Learning Human Actions: A Walk Through

K. Jagrutha Aditya[1*], K. Narendra[2], D. Karthik Reddy[3], T. Anil[4] and Eva Patel[5]
{ kja8586@gmail.com[1], narendrakhareedu9@gmail.com[2], dkarthikreddy44@gmail.com[3], thiriveedhianil16@gmail.com[4], evapatel08@gmail.com[5] }

Department of Advanced Computer Science and Engineering, Vignan' Foundation for Science Technology and Research, Guntur, 522213, Andhra Pradesh, India[1, 2, 3, 4, 5]

Abstract. Human action recognition in video is essential for numerous intelligent systems, ranging from surveillance to medical applications. In this contribution, we describe a comparative evaluation of four Deep Neural Networks (DNN) architectures intended to efficiently learn and recognize human actions: 3D Convolutional Neural Networks (3D CNN), 3D CNN with ResNet backbone, Inflated 3D ConvNet (I3D), and the Video Vision Transformer (ViViT). These architectures are compared in terms of how well they are capable of learning rich spatial-temporal representations needed to understand dynamic human activities. Through the examination of the performance and shortcomings of every design, this research offers insights into the developing scenario of video-based HAR and indicates the advantages of transformer-based attention mechanism over conventional convolutional methodologies.

Keywords: Human Activity Recognition, 3D Convolutional Neural Network, ResNet, Spatiotemporal Features, Inflated 3D ConvNet, Video Vision Transformer.

1 Introduction

Human Action Recognition (HAR) is a core Computer Vision (CV) task that deals with detecting and classifying human actions from video streams. HAR is an important task in numerous applications including security services, human-machine interaction, sports analytics, and health monitoring. The goal of HAR is to inspect spatial and temporal patterns in videos to effectively identify various human activities in complex environments. Pre-vision methods focused on HAR considerably relied on heuristically created features, such as motion trails, optical flows, and space-time interest points. Although older methods yielded acceptable success, the methods were weak in terms of robustness and scalability for scenarios in real- life applications because of their reliance on domain- specific heuristics as well as inferior generalization capacities.

With increased use of DL, paradigm shifts have also occurred in methods of HAR. CNNs have played a crucial role in extracting spatial information from single frames, while 3D CNNs capture both spatial and short-term temporal behavior. CNN based approaches, though, tend to have difficulty modeling long-range temporal behavior, constraining their application to analyzing complicated or extended actions. Equally, RNNs and LSTM networks introduced sequence learning for videos but are plagued by vanishing gradients and heavy computational requirements for longer sequences. In the face of these challenges, recent breakthroughs in

ICITSM-Part I 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2357848

attention-based models, especially Transformer architectures, have performed well on HAR tasks owing to their improved capacity to handle global temporal dependencies. These models take advantage of self-attention mechanisms to handle full sequences in parallel, alleviating many drawbacks of sequential traditional architectures.

In this paper, we provide a thorough comparison of four recent models for video-based HAR: 3D CNN, 3D CNN with ResNet backbone, Inflated 3D ConvNet (I3D), and the Video Vision Transformer (ViViT). These models are tested on the popular UCF101 dataset, which consists of around 13.32k videos across 101 classes. Our objective is to determine the merits and demerits of every architecture in learning spatial- temporal representations and precisely classifying hu- man actions. The comparative study seeks to gain insights into developing effective and scalable HAR systems for practical applications.

2 Related works

Human Action Recognition (HAR) has been increasingly focusing on due to its use in surveillance, health care, human-computer interaction, and smart environments. Researchers over the years have submitted different models in order to enhance HAR in spatial-temporal feature learning, robustness, and computational complexity. The literature has progressed from initial CNN-RNN hybrids to even more advanced attention-based models such as Transformers and generative models such as VAEs and Graph Convolutional Networks (GCNs).

Early attempts at HAR with video were made using deep learning models where spatial or temporal learning was concentrated upon. Surek et al. [1] performed an extensive review on deep learning HAR methods that contrasted between basic CNN models and sophisticated neural models, demarcating their limitations and potentials over benchmarks. Babiker et al. [2] likewise discussed neural network structures for general human activity identification in everyday circumstances, especially within surveillance scenarios, but with respect to their functionality in basic activities of detection.

