

# Comparative Analysis of Deep Learning Architectures for Human Action Recognition using the Stanford 40 Dataset

K. Harshawardhan <sup>1\*</sup>, Dr. M. Senthil Kumaran <sup>2</sup>

ORCID: <sup>1</sup>[0009-0008-8496-1605](https://orcid.org/0009-0008-8496-1605), <sup>2</sup>[0000-0001-5316-9983](https://orcid.org/0000-0001-5316-9983)

<sup>1</sup>Research Scholar, Department of Computer Science and Engineering,

<sup>2</sup>Associate Professor, Department of Computer Science and Engineering,

<sup>1 2</sup> Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya (SCSVMV) Deemed to be University, Kanchipuram, India

**Abstract.** Human Action Recognition (HAR) in still images is a well-established computer vision task with applications in sports, security, and human-computer interaction scenarios. Current trends in HAR research indicate that deep learning applications are proliferating, and a few publication hot spots are leading to systematic fair evaluations of the state of the art in convolutional neural networks (CNNs). This work conducts a systematic evaluation of seven popular CNN families, ResNet, Inception, MobileNet, DenseNet, VGGNet, EfficientNet, and EfficientNetV2, as well as representative instance variants within each family. The models are trained and evaluated on the Stanford-40 Human Action Recognition dataset using a common experimental setup, constrained to a limited training budget (three epochs) to allow comparability across a large number of model evaluations. The models are assessed utilizing a variety of accuracy, precision, recall, and F1 score metrics. The experiment results reveal model performance trends that are related to model family and model depth. Mid-range models, like ResNet-50 and DenseNet variants yield the best trade-offs between performance and resource consumption. Lightweight models perform the worst but justify the sacrifice in performance with training efficiency. The best overall model is EfficientNetV2-L, which achieves the best performance across all evaluated metrics. It achieves this performance through training-aware model design, improved compound model scaling, and fused MB-Conv blocks as well as greater input resolution, all of which enable effective learning even under the study's low training budget. In contrast with previous studies that advocate for highly hybridized specialized models, this study provides a framework for systematic evaluation of CNN families under the same training budget. Overall, this study provides realistic HAR baselines for images and informs the reader of relative model selection strategies and trade-offs that may be employed in budget-limited training contexts.

**Keywords**-Convolutional Neural Networks, Human Action Recognition, Performance Evaluation, Stanford 40 Actions Dataset, Image Classification, Deep Learning

## 1. Introduction

Human action recognition (HAR) in images is a challenging computer vision task because it lacks a temporal dimension. Image-based HAR thus needs to extract subtle signals and context from a single image. However, Applications like automated surveillance, sports analysis, and human-computer interaction greatly benefit from capability for recognizing human actions from still photos. Early HAR approaches relied on hand-crafted features, which inadequately captured the variations in images due to pose, appearance, scale, and lighting. By automating the hierarchical feature extraction process from pixel values, convolutional neural networks (CNNs) and deep learning significantly enhanced the capacity to identify actions in images.

While CNN-based HAR has been explored in depth, previous research has focused on a narrow selection of architectures over lengthy training processes, usually biased toward selecting large, resource-hungry models.

This research instead explores a resource-constrained training process (three epochs), focusing on systematically evaluating resource-limited CNN architecture families in resource-constrained situations typical of rapid prototyping. Six major families of CNN architectures (ResNet, InceptionNet, MobileNet, DenseNet, VGGNet, and EfficientNet, including successors) are evaluated in a variety of implementations in the Stanford 40 dataset for a thorough comparison of architectures.

State-of-the-art CNN architectures leverage many architectural features, from residual connections to multiscale feature processing and lightweight, parameter-efficient models. Image classifiers trained on the ImageNet dataset provide an off-the-shelf transfer learning technique that significantly improves convergence dynamics. Since deep learning models are the state of the art for image-based HAR, a comprehensive evaluation of CNN architectures in different families is justified.

This paper describes related work and its goals and contributions, datasets and experiments, evaluates the

\*Corresponding author: [rm24ce007@kanchiuniv.ac.in](mailto:rm24ce007@kanchiuniv.ac.in)

findings, analyses the findings, offers suggestions for future work, and concludes with the main findings.

