2023 6th International Conference on Information Systems and Computer Networks (ISCON)
GLA University, Mathura, India. Mar 3-4, 2023

Light-Weight Deep Learning Model for Human Action Recognition in Videos

Rahul Kumar
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
rahuldtucs@gmail.com

Shailender Kumar
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
shailenderkumar@dce.ac.in

2023 6th International Conference on Information Systems and Computer Networks (ISCON) | 979-8-3503-4696-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISCON57294.2023.10111975

***Abstract*— Human Action Recognition (HAR) from a visual stream has recently attained much researcher consideration in the domain of computer vision. Due to its large applications like monitoring of health, home automation, and tele-immersion, among others. However, it still faces human variances, occlusion, lighting changes, and complicated backgrounds. The evaluation criteria rely on the features collection approach as well as learning data being performed correctly. The success of Deep Learning (DL) has resulted in a variety of impressive outcomes, including neural networks. Nonetheless, a robust features vector is required for an efficient classifier to give the class label. Features serve as the essential component of any data set. Indeed, feature extraction may affect the algorithm's performance and computational cost. For this research framework, we used pre-trained deep learning models VGG19, Dense Net and Efficient Net for feature extraction from the sequence of images and classified each action with the help of the SoftMax layer. UCF50 action dataset used, which contains 50 sections and evaluates performance with the help of precision, recall, f1-score and AUC score. Testing accuracy from models achieved VGG19-90.11, DenseNet-92,57 and EfiicientNet-94.25.**

Keywords—Transfer Learning, CNN, VGG19, UCF50

I. INTRODUCTION

In HAR, an action is an observable thing that may be seen by the human eye or by some sensor device. In fact, engaging in an activity like walking requires constant attention to a person located inside the field of view. Actions may be separated into four categories depending on the body components needed to do them. [1].

- Gesture: It's based on facial expression. Don't need any action or verbal way of communication.
- Action: Walking, playing, punching human action comprised.
- Interaction: It Consist human object interaction as well human interaction like handshake, hugging is example of interaction.
- Group activity: when more than two action is happening is known as group activity like combination of gesture and interaction. Two or more than two actor is involved for performing action.

Over the course of the last two decades, HAR has emerged as an essential component of research into computer vision. Based on a collection of observations, HAR is designed to detect and identify activities carried out by one or more persons. This may be done for any number of people. This had developed into a need for the further development of human–computer interaction. Many researchers from different parts of the globe are drawn to this field of study due to the wide range of domains in which it may be used. Surveillance video, the labeling and retrieving of visuals, monitoring of health, automation, and environmental modelling are but a few of its most prominent applications[1]. Human activities have an intrinsic hierarchical structure that denotes their many levels, which may be categorized at three levels. First, there is an atomic element at the lowest level, and these action primitives represent increasingly complicated human actions. The actions/activities level follows the action primitive level as the second level. The greatest level of categorization for human activities is represented by complex interactions. Each of these categories is sufficiently extensive to need its own area of study. This is mostly due to the unpredictability and ambiguity of human acts in a real-world situation. HAR faces several barriers and obstacles. Gender discrimination, multi-subject interactions, and disparities in inter-class activity are some examples. Human action recognition form videos have four stage process. In first step we extract feature from given sequences of images. In the feature extraction process, handcrafted methods just like SIFT (scale invariant feature transform), SURF (Speed up Robust Feature), Shape-based, Pose-based, and optical flow, to mention a few, may be used [1]. Feature extraction can be done with deep learning method automatically in this approach model learn itself all feature from sequences of images. It involves the extraction of poses and gesture patterns from frames and sequences of visuals depicting human activities. Therefore, it is a challenging task due to obstacles such as size fluctuations, poor illumination, incorrect viewpoints, and backdrop clutter.