One significant change came with the incorporation of spatiotemporal modeling. Wu et al. [3] offered an in-depth survey of deep learning-based HAR models, highlighting how temporal modeling using LSTM and CNN hybrids enhanced activity detection for untrimmed videos. Additional advancements saw hybrid CNN-LSTM architectures introduced that effectively learned spatial patterns through CNNs and sequential motion using RNNs [7]. Following generative modeling, Sharma et al. [4] surveyed the use of VAEs and other deep architectures for HAR on benchmark video datasets. These generative approaches were especially effective in semi- supervised settings where labeled data is scarce. The application of residual learning, as utilized in ResNet-based HAR models, brought tremendous improvements in the depth of feature extraction and gradient flow. Mihanpour et al. [8] applied a ResNet–DB- LSTM hybrid, demonstrating how deeper spatial features and bidirectional temporal modeling enhanced recognition accuracy.

To deal with issues of real-time performance, Archana and Hareesh [10] suggested a ResNet

blended with 3D CNNs for rapid and trustworthy HAR. The integration of temporal and spatial modeling within one architecture was effective for on-the-fly classification of actions. With the emergence of attention mechanisms, Transformer-based models have accelerated HAR significantly. Arnab et al.'s [11] ViViT model is one of the significant milestones, which uses pure attention- based processing across video sequences to capture global spatial-temporal dependencies. This beats regular CNNs and RNNs, especially in the case of long- range temporal context capture.

Other deep learning-oriented HAR systems have been concerned with domain-specific applications. Sathya et al. [5] , for instance, proposed an intelligent surveillance system based on real-time action recognition using deep learning, combining CNNs and optimization methods for field deployment. Khurana and Kushwaha [7] also highlighted the importance of DL for HAR in video surveillances, surveying several architectures and their flexibility in challenging environments.Graph-based approaches have also come into existence, concentrating on human pose estimation and skeleton information. Ahmad et al. [6] gave an overview of GCNs for HAR, showing how they excel at modeling human joint correspondence and motion dynamics.

Recently, R. D. R. and P. C. J. [9] suggested models like Inception-v4 for HAR, which sought to balance depth and computational cost. Such models utilized sophisticated CNN architectures for real-world video action classification with high accuracy.Some Other recent developments in human activity recognition (HAR) have shown the strength of deep learning models in extracting intricate spatiotemporal patterns from video data. Ullah et al. [12] carried out a thorough systematic literature review comparing different deep neural networks for HAR, discussing trends, challenges, and performance measures across architectures. Zaidi et al. [13] discussed the use of deep learning to identify suspicious human actions from surveillance footage with a focus on the impact of precise HAR systems in public safety in real life. Sun et al. [14] presented an end-to-end approach to the Convolutional LSTM model and its ability to capture temporal dependencies to improve robust activity recognition.

In brief, the HAR field has evolved from early CNN-RNN hybrids [2], [3], [7] to strong generative models [4], [6], Transformer-based networks [11] and real-time efficient solutions [5],[8], [10] . Comparative analysis in our work compares four top-performing models 3D CNN, 3D CNN ResNet, I3D, and ViViT on the demanding UCF101 dataset. In this exploration, we seek to point out the performance-complexity-temporal modeling capability trade-offs in present deep learning techniques for HAR.

3 Methodology

This research explores and compares the performance of four deep learning models for HAR on the UCF101 video dataset. The selected model's 3D CNN, 3D CNN ResNet, (I3D) and (ViViT)are chosen due to their varied capabilities in spatial-temporal feature extraction. The methodology includes the following steps: data preprocessing, model architectures and evaluation.

3.1 Preprocessing

Preprocessing is an important step in transforming raw video data of the UCF101 dataset into a deep learning model-friendly format. First, the structure of the dataset is used to parse both video paths and related class labels, in which every subdirectory contains a distinct human action. These video paths and labels are arranged into NumPy arrays and divided into train-test sets with an 80:20 ratio to provide strong model evaluation. For every video, a fixed number of frames are sampled at regular intervals, namely every 15 frames, to capture significant temporal dynamics while maintaining the data size reasonable. The frame extraction is done using OpenCV, where a random initial point is chosen in every video, followed by sequential sampling of frames. If the number of frames needed is more than the available frames, then zero-padding is applied to ensure uniformity. The frames are then converted from BGR to RGB format and resized to 224×224 pixels using TensorFlow utilities. To augment the learning process, all the frames are normalized in the [0,1] range and then sent through a FrameGenerator class that produces pairs of frame sequences and their respective class labels. These pairs are used in TensorFlow's tf.data. Dataset API, with some optimizations added in the form of caching for efficiency, shuffling to introduce randomness, and prefetching to allow asynchronous data loading. Lastly, the preprocessed frame sequences are grouped into batches ideally of size 32 and thus creating a uniform input shape of type (batch size, frames, height, width, channels) which is acceptable for inputting into models.