## 1.1 Literature Review

Finding human actions from a single frame is the goal of HAR from Static Images, a fundamental problem in computer vision. The success of HAR systems is largely driven by the power of deep learning models for learning rich, discriminative features from visual data. This review traces the evolution of these models, from foundational Convolutional Neural Networks (CNNs) to the latest architectural and training paradigms from 2014–2025.

Foundational CNN Architectures of Feature Extraction.

Deep CNNs have played a significant function in development of image recognition activities, among which HAR belongs. Were the first to achieve good results with VGGNet, showing that deep, but architecturally simple,  $3 \times 3$  convolution stacks could be used to achieve high performance [1]. The groundbreaking ResNet architecture then took its place, including residual connections for addressing the challenges of disappearing gradients and enable the successful training of incredibly deep networks. [2].

Later studies were aimed at enhancing computational efficiency and multi-scale feature representation. Parallel convolutional filters of varying sizes in inception networks[3] were used to extract features at various scales on a block basis. Introduced DenseNet, which facilitated the flow of information and reused features by linking each layer to all other subsequent layers[4].

Mobile and embedded applications demanded architectures such as MobileNet.[5],[6]. They applied depthwise separable convolutions to reduce the costs of computing. Expanded upon this to develop EfficientNet, which established a precedent of expanding network depth, width, and resolution methodically in a balanced manner [7]. The line of work was continued further in EfficientNetV2[8], which ameliorated the architecture and training process to demonstrate faster training speeds and elevated performance. Although these models are the foundation of contemporary computer vision, there are few comparative experiments available in the literature concerning HAR, especially in light of constrained training scenarios.

Recent Architectural Innovations (2023–2025): Fusing Local and Global Context

Based on these canonical CNNs, new paradigms have been experimented with to extend performance by explicitly modelling context, human structure, and long-range dependencies. A thorough survey proves that there has been a strong shift towards transformers, multi-modal fusion, and deployment-oriented approaches[9].

- **Hybrid CNN-Transformer Models:** The combination of CNNs with Vision Transformers (ViTs) is one of the most popular developments. suggested ConViT, a hybrid module that affects a transformer's capacity to model global relationships and neural networks' capacity to extract local features[10]. Their good performance demonstrates the synergistic value of combining the two paradigms in single-frame HAR.

- **Pose-Assisted Networks:** Explicit representation of the human body structure is a very potent method and it presented a key-points-assisted network, which integrates human pose information with the image in a bottom-up manner[11]. By informing the network with structural knowledge, they were able to obtain higher accuracy, which highlights why domain-specific knowledge is valuable in the recognition pipeline.

- Researchers are currently working on attention mechanisms that are specifically tailored for human action recognition. One such example is the “Region-aware Image-based Human Action Retrieval with Transformers” (RIART) model[12], focusing on incorporating region specific information into the transformer architecture to enhance action retrieval performance. This design clearly integrates characteristics from person-anchored regions, adjacent contextual areas, and the worldwide image scene, enhancing its capacity to distinguish activities by comprehending the interaction between the subject and their surroundings. This design clearly integrates characteristics from person-anchored regions, adjacent contextual areas, and the worldwide image scene, enhancing its capacity to distinguish activities by comprehending the interaction between the subject and their surroundings.

Recent Training Strategies (2023–2024): Efficiency and Robustness

Besides the work on architectures, much effort has been put into making HAR models more practical and resilient, especially in resource-constrained environments.

- **Ensemble and Transfer Learning:** Ensemble techniques have been employed to make models more robust by essentially evading large-scale fine-tuning of the model. Recent studies have highlighted the benefits of ensemble transfer learning. In particular, [13] demonstrated that combining multiple pre-trained CNN backbones leads to improved model performance., provides a significant robustness gain and provides an effective solution in conditions with few computational resources.

- **Knowledge Distillation:** The other significant objective is the creation of small yet mighty models. The research focused on knowledge distillation, a technique for training a smaller ‘student’ model to mimic the predictions of a larger ‘teacher’ model”[14]. They

effectively minimised the differences in knowledge between CNN and transformer models, thereby creating light solutions with high performance, which can be deployed to edge devices.

Although many contributions of new architectures for HAR, no cross-family evaluations have been performed comprehensively and under the same training conditions to date. Most works perform tests on a few models at most or vary the training setups, which makes comparison difficult. This work does so by providing a systematic and fair evaluation of seven CNN families.