In a subsequent stage, actions are learned and recognized based on extracted features. Learning new models that are taught by extracted features is an important part of action learning and recognition, as is determining which features are relevant to which action classes and assessing those features with the aid of classifiers. The Machine Learning (ML) technique and the DL method are two of the most notable approaches to solve the HAR issue. The former, the most conventional kind of Artificial Intelligence, takes a more deterministic approach, requiring the user to layout, dictate, and fine-tune the extracted features and characterize action. Using the latter strategy, however, we anticipate that the deep neural network (DNN). When we use the latter approach, on the other hand, we predict that the DNN will solve all of the attributes on its own by act as the function of the human brain [1][2].

Figure 1 depict ML and DL base classification for HAR.

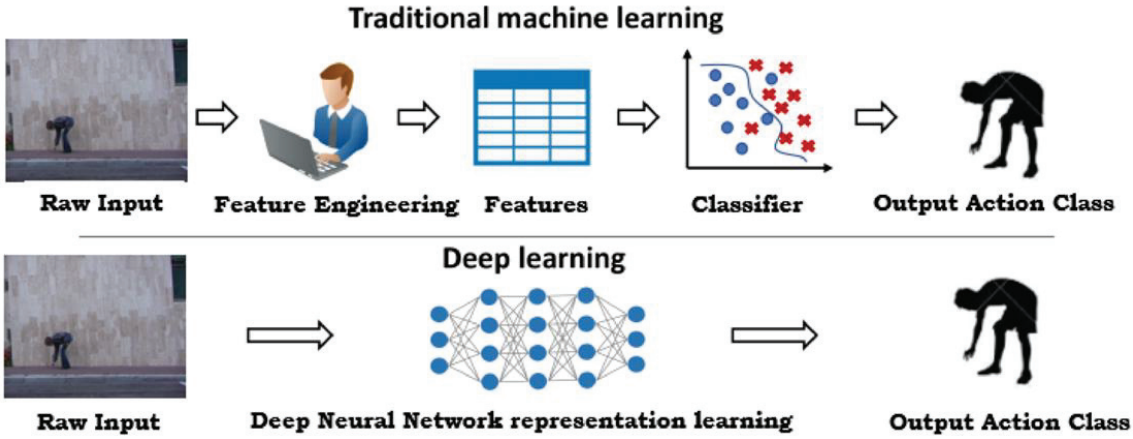

Fig. 1. A graphical representation of the conventional ML methods and the cutting-edge DL methods employed for HAR [2].

ML base methods such as random forest (RF), Bayesian networks (BN), Markov models (MM), and support vector machine (SVM) have been employed for decades to try to solve the HAR issue associated with it like clutter background, noise issue, class similarity issue. Accustomed ML algorithms have been able to attain excellent performance in settings with limited data inputs and stringent restrictions. Machine learning algorithm time consuming and need special attention due to preprocessing step with the handcrafted feature; they need to perform better When the data size is huge. DL has attained lot of gained in last years. This is due to the fact that research that is focused on deep learning has achieved excellent performance in a variety of study fields, such as detection of object in frames and recognition of action, classification of frames, and natural language processing.

By unmanned features pull out through numerous hidden layers, DL significantly reduces the effort required to select the appropriate features compared to traditional ML algorithms, and its structure has been shown to be effective for unsupervised learning and reinforcement learning. Because of this, there has been a rise in the number of proposed deep learning-based HAR frameworks.

The following is a summarization of the research paper:

In the first section, we provide a high-level overview of human action identification and then examine machine and deep learning techniques for this task. Following that, in Section 2, we will talk about the modalities and accuracy levels of previously proposed techniques to human action recognition. Methods and results on the dataset are discussed in Section 3. Section 4 summarizes the work done and the planned future development in the area of computer vision.

II. Related Work

In domain of human action recognition lots of work has been done. Due to its wide variety of application, there is lot of scope to improve the prediction of human action. Many manually produced and automated learned feature-based techniques for human activity detection in visuals have been developed during the last decade. Earlier techniques to human activity identification relied on handmade characteristics that were primarily focused on small atomic activities that seem to be less relevant for actual applications [3].Despite obtaining a model with a high degree of accuracy, the primary downside of these methods is that they need extensive data preprocessing and are difficult to generalize in practice. Following the success of convolutional neural networks (CNNs) in text and visual classification, many spatiotemporal approaches for video activity analysis have been created; these algorithms can automatically train and classify from raw RGB video [4].