3.2 Model Architecture

In this section, we discuss four deep learning architectures that illustrate the progression and effectiveness of deep learning architectures in HAR.

3.2.1 3D CNN

The 3D Convolutional Neural Network (3D CNN) architecture implemented in this study is a sequential model designed to effectively capture spatiotemporal features from video sequences. The input to the model is a video clip consisting of 10 consecutive frames, each of resolution 224×224 with 3 color channels, resulting in an input shape of (10, 224, 224, 3). The architecture begins with a 3D convolutional layer comprising 32 filters with a kernel size of 3×3×3, using the ReLU activation function and same padding to preserve spatial and temporal dimensions. This is followed by a 3D max pooling layer to reduce dimensionality and a batch normalization layer to stabilize and accelerate training.

Subsequent layers follow a similar pattern, progressively increasing the number of filters to 64 and 128 in the next two convolutional blocks. Each block includes a Conv3D layer, a MaxPooling3D layer, and Batch Normalization to ensure robust learning. These layers help the model extract complex motion and appearance features across frames by performing convolution operations in three dimensions height, width, and time.

After the feature extraction layers, a Global Average Pooling 3D layer is applied to flatten the 3D feature maps into a 1D vector, summarizing the spatial-temporal information. This is

followed by a fully connected dense layer with 256 neurons and ReLU activation, along with a dropout layer (dropout rate = 0.5) to mitigate overfitting. Finally, the model concludes with a dense output layer consisting of 101 neurons and a softmax activation function to classify each input sequence into one of the 101 action classes from the UCF101 dataset. This architecture strikes a balance between computational efficiency and feature richness, making it a suitable baseline for action recognition tasks in video data. Fig. 1 shows the 3D CNN Architecture.

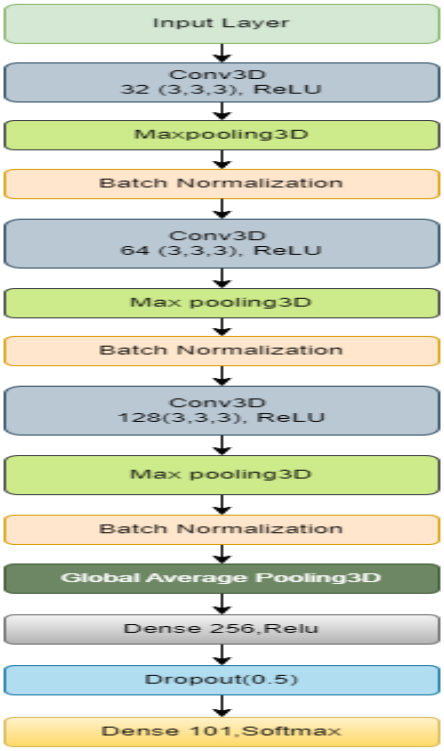

Fig. 1. 3D CNN Architecture.

3.2.2 3D CNN Resnet

The architecture used for human activity recognition in [10] is based on a 3D CNN, motivated by the ResNet architecture and augmented with Squeeze-and-Excitation (SE) blocks to enhance its capacity to learn intricate features. The network accepts an input of 10 consecutive frames, each resized to 224×224 pixels with 3 RGB channels, yielding an input shape of (10, 224, 224, 3). The first layer is a 3D convolution with 64 filters of size 7 as the kernel, batch normalization, a ReLU activation, and a max pooling layer with stride 2.

The central part of the model consists of a sequence of residual blocks, with each block having

two 3D convolutional layers, batch normalization, and ReLU activation. SE blocks are optionally inserted before every residual block to refine feature representation. These blocks make use of global average pooling and fully connected layers to learn channel-wise relationships and adaptively recalibrate feature maps. Fig. 2 shows the 3D CNN Resnet Architecture.

