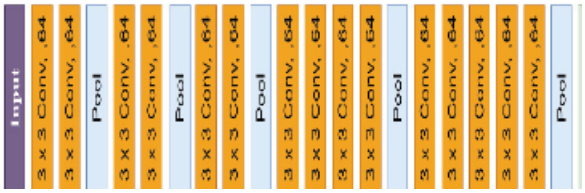


Fig. 3. VGG19 Architecture

EfficientNet

An architecture and scaling technique for convolutional neural networks, Efficient Net[21] uses a compound to uniformly scale all parameters of depth, breadth, and resolution.

coefficient. Efficient Net scaling uses a set of specified scaling factors to evenly modify the breadth, depth, and resolution of the network, as opposed to the existing method that randomly scales these parameters. Optimal Network [21] Unique among convolutional neural networks (CNNs), it quickly and efficiently estimates parameters. In order to more methodically scale up CNN models, Efficient Net [21] uses a simple and difficult scaling methodology to evenly scale network features including depth, breadth, and resolution. As a spatial feature extraction network, Efficient Net [21] was also used in classification tasks. With the names EfficientNet-B0 through EfcientNet-B7, the Efficient Net family included seven convolutional neural network (CNN) models. Even though EfcientNet-B0 only had a fraction of the parameters and FLOPs (floating-point operations per second) of Resnet-50[22], it nevertheless managed to beat it when it came to feature extraction efficiency, using the same input size.

D. Dataset The UCF50[23] dataset was used to assess the performance of the model. Reddy et al. (2012) first suggested this dataset. Videos are gathered via online channels such as YouTube. No footage has ever been shot in a studio; instead, it all takes place in an authentic setting. Compared to the UCF11 dataset, this one has been revised and improved. Basketball, shooting, riding, tabla playing, violin playing, and 50 more activity lessons are all part of it. There are 6618 videos covering a wide range of topics, from basic sports to commonplace life. There are a minimum of four films assigned to each activity, and each class is further divided into 25 similar groups. Characters, settings, and points of view are commonalities across films that fall under the same genre. Figure 4 shows the UCF 50 dataset's action snippets.

DISCUSSION AND RESULTS

We used three pre-trained deep learning models—Dense Net, VGG19, and Efficient Net—to categorize each action. Our team took advantage of the situation by using pre-trained deep learning to

data collected from large databases like ImageNet. The idea behind the transfer learning method is to feed data from an existing trained model into a new neural

network to train a new domain. Evaluation of the UCF50

Fig. 4. UCF50 Action Dataset Frames.

activity dataset, which includes several picture categories. Our goal in using this method was to examine the accuracy of several deep learning models on the aforementioned dataset.

contemporary approaches. An first step was to feed each set of action movies' extracted frames into a deeplearning model that had already been trained. Confusion matrices for 50 activities recognition from the UCF 50 dataset employing the VGG19 model, Dense Net 161, and EfficientNet b7 are shown in Figures 5-7.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

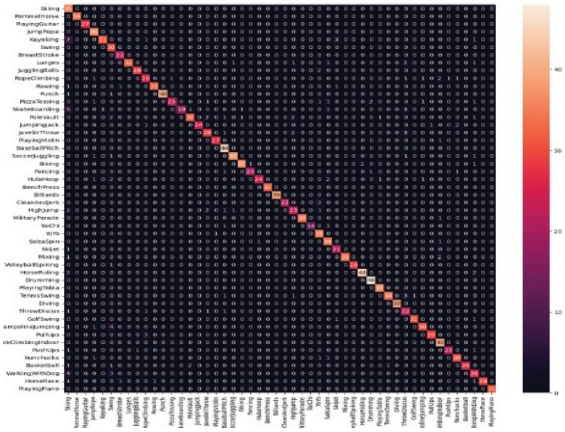


Fig. 5. VGG19 model confusion matrix for action recognition.

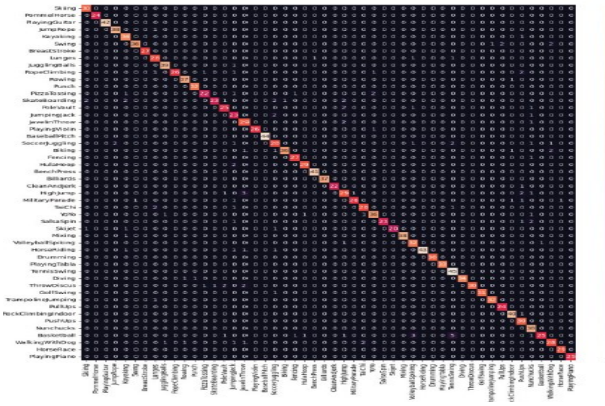


Fig. 6. Utilizing Dense Net 161 model, a confusion matrix for action recognition.