Shuiwang Ji et al.[5] introduced a 3D convolution method for extracting spatial and temporal video data for action identification. As a result, the envisioned architecture creates many channels of data from the video sequence and applies convolution and subsampling to each channel individually. Gu et al. offered a DL based method to identify locomotive motions relevant to indoor localization and navigation. Their solution made use of stacked denoising auto-encoders that learnt data characteristics automatically, decreasing the requirement to manually build the relevant features [6]. Higher precision attained claimed in proposed research framework as compare to another classifier. Aubry et al.[7] came up with a new way to figure out what an action is by looking at RGB (Color model) footage. To do this, first the video's motion has to be taken out and provide the human skeleton extraction. Open Pose [8] , a Deep Neural Network (DNN)-employ identification method that pulls out a 2-D skeleton with 18 known joints from each body, was used to do this extraction. In the second instance, motion patterns are turned into RGB pictures with the help of an image classifier. R, G, B Channels used to store motion information. This makes an RGB picture for an action sequence. In the future, neural networks that are used to classify images could be used to recognize actions.

Dual stream model suggested by researcher Dai et al.[9] which use attention-based LSTM structure for localizing action in frames of visuals. They said they had solved the problem of ignoring visual attention. With the UCF11 dataset, the architecture was 96.9% accurate, with the UCF Sports dataset it was 98.6% accurate, and with the j-HMDB dataset it was 76.3% accurate. Du et al. [10] came up with a skeleton-based method for recognizing actions employing a hierarchical RNN model. Also, their suggested techniques was evaluated(compared) to five different deep RNN designs that were based on it. During their whole evaluation, they used the MSR Action-3D dataset, the Berkeley MHAD dataset, and the HDM05 dataset. Majd and Safabakhsh[11] devised the Correlational Convolutional LSTM by incorporating spatial and motion information into a

preexisting LSTM module in addition to establishing temporal linkages. Their work was assessed on the UCF101 and HMDB51 benchmark datasets, both of which are widely used, and they achieved correctness(accuracy) of 92.3% and 61.0%, respectively. For the purpose of recognizing both group and individual activities, Qi et al.[12] offered a different approach to building a semantic RNN known as stag-Net. They added a fourth dimension to their semantic network model—time—with the help of a structural RNN. Using this method, 90.5% of the Volleyball dataset was completed as a team and 8.5% as an individual effort.

Posture-based characteristics are extracted from a 3D convolutional neural network (ConvNet) in Huang et. al.[13] by fusing 3-D pose, 2-D appearance, and information of motion. The 3-D CNNs attained of features of color joint in frames is expected to be computationally intensive, thus we perform convolution in each of the heatmap's 15 channels to reduce the noise. The (BN-inception) network architecture was used by Wang et al. in Inception and Batch Normalization [14]. The aforementioned approach, like two-stream networks, employs RGB variation frames (to simulate appearing alteration) and optical flow fields in conjunction to RGB and optical flow frames (to inhibit background motion).

The author in [15] utilized the GCN with channel attention strategy for joints and graph pooling network. Finally, the SGP design incorporated the human skeletal network and improved the convolution. Kernel receptive regions are used to collect particular human body data. The suggested SGP technique has the potential to considerably improve GCNs' capacity to gather based on motion characteristics while also lowering computation costs.

Context stream and fovea stream designs employed in research article [16]. The context channel receives frames at half their original resolution, but the fovea channel receives the center region at full resolution. The study trains a model to detect three separate pattern classes: Early Fusion, Late Fusion, and Slow Fusion using each video as a collection of small, fixed-length clips. CNN may create single-frame animations by combining time and space in various ways. Singh et. al.[17] proposed a highly linked ConvNet with RGB frames as the top layer for identifying human activities-term, bidirectional-LSTM. The lowest layer of the ConvNet model is learned(trained) with the help of individual DMI. The ConvNet-Bi-LSTM model is trained from scratch for RGB frames to improve the features of the pre-trained CNN, while the uppermost layers of the pre-trained ConvNet are adjusted(fine-tune) to extract temporal information from video streams. Features are combined at the decision layer to get a higher precision value using a late fusion technique following the SoftMax layer. Four RGB-D (depth) datasets including both single-person and multiple-person activities are employed to examine the effectiveness of the suggested model.