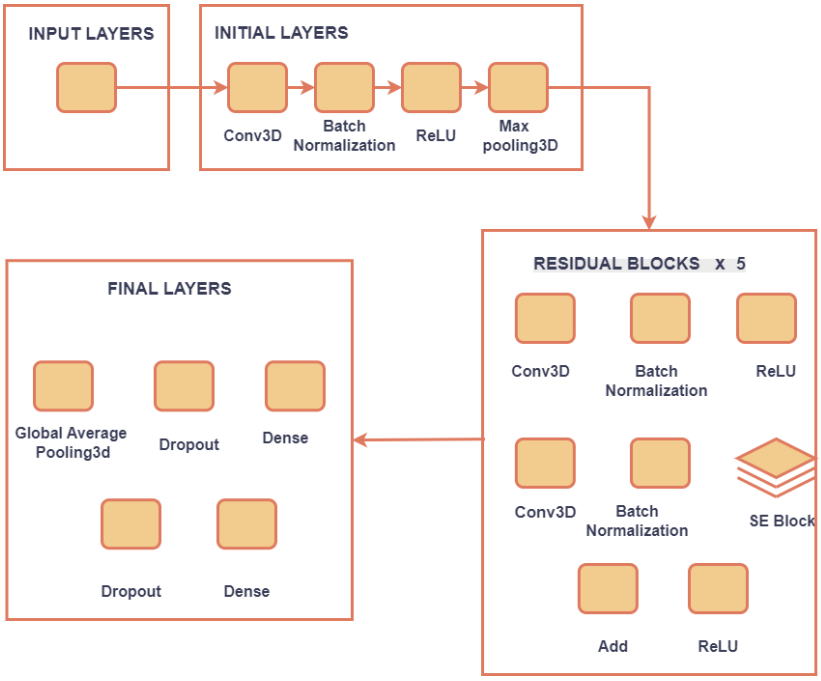

Fig. 2. 3D CNN Resnet Architecture.

As the network becomes deeper, the number of filters doubles from 64 to 1024 to capture hierarchical spatial and temporal features. Each residual block also includes shortcut connections to promote gradient flow and counter problems such as vanishing gradients. Following the stack of residual blocks is a global average pooling layer that shrinks the high-dimensional feature map into a 1D vector and then feeds it through fully connected layers with dropout used to prevent overfitting.

The last output layer employs a SoftMax activation function to generate probability distributions over 101 action categories. The model is trained with the Adam optimizer with weight decay, sparse categorical cross-entropy as the loss function, and accuracy as the main evaluation metric. Early stopping on the training loss is used to prevent overtraining. This 3D CNN structure, based on ResNet and augmented with SE blocks, efficiently extracts both spatial and temporal information and is thus highly appropriate for action recognition tasks.

3.2.3 Inflated 3D ConvNet

Aside from the 3D CNN ResNet, this work also reviews an Inflated 3D ConvNet (I3D) architecture similar to the work in [17], to improve video-based human activity recognition. The I3D model inflates regular 2D convolutional filters to 3D, allowing the network to extract both spatial and temporal information from a sequence of video frames. The model input is 10 frames of a video each of resolution 112×112 pixels with three color channels, giving an input shape of (10, 112, 112, 3). The model starts with a 3D convolutional layer with 64 filters and kernel size (7,7,7), followed by batch normalization and application of the ReLU activation function. It is further applied to max pooling in order to decrease the spatial dimensions.

The network goes on to include a few convolutional blocks and then gets into the middle of the structure - Inception modules. Such modules include more than one concurrent convolutional route: a 1×1×1 convolution, a 3×3×3 convolution that is preceeded by a 1×1×1 dimensionality reduction, a 5×5×5 convolution pathway similarly preceded by a 1×1×1 convolution, and a pooling pathway post-proceeded with a projection convolution. These paths are concatenated to enable the model to simultaneously process features of various scales and complexities. The architecture consists of a sequence of such inception modules, with progressively larger filter sizes to increasingly learn higher-level spatiotemporal features. Max pooling is placed strategically to decrease spatial and temporal dimensions without losing important feature information.

Following the inception blocks, the model performs global average pooling to compress the feature maps, then a dense layer with 512 units and ReLU activation. For regularization against overfitting, dropout is used. A softmax output layer finally predicts probabilities over 101 activity classes. The I3D model architecture by integrating the best of convolutional feature extraction along with multiscale spatiotemporal processing is well suited for action recognition in video data. Fig. 3 shows the Inflated 3D ConvNet Architecture.

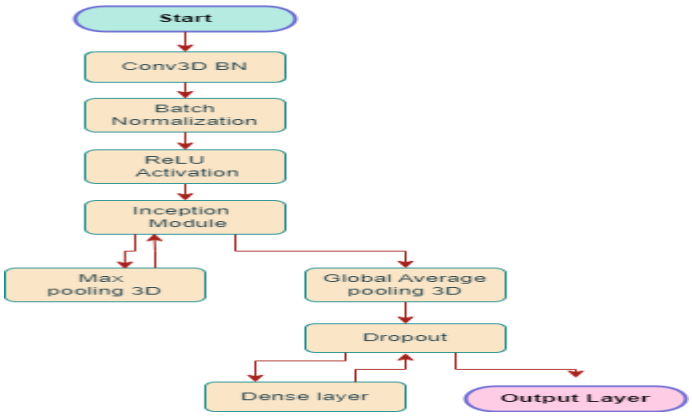

Fig. 3. Inflated 3D ConvNet Architecture.

