ISSN NO: 9726-001X

Volume 13 Issue 02 2025

JOURNAL OF CURRENT SCIENCE

Human Action Recognition in Videos by using Optimized Deep Learning Architecture

[1] Gosala Sweeti, [2] Mudhunuri Sandhya Rani, [3] Araja Veeranjaneyulu, [4] Anegandhi Lakshmi Sai Hemanth [5] Mr. K.T.V. Subbarao,

[1,2,3,4] Students, Dept. of CSE, DNR College of Engineering & Technology, Balusumudi, Bhimavaram, India.

[5] Assistant Professor, Dept. of CSE, DNR College of Engineering & Technology, Balusumudi, Bhimavaram, India.

Abstract—

In the field of computer vision, Human Action Recognition (HAR) from a visual stream has lately garnered a lot of attention from researchers. Because of all the amazing things it can do, including teleimmersion, home automation, and health monitoring. 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). To be sure, a reliable classifier can't assign a label without a strong features vector. The backbone of any data collection is its features. It is possible that the algorithm's efficiency and computing cost are 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. The UCF50 action dataset was used; it is 50 sections long and uses f1-score, AUC, precision, and recall to measure performance. The models' tested accuracy was 90.11 for VGG19, 92.57 for DenseNet, and 94.25 for EfiicientNet.
Keywords: UCF50, VGG19, CNN, Transfer Learning

INTRODUCTION

Any event that can be observed by either the naked eye or a sensing device is considered an action in HAR. Actually, paying close attention to someone in your line of sight is essential while you're doing anything like walking. It is possible to classify actions into four groups based on the parts of the body that are required to carry them out. [1]. The foundation of gesture is the look on the face. Requires neither physical nor verbal means of expression. Human activity included walking, playing, and punching. Interaction: It includes the interaction of humans with objects and with each other, such as when people embrace or shake hands. When more than two actions are taking place, such as a mix of gestures and interaction, it is referred to as group activity. In order to carry out an action, two or more actors are required. 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. Many notable applications use it, including environmental modeling, health monitoring, automation, and surveillance video classification 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. As a starting point, we have an atomic element; these action primitives stand in for the ever more complex human acts. The actions/activities level is the second level after the action basic level. Complex interactions reflect the highest degree of human activity classification. Because of how vast each of these groups is, research on them has to be separate. The main reason for this is because human actions in real life are often unpredictable and ambiguous. There are a number of challenges that HAR must overcome. Interactions involving many subjects, gender bias, and differences in inter-class activity are all instances of this. There is a four-step technique for human activity recognition in videos. We begin by extracting features from provided picture sequences. An assortment of handmade approaches may be used in the feature extraction process, including but not limited to SIFT (scale-invariant feature transform), SURF (speed up robust feature), shape-based, pose-based, and optical flow [1]. This technique uses deep learning to autonomously extract features from picture

ISSN NO: 9726-001X

Volume 13 Issue 02 2025

JOURNAL OF CURRENT SCIENCE

sequences. The model learns all the features on its own. It entails identifying human actions in moving images and then extracting related postures and gesture patterns. So, it's not an easy process because of things like size variations, bad lighting, wrong perspectives, and background clutter. The next step involves using the collected information to learn and identify actions. Understanding which features are pertinent to which action classes and evaluating those characteristics using classifiers are crucial components of action learning and recognition, as is learning new models that are instructed by extracted features. Some of the most well-known ways to address the HAR problem are the DL method and the Machine Learning (ML) methodology. 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]. Base classification for HAR using ML and DL is shown in Figure 1.

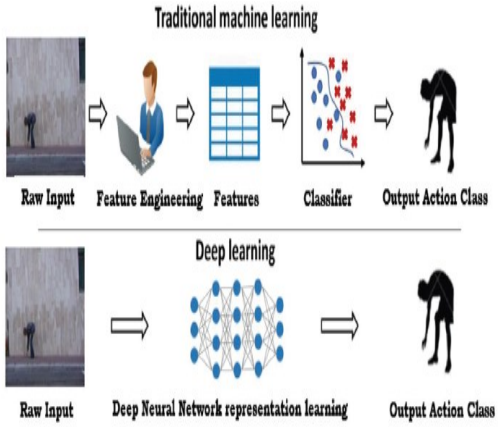

Fig. 1. A graphical representation of the conventional ML methods and the cutting-edge DL methods employed for HAR.[2].