III. METHODOLOGY

The DL model for HAR shows the impactful result on the classification of each activity. We discussed some deep-learning models, how they work and classify each action accurately. Deep learning model training is required from scratch and requires lots of computing power. Compared to transfer learning models are trained, learning models. They are learned on the huge amount of data ImageNet [18]. ImageNet has more than 1 M of images suitable for training transfer learning models. This research paper used various transfer learning models to classify each action & compared them with state-of-the-art methods. This research paper compared various transfer learning models to recognizing human action. Figure 2 represents the Human Action Recognition model with the pre-trained deep learning model.

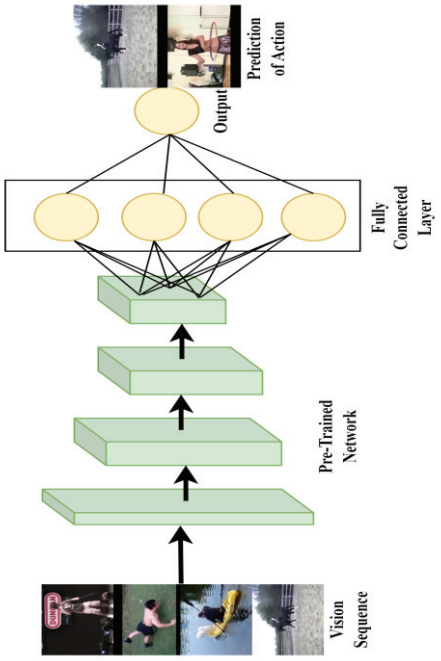

Fig. 2. HAR using pre-trained DL method.

Dense Net[19] are used to evaluate transfer learning (TL)-based methodologies. Dense Net neural networks were chosen for their new techniques to coping with vanishing or growing gradients, as well as their particular design, which allows one layer to learn from the feature maps of preceding layers, allowing feature reuse. VGG[20] is also trained via transfer learning-based HAR approach because to its very deep architecture, which is accomplished by using tiny (3 3) filters. As a result of their complexity, VGG models often experience gradient explosions. To address this problem, we employed VGG models with batch normalization layers to keep gradients under control. Efficient Net[21] method is also used to evaluate performance of framework.

A. *Dense Net*

A dense Convolution neural network (Dense Net)[19] uses a feed-forward fashion to interconnect each subsequent network layer; this extensive interconnectedness has earned it the moniker Dense Net. The data is first routed via a Conv2D layer with a high filter size, then through a dense block that establishes dense connections with all following layers. Each Dense Net layer receives new inputs from all previous layers and broadcasts its feature-maps to all subsequent layers.

B. *VGG*

VGG [20] is a CNN architecture that we also included into the TL-based approach to identify human actions. The pictures provided to VGG for training are images of a strict ratio, i.e., 512 by 512 pixels (224, 224, 3). These images have been processed using a stack of convolutional layers with 3-by-3-pixel filters. Following specific conv2D layers,

five max-pooling layers do spatial pooling. A stack of convolutional layers is followed by dense layers with full connection and a SoftMax prognostication layer.