3.2.4 Video Vision Transformer

The Video Vision Transformer model [11] borrows the success of Vision Transformers to video understanding by projecting their ability to process spatiotemporal data. The model accepts as input a sequence of video composed of several consecutive frames and starts by splitting these frames into smaller spatiotemporal patches, which are embedded into token representations. Every token is augmented with positional and token embeddings to preserve spatial and temporal structure. A classification token ([CLS]) is added to the beginning of the sequence for global information aggregation across the video.

These tokens are then passed into a basic Transformer Encoder architecture that consists of stacked layers (L) of multi-head self-attention and feed-forward neural networks wrapped with layer normalization and residual connections. At each encoder block, the model computes multi-head dot-product attention to enable the capture of advanced interdependencies across both time and space. Through this attention process, the model gains an insight into motion and object interaction along the length of the video. Fig. 4 shows the ViViT Architecture.

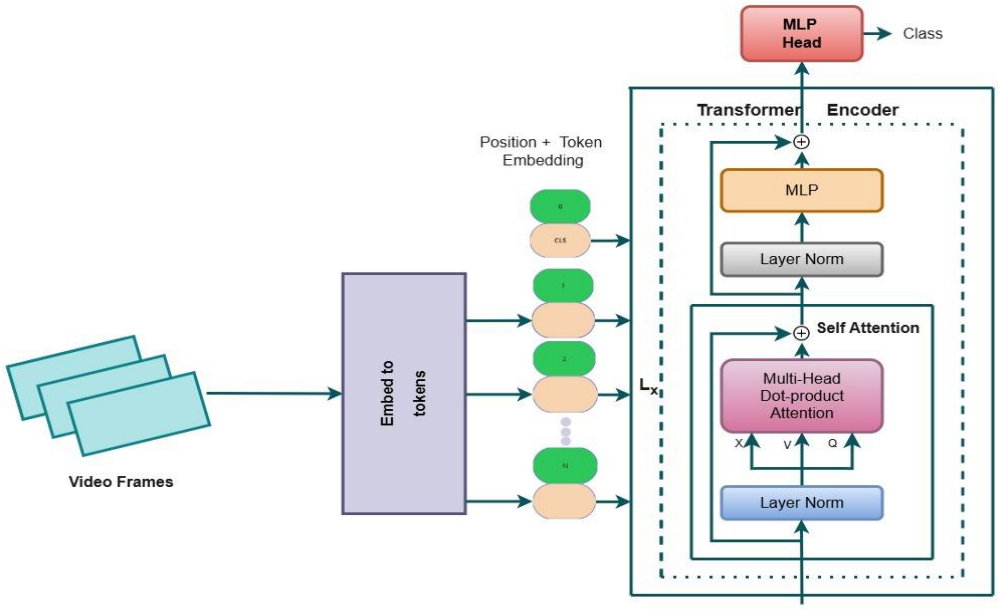

Fig. 4. ViViT Architecture.

Finally, the last output for the [CLS] token, after going through the encoder stack, is fed into a Multi-Layer Perceptron (MLP) head that projects the high-level features into the target output classes. This allows ViViT to learn video-rich, contextually aware representations purely

through attention without depending on the classical convolutions or recurrent architecture, a very effective replacement for video classification

4 Experimental Observations

For evaluating our model, we use various metrics that play a key role in determining how model performs on testing data. Among the numerous evaluation metrics available, we focus on some of the most commonly used: accuracy, precision, and recall.

4.1 Accuracy

Accuracy measures how frequently a model makes correct predictions, and it is determined using a specific formula.

$$Accuracy = \frac{T.P+T.N}{T.P+T.N+F.P+F.N} \quad (1)$$

4.2 Precision

Precision indicates the fraction of predicted positive cases that are truly positive. It is calculated using the formula:

$$Precision = \frac{T.P}{T.P+F.P} \quad (2)$$

4.3 Recall

Recall, also referred to as sensitivity or the true positive rate, represents the percentage of actual positive cases that the model correctly identified:

$$Recall = \frac{T.P}{T.P+F.N} \quad (3)$$

5 Results and Discussion

We evaluated the models using metrics such as accuracy, precision, recall and Loss in table 1.