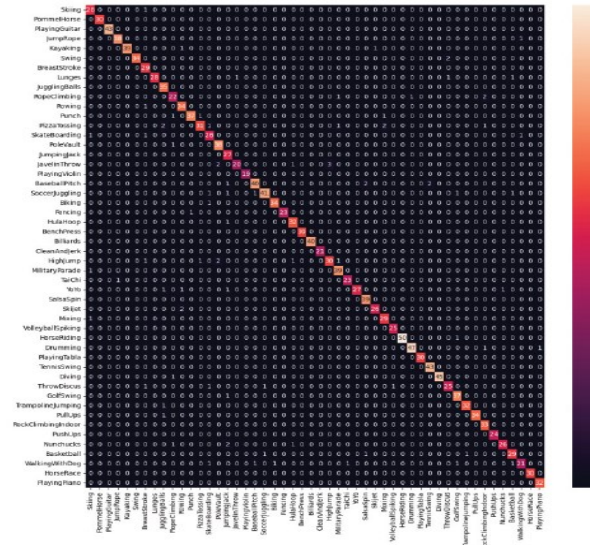


Fig. 7. Confusion matrix for action prediction from Efficient Net b7 model.

Findings from the activity dataset's classification A Confusion matrix displays UCF 50. We can confidently and properly categorize the majority of the actions. Regarding the UCF50 action model assessment metrics using TL approaches are compared in Table 1 of the dataset. The recovered frames were partitioned during the implementation phase using the methods employed in training, validation, and testing. A visual representation of this may be seen in Figure 8. Compared to other cutting-edge approaches shown in Table2:

TABLE I. COMPARISON OF VARIOUS LIGHT WEIGHT DL METHOD

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
VGG19	90.11	91.92	90.34	90.53
Dense Net 161	92.57	93.06	92.45	92.43
Efficient Net b7	94.25	94.92	94.79	94.71

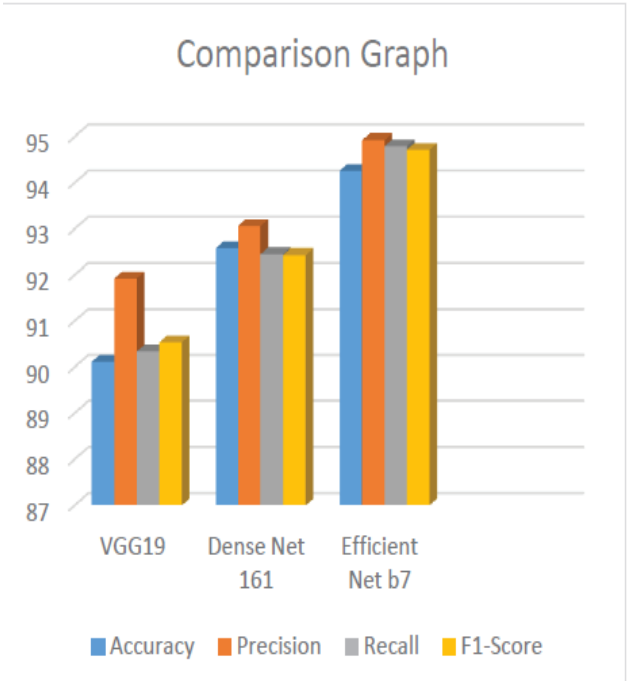


Fig. 8. Comparison graph for evaluation metrices.

TABLE II. COMPARISON OF LIGHTWEIGHT DL METHOD WITH

EXISTING APPROACH

Researcher	Dataset	Accuracy (%)
L. Zhang et al[24]	UCF50	88.0
H. Wang et al[25]	UCF50	89.1
Q. Meng et. al[26]	UCF50	89.3
Ahmad Jalal et. al[27]	UCF50	90.48
VGG19_bn	UCF50	90.11
Dense Net 161	UCF50	92.57
Efficient Net_b7	UCF50	94.25

We compared the efficiency of our strategy to that of many other approaches that avoided transfer learning on the UCF 50 dataset. The results of the study

tested if using a different dataset with transfer learning improves the recognition score. Its classification performance is improved by 1-4 percent when pretrained DL is used.

CONCLUSION

The UCF 50 action dataset is used to develop deep learning algorithms that can categorize human actions. The UCF50 action dataset is structured into 25 groups and comprises 50 distinct action types.

every one including a minimum of four videos. The accuracy and efficacy of the model were tested using several evaluation matrices, such as recall, f1 score, and AUC score. Every activity in the dataset is categorized by VGG19, Dense Net 161, and Efficient Net programs. Additionally, this study contrasted cutting-edge approaches that were used with the UCF50 dataset. When compared to state-of-the-art approaches, these pretrained deep learning models outperform them. When compared to other pre-trained deep learning models, Efficient Net's 94% accuracy is superior. Adding more datasets, real-time action monitoring, aberrant action detection, and crowd behavior classification capabilities to this work is possible down the road. This study then modifies the pre-trained deep learning model's architecture, for example by adding an attention layer, so that it may be used with Bi-LSTM.