For decades, ML-based algorithms like support vector machines (SVM), random forests (RF), and Bayesian networks (BN) have been used to attempt to overcome the HAR problems that come with it, such as the noise issue, clutter backdrop, and class similarity issue. Skilled ML algorithms have shown themselves capable of producing top-notch results even when faced with severely constrained data inputs. Improving the performance of machine learning algorithms is a top priority because of the time and effort required for the preprocessing stage that involves handmade features. 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. Deep learning (DL) is an efficient framework for both unsupervised and reinforced learning, and it drastically cuts down on the work needed to choose the right features compared to conventional ML algorithms by using a cascade of hidden layers to extract them. 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: We begin with a brief introduction to human action recognition before moving on to a discussion of machine and deep learning approaches to the problem. After that, in Section 2, we will discuss the methods and degrees of accuracy of earlier approaches to human action recognition. In Section 3, we talk about the methods and the outcomes on the dataset. In Section 4, the current state of computer vision research and its anticipated future trajectory are reviewed.

RELATED WORK

The field of human action recognition has accomplished a great deal. There is a lot of room to enhance human action prediction because of its broad range of applications. Human activity identification in pictures has seen a proliferation of feature-based approaches in the last decade, both manually and automatically trained. Earlier methods of human activity recognition depended on manually-entered traits, with an emphasis on insignificant atomic operations that don't seem to have any practical utility [3]. Although these approaches provide very accurate models, their main drawback is the difficulty in generalizing their results and the significant data preparation that is required. 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]. To achieve action recognition in videos, Shuiwang Ji et al.[5] presented a 3D convolution technique for extracting spatial and temporal data. Consequently, the proposed architecture uses the video sequence to generate several data channels, each of which is subjected to subsampling and convolution. For indoor navigation and localization, Gu et al. presented a DL-based approach to detect locomotive movements. 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 research framework claims to have achieved higher precision. By analyzing RGB (Color model) video, Aubry et al.[7] developed a novel approach to identifying actions. In order to do this, the motion in the film must be removed and the human skeleton extracted. Extracting a 2-dimensional skeleton with 18 known joints from each body was done using Open Pose [8], a Deep Neural Network (DNN)-employ identification approach. In the second case, an image classifier is used to transform motion patterns into RGB images. R, G, B Channels used to store motion information. An RGB image for a scene of activity is created in this way. Neural networks now used for picture classification may one day be trained to identify human behaviors.
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. Using a hierarchical RNN model, Du et al. [10] developed a skeleton-based approach to action recognition. In addition, five distinct deep RNN designs were tested against their proposed methods. Their comprehensive review made use of the following datasets: HDM05, the Berkeley MHAD, and MSR Action-3D. By creating temporal links and adding spatial and motion information to an existing LSTM module, Majd and Safabakhsh[11] created the Correlational Convolutional LSTM. They got 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. With this strategy, teams were able to finish 90.5% of the volleyball dataset, whereas individuals only managed 8.5%. In the study by Huang et al. [13], features based on posture are retrieved from a 3D convolutional neural network (ConvNet) by combining information about motion, 2-dimensional appearance, and 3-dimensional stance. 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]. This method, similar to twostream networks, uses RGB variation frames to make it seem like something has changed, and it uses optical flow fields with RGB and optical flow frames to cut out moving backgrounds. In [15], the author used a graph pooling network and a GCN with a channel attention method for the joints. In the end, the SGP architecture enhanced convolution by include the human skeletal network. In order to get specific information about the human body, kernel receptive areas are used. With the proposed SGP method, computation costs can be reduced and GCNs' ability to gather based on motion characteristics can be significantly enhanced. This study used two different stream designs: context stream and fovea stream [16]. Although the fovea channel receives the central region 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 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. In their proposal for a highly connected ConvNet, bidirectional-LSTM, Singh et al.[17] used RGB frames as the top layer to recognize human activities. Each DMI contributes to the learning process that occurs in the bottom layer of the ConvNet model. In order to enhance the pre-trained CNN's features, the ConvNet-Bi-LSTM model is trained from the ground up for RGB frames. In the meanwhile, the pre-trained ConvNet's highest layers are fine-tuned to extract temporal information from video streams. 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 a few deep learning models, how they function, and how precisely they categorize each action. Training a deep learning model from start requires a lot of processing power. Learning models are taught differently from transfer learning models. They are trained using ImageNet's massive dataset [18]. For the purpose of training transfer learning models, ImageNet contains over 1 million photos. In order to categorize each action, this study contrasted many transfer learning models with state-of-the-art approaches. In order to identify human actions, this