Fig 5 to 12 show how well each model worked by displaying their accuracy and error (loss). These include results for 3D CNN ResNet (Figs. 5–6), I3D (Figs. 7–8), 3D CNN (Figs. 9–10), and ViViT (Figs. 11–12).

Table 1: Performance Comparison of HAR models.

Method	Accuracy	Precision	Recall	Loss
3D CNN	91.04%	0.91	0.90	0.43

3D CNN ResNet	94.02%	0.93	0.92	0.34
Inflated 3D ConvNet	86.07%	0.83	0.85	1.23
Video Vision Transformer	95.62%	0.94	0.93	0.24

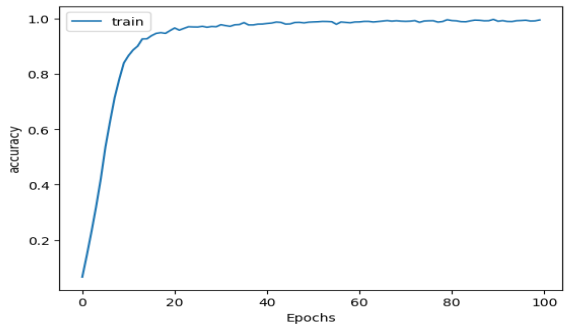

Fig. 5. Accuracy of 3D CNN Resnet.

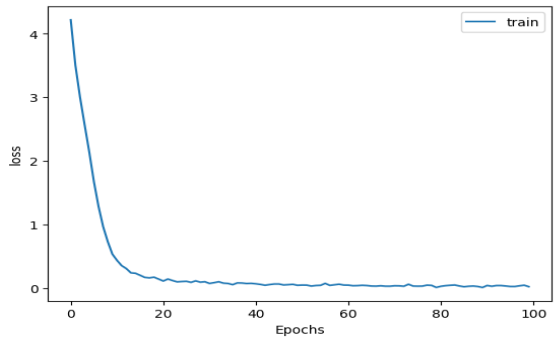

Fig. 6. Loss of 3D CNN Resnet.

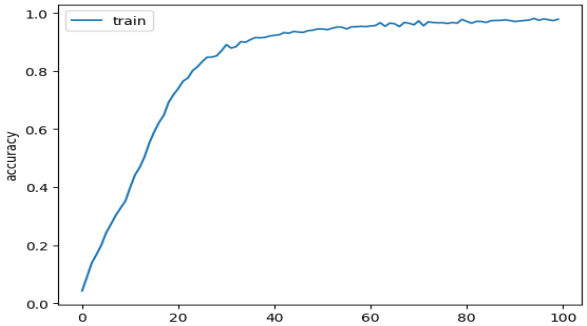

Fig. 7. Accuracy of I3D.

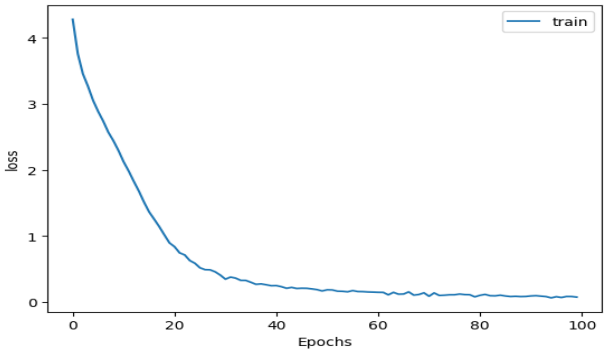

Fig. 8. Loss of I3D.

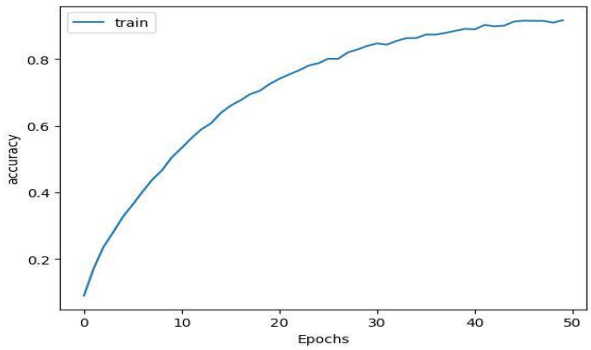

Fig. 9. Accuracy of 3D CNN .