REFERENCES

[1]. P. Pareek and A. Thakkar, “A survey on video-based Human ActionRecognition: recent updates, datasets, challenges, and applications,”Artif Intell Rev, vol. 54, no. 3, pp. 2259–2322, Mar. 2021, doi:10.1007/s10462-020-09904-8.

[2]. P. K. Singh, S. Kundu, T. Adhikary, R. Sarkar, and D. Bhattacharjee,“Progress of Human Action Recognition Research in the Last TenYears: A Comprehensive Survey,” Archives of ComputationalMethods in Engineering, vol. 29, no. 4, pp. 2309–2349, Jun. 2022,doi: 10.1007/s11831-021-09681-9.

[3]. A. Ladjailia, I. Bouchrika, H. F. Merouani, N. Harrati, and Z.Mahfouf, “Human activity recognition via optical flow: decomposingactivities into basic actions,” Neural Comput Appl, vol. 32, no. 21,pp. 16387–16400, Nov. 2020, doi: 10.1007/s00521-018-3951-x.

[4]. K. Simonyan and A. Zisserman, “Two-Stream ConvolutionalNetworks for Action Recognition in Videos.”

[5]. S. Ji, W. Xu, M. Yang, and K. Yu, “3D Convolutional neuralnetworks for human action recognition,” IEEE Trans Pattern AnalMach Intell, vol. 35, no. 1, pp. 221–231, 2013, doi:

[6]. 10.1109/TPAMI.2012.59.

[7]. F. Gu, K. Khoshelham, and S. Valaee, “Locomotion activityrecognition: A deep learning approach,” in IEEE InternationalSymposium on Personal, Indoor and Mobile Radio Communications,PIMRC, Feb. 2018, vol. 2017-October, pp. 1–5. doi:10.1109/PIMRC.2017.8292444.

[8]. S. Aubry, S. Laraba, J. Tilmanne, and T. Dutoit, “Action recognitionbased on 2D skeletons extracted from RGB videos,” MATEC Web ofConferences, vol. 277, p. 02034, 2019, doi:10.1051/matecconf/201927702034.

[9]. Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, “OpenPose:Realtime Multi-Person 2D Pose Estimation Using Part AffinityFields,” IEEE Trans Pattern Anal Mach Intell, vol. 43, no. 1, pp. 172–186, Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[10]. C. Dai, X. Liu, and J. Lai, “Human action recognition using twostreamattention-based LSTM networks,” Applied Soft

ComputingJournal, vol. 86, Jan. 2020, doi: 10.1016/j.asoc.2019.105820.

[11]. Y. Du, W. Wang, and L. Wang, "Hierarchical Recurrent NeuralNetwork for Skeleton Based Action Recognition."

[12]. M. Majd and R. Safabakhsh, "Correlational Convolutional LSTM forhuman action recognition," Neurocomputing, vol. 396, pp. 224–229,Jul. 2020, doi: 10.1016/j.neucom.2018.10.095.

[13]. M. Qi, Y. Wang, J. Qin, A. Li, J. Luo, and L. van Gool, "StagNet: AnAttentive Semantic RNN for Group Activity and Individual ActionRecognition," IEEE Transactions on Circuits and Systems for VideoTechnology, vol. 30, no. 2, pp. 549–565, Feb. 2020, doi:10.1109/TCSVT.2019.2894161.

[14]. Y. Huang, S.-H. Lai, and S.-H. Tai, "Human Action RecognitionBased on Temporal Pose CNN and Multi-Dimensional Fusion."

[15]. Wang Limin et al., Computer Vision – ECCV 2016, vol. 9912. Cham:Springer International Publishing, 2016. doi: 10.1007/978-3-319-46484-8.

[16]. Y. Chen et al., "Graph convolutional network with structure poolingand joint-wise channel attention for action recognition," PatternRecognit, vol. 103, Jul. 2020, doi: 10.1016/j.patcog.2020.107321.

[17]. A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing andUnderstanding Recurrent Networks," Jun. 2015, [Online]. Available:http://arxiv.org/abs/1506.02078

[18]. T. Singh and D. K. Vishwakarma, "A deeply coupled ConvNet forhuman activity recognition using dynamic and RGB images," NeuralComput Appl, vol. 33, no. 1, pp. 469–485, Jan. 2021, doi:10.1007/s00521-020-05018-y.

[19]. O. Russakovsky et al., "ImageNet Large Scale Visual RecognitionChallenge," Int J Comput Vis, vol. 115, no. 3, pp. 211–252, Dec.2015, doi: 10.1007/s11263-015-0816-y.

[20]. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger,"Densely connected convolutional networks," in Proceedings - 30thIEEE Conference on Computer Vision and Pattern Recognition,CVPR 2017, Nov. 2017, vol. 2017-January, pp. 2261–2269. doi:10.1109/CVPR.2017.243.