ISSN NO: 9726-001X

Volume 13 Issue 02 2025

JOURNAL OF CURRENT SCIENCE

study examined several transfer learning models. The combination of the pre-trained deep learning model with the Human Action Recognition model is shown in Figure 2. To assess methods that rely on transfer learning (TL), Dense Net[19] is used. Dense Net neural networks were selected because to their innovative methods for handling decreasing or increasing gradients and its unique architecture, which enables a single layer to acquire knowledge from the feature maps of previous layers, enabling the reuse of features. 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. Here, we used VGG models with batch normalization layers to rein in gradients and solve the issue. To assess the efficiency of the framework, the Efficient Net[21] approach is also used. A. Dense Network The name "Dense Net" comes from the fact that each layer of a dense convolution neural network (Dense Net)[19] is connected to every layer of the network in an extremely dense fashion.

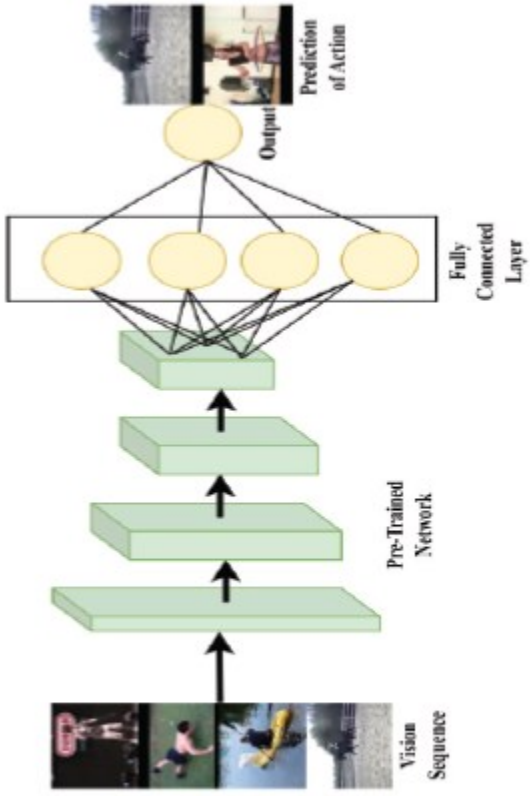

Fig. 2. HAR using pre-trained DL method.

A large-filter-size Conv2D layer is the first port of call for the data, followed by a dense block that forms dense connections to every subsequent layer. Every Dense Net layer takes data from every layer below it and sends its feature maps to every layer above it. Option B: VGG Our TL-based method for action recognition additionally makes use of VGG [20], a CNN architecture. Images sent into VGG for training must adhere to a certain ratio, meaning they must be 512 pixels wide by 512 pixels high (224, 224, 3). Several convolutional layers with filters of size 3 by 3 pixels were used to process these photos. Spatial pooling is performed via five max-pooling layers following particular conv2D layers. Dense layers with complete connectivity and a SoftMax prediction layer follow a stack of convolutional layers. Figure 3 shows the VGG19 architecture, with the words "conv," "pool," and "FC" standing for the many layers that make up the network.

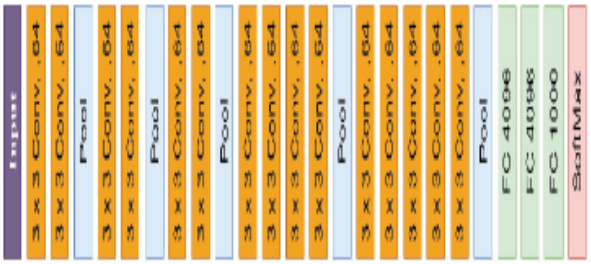

Fig. 3. VGG19 Architecture

Efficient Net

Efficient Net[21] is a method for designing and scaling convolutional neural networks that uses a compound coefficient to uniformly scale the depth, width, and resolution parameters. 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. Efficient Net [21] is a one-of-a-kind convolutional neural network (CNN) that 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. Due to its superior feature extraction capabilities, EfcientNet-B0 surpassed Resnet-50[22] with less parameters and FLOPs (floating-point operations per second) accuracy, all while using the same input size. The dataset used to evaluate the performance of the

ISSN NO: 9726-001X

Volume 13 Issue 02 2025

JOURNAL OF CURRENT SCIENCE

model is UCF50[23]. Reddy et al. (2012) first suggested this dataset. Videos are gathered via online channels such as YouTube. Nothing in these films is staged; instead, they all feature natural settings. Compared to the UCF11 dataset, this one has been revised and improved. It has fifty activity lessons like shooting, riding, shooting, playing the tabla, violin, etc. 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. Similarities in characters, settings, and points of view are common in films that fall into the same genre. The UCF 50 dataset's action snippets are shown in Figure 4.