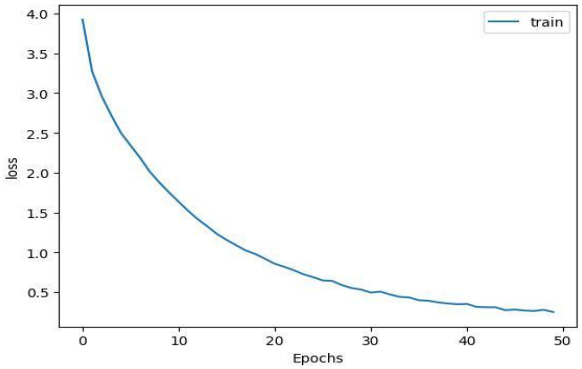

Fig. 10. Loss of 3D CNN.

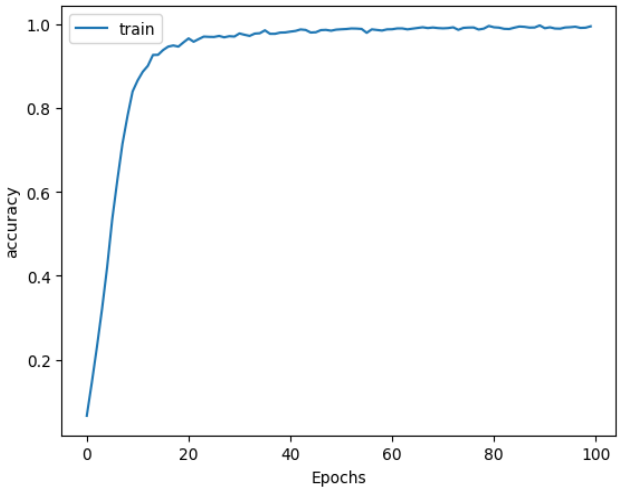

Fig. 11. Accuracy of ViViT.

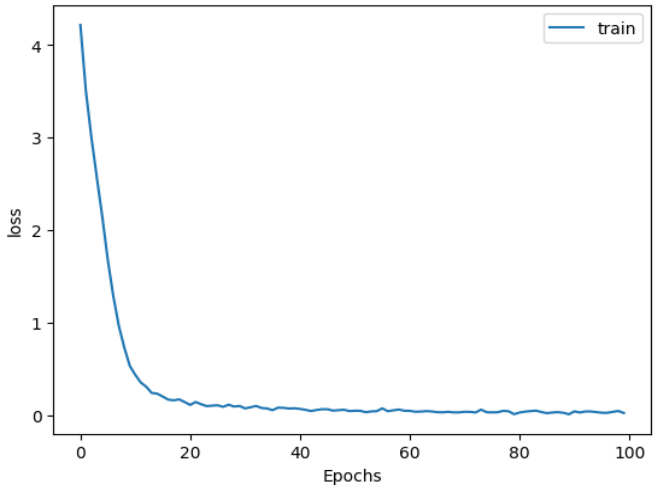

Fig. 12. Loss of ViViT.

5.1 Comparison of Models

The systematic evaluation provided valuable insights into the strengths and limitations of each model which are summarized in Table 2.

Table 2. Summary of Deep Learning Methods for HAR.

Method	Model Architecture	Accuracy	Scalability	Feature Extraction	Limitations
3D CNN (Proposed baseline)	3D Convolutional Layers with Max Pooling and Dropout layers.	Low Accuracy compared to Advanced Methods like Transformers and 3D CNN Resnet. Low Accuracy compared to Inflated 3D ConvNet.	Low to moderate	Efficient in capturing Spatiotemporal Features	Struggles with complex Tasks.
3D CNN Resnet [10]	3D CNN with Residual connection.	High Accuracy and Recall compared to 3D CNN and I3D Convnet.	Low to moderate.	Improved due to Residual connections	Computationally Expensive, Limited ability to capture long-range dependencies.
Inflated 3D ConvNet [15]	Leverages pretrained ImageNet models for Strong Feature initialization.	Low Accuracy compared to other methods.	Moderate	Improved Motion understanding due to the use of Dual-Stream input.	Computationally expensive due to dual input streams.
Video Vision Transformer [11]	Transformer	High Accuracy compared to above listed methods.	High	Efficient learning due to capability to capture long-range dependencies with self - attention.	High computational cost, large datasets for optimal performance, slower training.

6 Conclusion

This comparative study highlights a clear evolution from traditional models like 3D CNN to more advanced approaches such as Video Vision Transformer. While 3D CNN provides a strong baseline with relatively low complexity, it struggles to match the accuracy and recall of more sophisticated models like 3D CNN ResNet and I3D, which incorporate residual connections and pretrained features to improve learning. The Video Vision Transformer emerges as the most

effective model, achieving the best results in all key metrics, though it is computationally more expensive.