Figure 3 represent the VGG19 architecture where conv belongs to convolution layer, pool represent pooling layer and FC belong to fully connected layer

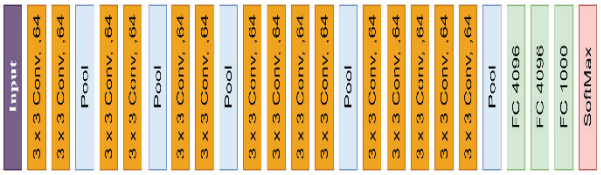

Fig. 3. VGG19 Architecture

C. *EfficientNet*

Efficient Net[21] is an architectural and scaling strategy for convolutional neural networks that scales all parameters of depth/width/resolution evenly utilizing a compound coefficient. Unlike current practice, which randomly scales these elements, the Efficient Net scaling technique uniformly adjusts network breadth, depth, and resolution using a set of predefined scaling factors. Efficient Net [21] It is a one-of-a-kind CNN network with high parameter estimation efficiency and speed. Efficient Net [21] used a simple and complicated scaling approach to more systematically scale up CNN models by scaling network characteristics like as depth, breadth, and resolution uniformly. In classification tasks, Efficient Net [21] was utilized also as spatial feature extraction network. There were seven CNN models dubbed EfcientNet-B0 through EfcientNet-B7 in the Efficient Net family. With the same input size, EfcientNet-B0 outperformed Resnet-50[22] with less parameters and FLOPs (floating-point operations per second) precision, indicating that EfcientNet-B0 is capable of efficient feature extraction.

D. *Dataset*

UCF50[23] dataset used for evaluation of model performance. This dataset was proposed by Reddy et al. in 2012. Online platforms like YouTube are used for the collection of videos. All videos have a realistic environment and are not taken from a controlled environment. This dataset is an updated version of the UCF11 dataset. It contains 50 action classes like basketball, shooting, tabla playing, biking, violin playing etc. There are a total of 6618 films featuring a variety of activities ranging from general sports to everyday life activities. Each activity class is separated into 25 homogeneous groups, including at least four movies for each activity. Films in the same category may have some characteristics, such as the same person, backdrop, or perspective. Figure 4 represents the action snippets of the UCF 50 dataset.

IV. DISCUSSION AND RESULTS

To classify each activity, we used three pre-trained deep-learning models Dense Net, VGG19 and Efficient Net. We applied pre-trained deep learning to capitalize on the information gathered from massive datasets such as ImageNet. The transfer learning approach transfers information from a previously-trained model to a neural network to train a new domain. Evaluation of the UCF50

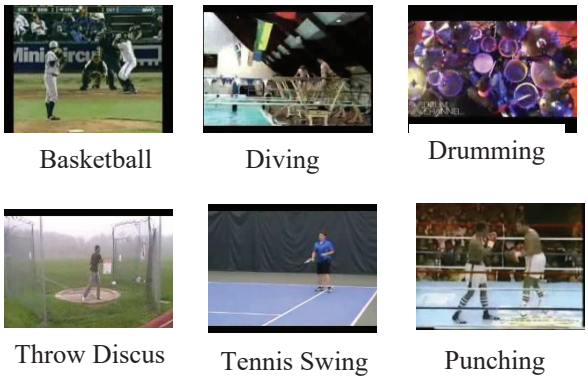

Fig. 4. UCF50 Action Dataset Frames.

action dataset, which contains various groups of images. We compared various deep learning model performances onthe above dataset in this approach and compared accuracy with state-of-the-art methods. Firstly, frames were extracted from each group of action videos and fed into a pre-trained deep-learning model. Figure 5-7 depicts the confusion matrix for recognizing 50 activities from UCF 50 dataset using the VGG19 model, Dense Net 161,EfficientNet b7.