DISCUSSION AND RESULTS

Dense Net, VGG19, and Efficient Net were the three pre-trained deep learning models utilized for activity classification. To make the most of the data collected from large datasets like ImageNet, we used pre-trained deep learning. In order to train a neural network for a new domain, the transfer learning method uses data from an existing model that has already been trained. The UCF50 was evaluated.

Fig. 4. UCF50 Action Dataset Frames.

activity dataset, which includes several picture categories. Using this strategy, we evaluated the accuracy of multiple deep learning models on the aforementioned dataset in comparison to state-of-the-art approaches. To begin, a pre-trained deeplearning model was fed frames collected from each set of action footage. 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.

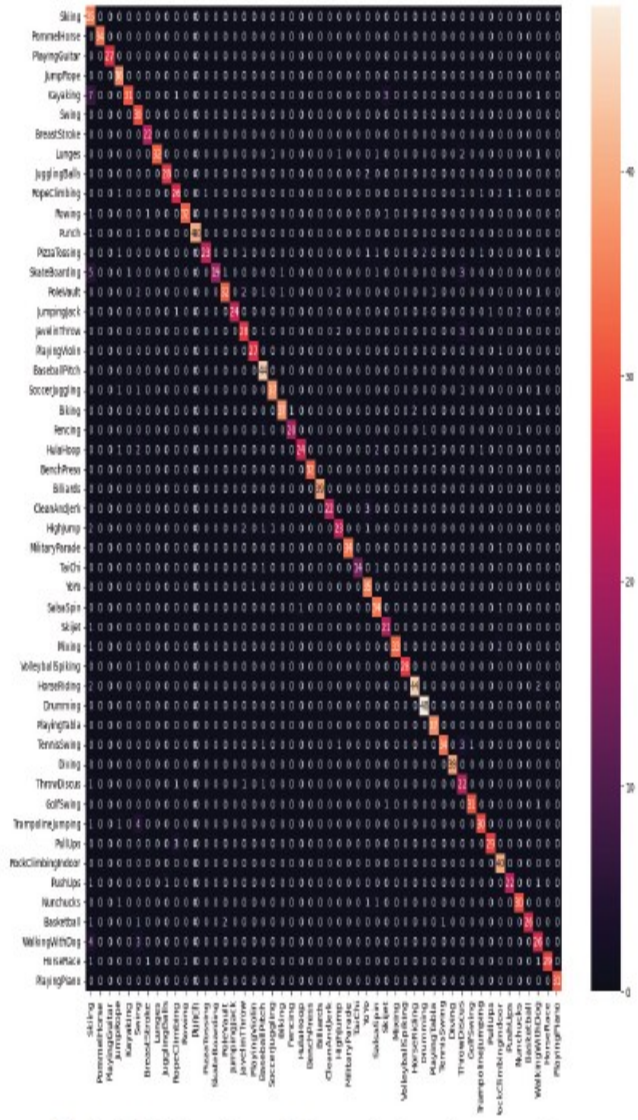

Fig. 5. VGG19 model confusion matrix for action recognition.