In conclusion, the progression from 3D CNN (base model) to ViViT (advanced model) demonstrates how incorporating residual connections, pretrained networks, and attention mechanisms can lead to significant performance gains, especially for complex tasks like human action recognition in video data. However, the trade-off between model performance and computational resources must be considered when choosing the optimal model for deployment.

References

[1] Surek, G.A.S.; Seman, L.O.; Stefenon, S.F.; Mariani, V.C.; Coelho, L.d.S. Video-Based Human Activity Recognition Using Deep Learning Approaches. Sensors 2023, 23, 6384.

[2] M. Babiker, O. O. Khalifa, K. K. Htike, A. Hassan and M. Zaharadeen , "Automated daily human activity recognition for video surveillance using neural network," 2017 IEEE 4th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA), Putrajaya, Malaysia, 2017, pp. 1- 5, doi: 10.1109/ICSIMA.2017.8312024.

[3] D. Wu, N. Sharma and M. Blumenstein," Recent advances in video-based human action recognition using deep learning: A review," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 2017, pp. 2865- 2872, doi: 10.1109/IJCNN.2017.7966210.

[4] Sharma, V., Gupta, M., Pandey, A. K., Mishra, D., & Kumar, A. (2022). A Review of Deep Learning-based Human Activity Recognition on Benchmark Video Datasets. Applied Artificial Intelligence, 36(1)

[5] R. Sathya, M. Mythili, S. Ananthi, R. Asitha, V. N. Vardhini and M. Shivaani," Intelligent Video Surveillance System for Real Time Effective Human Action Recognition using Deep Learning Techniques," 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2023, pp. 1826-1831, doi: 10.1109/ICACRS58579.2023.10404670.

[6] T. Ahmad, L. Jin, X. Zhang, S. Lai, G. Tang and L. Lin," Graph Convolutional Neural Network for Human Action Recognition: A Comprehensive Survey," in IEEE Transactions on Artificial Intelligence, vol. 2, no. 2, pp. 128-145, April 2021, doi: 10.1109/TAI.2021.3076974.

[7] R. Khurana and A. K. S. Kushwaha," Deep Learning Approaches for Human Activity Recognition in Video Surveillance - A Survey,"2018 First International Conference on Secure Cyber Computing and Communication (IC- SCCC), Jalandhar, India, 2018, pp. 542-544, doi: 10.1109/IC- SCCC.2018.8703295.

[8] A. Mihanpour, M. J. Rashti and S. E. Alavi," Human Action Recognition in Video Using DB-LSTM and ResNet," 2020 6th International Conference on Web Re- search (ICWR), Tehran, Iran, 2020, pp. 133-138, doi: 10.1109/ICWR49608.2020.9122304.

[9] D. R. R and P. C. J," Human Action Recognition in Videos Using Inception-v4 Deep Learning Model," 2024 Second Inter- national Conference on Advances in Information Technology (ICAIT), Chikkamagaluru, Karnataka, India, 2024, pp. 1-6, doi: 10.1109/ICAIT61638.2024.10690470.

[10] N. Archana and K. Hareesh," Real-time Human Activity Recognition Using ResNet and 3D Convolutional Neural Networks," 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS), Ernakulam, India, 2021, pp. 173-177, doi: 10.1109/ACCESS51619.2021.9563316.

[11] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucˇic´, and C. Schmid," ViViT: A Video Vision

Transformer," arXiv, 2021, doi: 10.48550/arXiv.2103.15691
[12] H. A. Ullah, S. Letchmunan, M. S. Zia, U. M. Butt, and F.H. Hassan,"Analysis of Deep Neural Networks for Human Activity Recognition in Videos—A Systematic Literature Re- view," IEEE Access, vol. 9, pp. 126366-126387, 2021, doi: 10.1109/ACCESS.2021.3110610.
[13] M. Mohamed Zaidi *et al.*," Suspicious Human Activity Recognition from Surveillance Videos Using Deep Learning," *IEEE Access*, vol. 12, pp. 105497-105510, 2024, doi: 10.1109/ACCESS.2024.3436653.
[14] Y. G. Sun *et al.*," Performance of End-to-End Model Based on Convolutional LSTM for Human Activity Recognition," *Journal of Web Engineering*, vol. 21, no. 5, pp. 1671-1689, July 2022, doi: 10.13052/jwe1540-9589.21512.
[15] Y. Huang, Y. Guo and C. Gao, "Efficient Parallel Inflated 3D Convolution Architecture for Action Recognition," in *IEEE Access*, vol. 8, pp. 45753-45765, 2020, doi: 10.1109/ACCESS.2020.2978223.