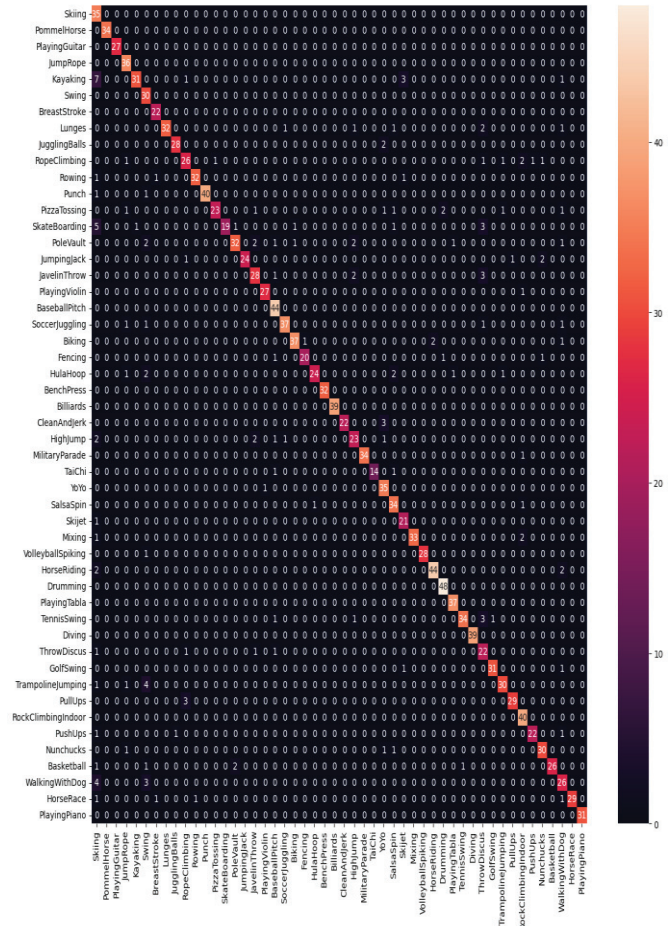

Fig. 5. VGG19 model confusion matrix for action recognition.

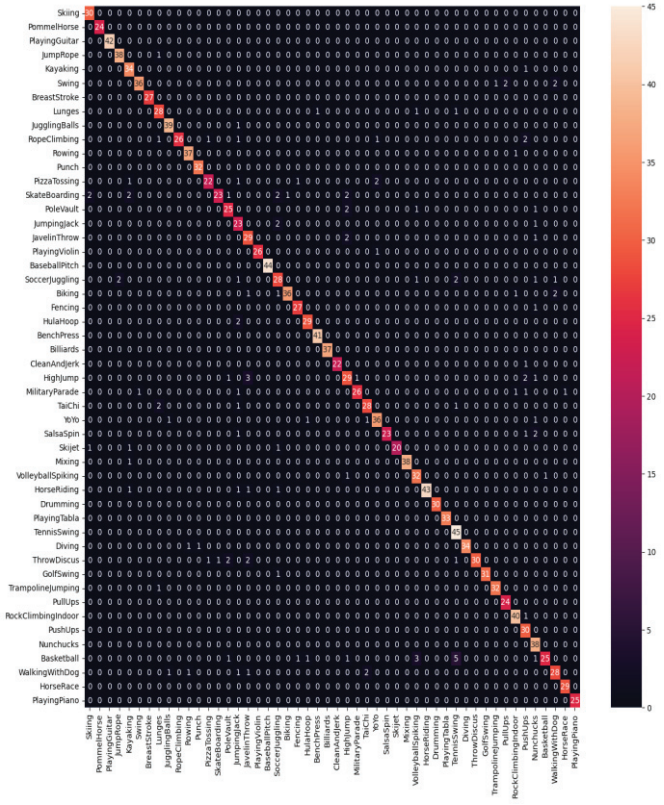

Fig. 6. Utilizing Dense Net 161 model, a confusion matrix for action recognition.