ISSN NO: 9726-001X

Volume 13 Issue 02 2025

JOURNAL OF CURRENT SCIENCE

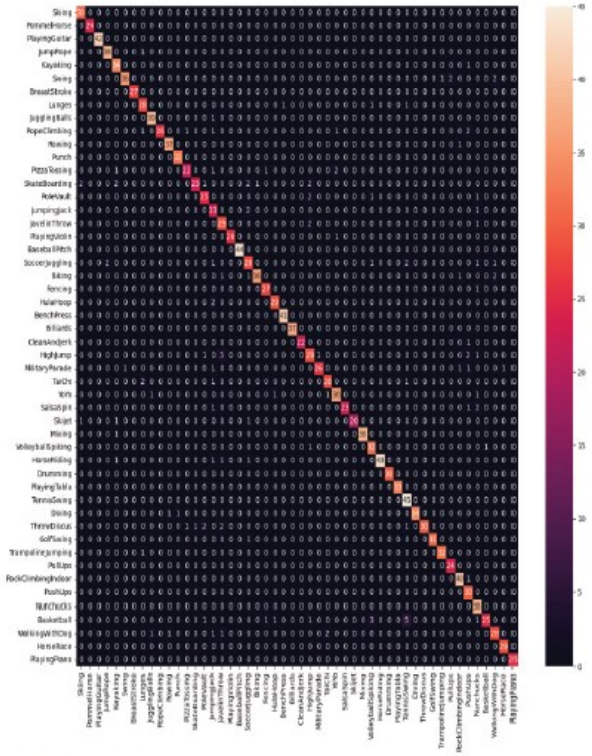

Fig. 6. Utilizing Dense Net 161 model, a confusion matrix for action recognition.

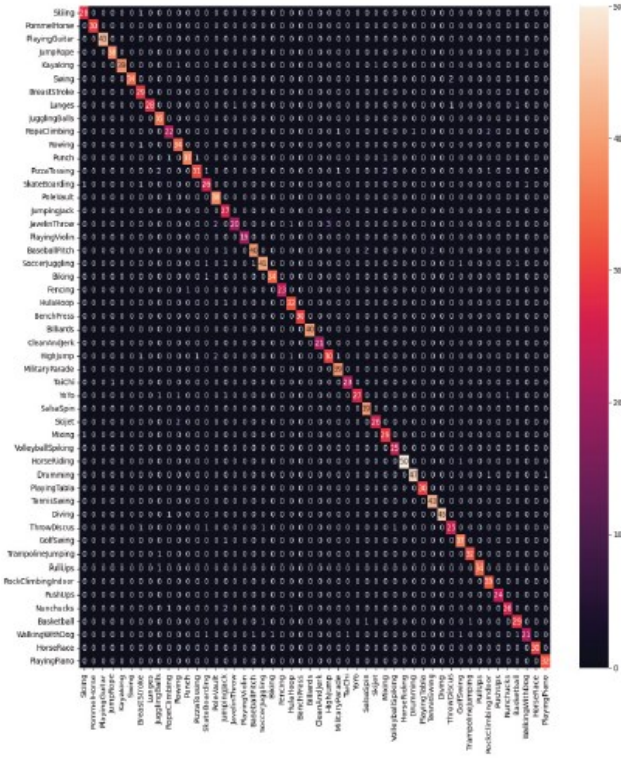

Fig. 7. Confusion matrix for action prediction from Efficient Net b7 model.

Using the UCF 50 activity dataset, the classification result is shown as a Confusion matrix. We can confidently and properly categorize the majority of the actions. Model assessment metrics using TL approaches are compared in Table 1 on the UCF50 action dataset. In the implementation step, the recovered frames were partitioned using the training, validation, and testing stages. An example of this is shown graphically in Figure 8. In Table 2 we may see a comparison with several state-of-the-art approaches:

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

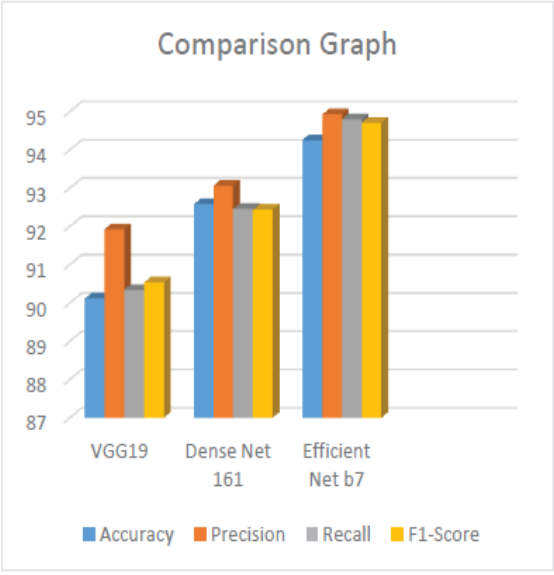

Fig. 8. Comparison graph for evaluation metrices.

TABLE II. COMPARISON OF LIGHT WEIGHT 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 our method's performance on the UCF 50 dataset to that of many other approaches that did not use transfer learning. Results showed that applying TL to a comparable dataset did really boost recognition score. When pretrained deep learning is applied, their classification performance is improved by 1-4 percent.