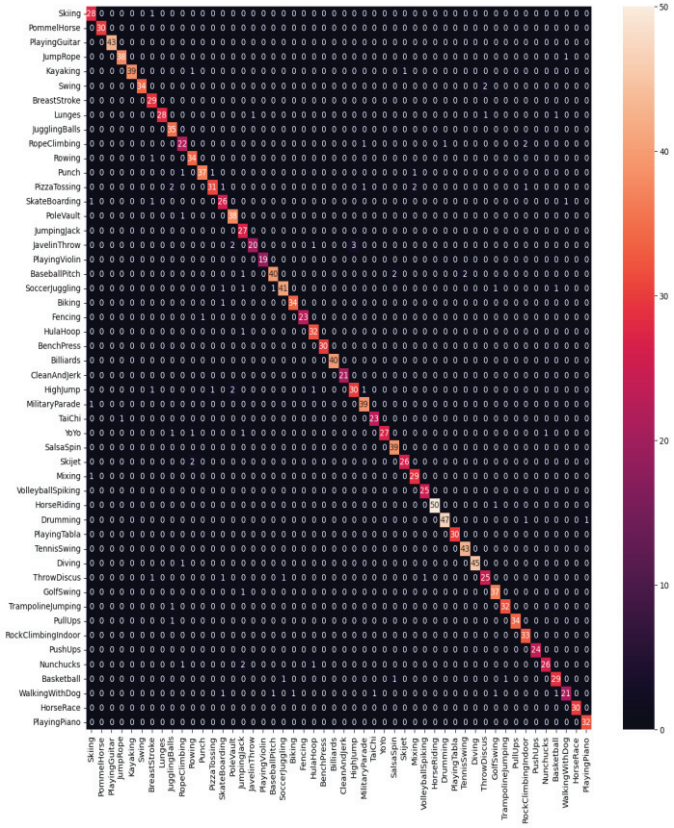

Fig. 7. Confusion matrix for action prediction from Efficient Net b7 model.

Classification result on activity dataset UCF 50 is shown as a Confusion matrix. Most of the activities classify accurately and with high confidence. On the UCF50 action dataset, Table 1 compares models evaluation metrices utilising TL techniques. Training, validation, and testing phases were used to partition the recovered frames during the implementation phase. Figure 8 shows a graphic illustration of this. Comparison with different state of the art methods 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

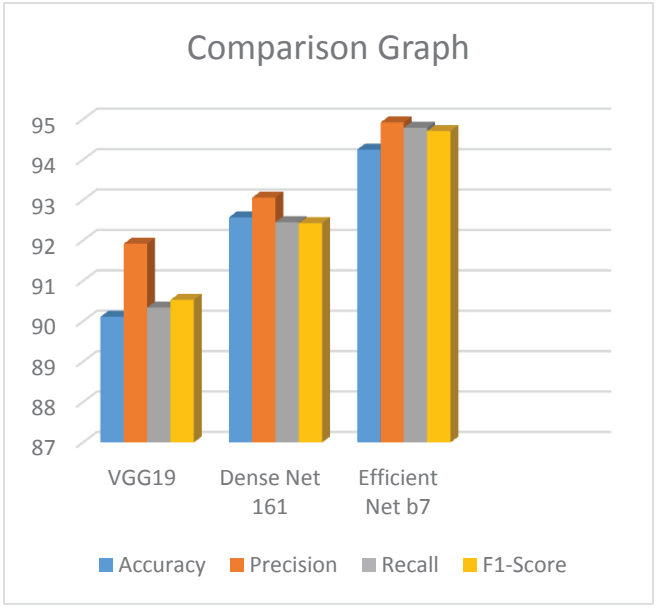

Fig. 8. Comparison graph for evaluation metrices.

TABLE II. COMPARISON OF 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**

On the UCF 50 dataset, we evaluated our approach's performance with that of a number of other methods that did not involve transfer learning. The findings of the experiment demonstrated whether employing transfer learning to a similar dataset improved recognition score. Their classification performance is enhanced by 1-4 % when pre-trained deep learning is used.

V. CONCLUSION

Pre-trained deep learning models are used to classify human action from UCF 50 action dataset. UCF50 action dataset contains 50 different action categories into 25 groups, each containing at least 4 videos. Various evaluation matrices were used for testing model accuracy and effectiveness like precision, recall,f1 score and AUC score.VGG19, Dense Net 161, and Efficient Net models classify each dataset action. This work also compared state-of-the-art methods applied to the UCF50 dataset. These pre-trained deep learning models perform better as compared to state-of-the-art methods. Efficient Net performs better than other pre-trained deep learning models with 94 % accuracy.

In the future, we can extend this work to classify another dataset action, real action monitoring, abnormal action detection, and crowd behavior. This research goes on to adjust the framework of the pre-trained deep learning model, such as adding an attention layer, such that the pre-trained deep learning model may be employed 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 two-stream 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 two-stage 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.