CONCLUSION

The UCF 50 action dataset is used to train deep learning models that can classify human actions. There are a total of fifty action categories in the UCF50 action dataset, organized into twenty-five groups. Each group has four videos. Precision, recall, f1 score, and area under the curve (AUC) were some of the evaluation matrices used to assess the efficacy and accuracy of the model. Models such as Efficient Net, Dense Net 161, and VGG19 categorize each dataset action. Applying cutting-edge techniques to the UCF50 dataset, this study also compared them. When pitted against current best practices, these pretrained DL models come out on top. With 94% accuracy, Efficient Net outperforms other pre-trained deep learning models. Our current work can be expanded to include the classification of actions in other datasets, real-time action monitoring, detection of anomalous actions, and crowd behavior in the future. This study then modifies the pre-trained deep learning model's architecture, for example by including an attention layer, so that it can be used with Bi-LSTM.

REFERENCES

[1] P. Pareek and A. Thakkar, “A survey on video-based Human Action Recognition: 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 Ten Years: A Comprehensive Survey,” Archives of Computational Methods 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: decomposing activities 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 Convolutional Networks for Action Recognition in Videos.”

[5] S. Ji, W. Xu, M. Yang, and K. Yu, “3D Convolutional neural networks for human action recognition,” IEEE Trans Pattern Anal Mach Intell, vol. 35, no. 1, pp. 221–231, 2013, doi: 10.1109/TPAMI.2012.59.

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

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

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

[9] C. Dai, X. Liu, and J. Lai, “Human action recognition using twostream attention-based LSTM networks,” Applied Soft Computing Journal, vol. 86, Jan. 2020, doi: 10.1016/j.asoc.2019.105820.

[10] Y. Du, W. Wang, and L. Wang, “Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition.”

[11] M. Majd and R. Safabakhsh, “Correlational Convolutional LSTM for human action recognition,” Neurocomputing, vol. 396, pp. 224–229, Jul. 2020, doi: 10.1016/j.neucom.2018.10.095.

[12] M. Qi, Y. Wang, J. Qin, A. Li, J. Luo, and L. van Gool, “StagNet: An Attentive Semantic RNN for Group Activity and Individual Action Recognition,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 2, pp. 549–565, Feb. 2020, doi: 10.1109/TCSVT.2019.2894161.

[13] Y. Huang, S.-H. Lai, and S.-H. Tai, "Human Action Recognition Based on Temporal Pose CNN and Multi-Dimensional Fusion."
[14] Wang Limin et al., Computer Vision – ECCV 2016, vol. 9912. Cham: Springer International Publishing, 2016. doi: 10.1007/978-3-319- 46484-8.
[15] Y. Chen et al., "Graph convolutional network with structure pooling and joint-wise channel attention for action recognition," Pattern Recognit, vol. 103, Jul. 2020, doi: 10.1016/j.patcog.2020.107321.
[16] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and Understanding Recurrent Networks," Jun. 2015, [Online]. Available: http://arxiv.org/abs/1506.02078
[17] T. Singh and D. K. Vishwakarma, "A deeply coupled ConvNet for human activity recognition using dynamic and RGB images," Neural Comput Appl, vol. 33, no. 1, pp. 469–485, Jan. 2021, doi: 10.1007/s00521-020-05018-y.
[18] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," Int J Comput Vis, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.
[19] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Nov. 2017, vol. 2017-January, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.
[20] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1556
[21] M. Tan and Q. v. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," May 2019, [Online]. Available: http://arxiv.org/abs/1905.11946
[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Dec. 2016, vol. 2016-December, pp. 770–778. doi: 10.1109/CVPR.2016.90.
[23] K. K. Reddy and M. Shah, "Recognizing 50 Human Action Categories of Web Videos."
[24] L. Zhang and X. Xiang, "Video event classification based on twostage neural network," Multimed Tools Appl, vol. 79, no. 29–30, pp. 21471–21486, Aug. 2020, doi: 10.1007/s11042-019-08457-5.
[25] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A Robust and Efficient Video Representation for Action Recognition," Int J Comput Vis, vol. 119, no. 3, pp. 219–238, Sep. 2016, doi: 10.1007/s11263-015-0846-5.
[26] Q. Meng, H. Zhu, W. Zhang, X. Piao, and A. Zhang, "Action recognition using form and motion modalities," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 16, no. 1s, Apr. 2020, doi: 10.1145/3350840.
[27] A. Jalal, I. Akhtar, and K. Kim, "Human posture estimation and sustainable events classification via Pseudo-2D stick model and K-ary tree hashing," Sustainability (Switzerland), vol. 12, no. 23, pp. 1–24, Dec. 2020, doi: 10.3390/su12239814.