### 1.2 Objectives and Contributions

The “main aim of this study is to perform a thorough evaluation of advanced deep learning models for human action identification, employing Stanford 40 Actions Dataset. The specific purpose of this research is to compare and assess the performance of different families of CNN models within a unified experimental environment.

This paper's principal contributions are:

- A unified and controlled cross-family comparison of seven major CNN architectures, including ResNet, Inception, MobileNet, DenseNet, VGGNet, EfficientNet, and EfficientNetV2 and their representative variants under identical experimental conditions
- A comparison of the performance of these models in terms of accuracy and efficiency.
- A systematic evaluation under constrained training (three-epoch setting), enabling fair convergence-based comparison
- Insights into the performance vs. complexity of the” recognition models
- Experimental results that guide other researchers in choosing an architecture that best suits their needs for real-world action recognition.
- A consolidated performance–efficiency benchmark for image-based HAR on the Stanford-40 dataset.

### 1.3 Research Trend and Study Selection Analysis

This subsection analyses recent research trends in deep learning–based human action recognition and outlines the systematic study selection process adopted for this work. The analysis provides context for the comparisons made in the following sections and demonstrates the evolving research environment that justifies an all-architectures comparison.

#### 1.3.1 Study Selection Process

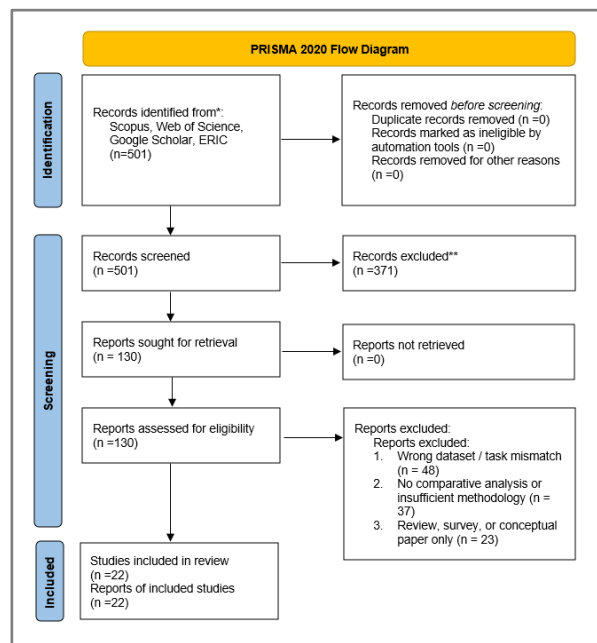


Fig. 1. PRISMA 2020 Flow Diagram

Figure 1 illustrates selection process as per PRISMA 2020. Searches from the academic databases yielded 501 records. In total, 22 studies were accepted on the basis of methodology and comparability. This rigorous process enhanced the current research with robustly designed, comparable studies. The literature processed in this subsection was gathered and processed according to PRISMA 2020 for systematic reviews and meta-analyses [15]

#### 1.3.2 Temporal Distribution of Publications

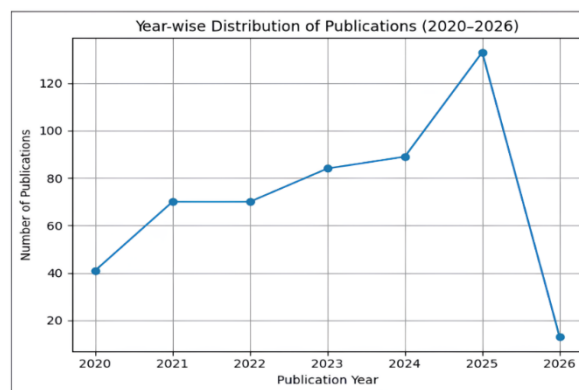


Fig. 2. Year-wise Distribution of Publications (2020–2026)

Figure 2 shows the yearly spread of publications on deep learning–based human action recognition. The spread of the yearly research volume from 2020 to 2024 and the immense spike in 2025 (along with the dip in numbers for 2026, which is not yet fully indexed) indicate this is an emerging field of study.

### 1.3.3 Geographical Distribution of Research Contributions

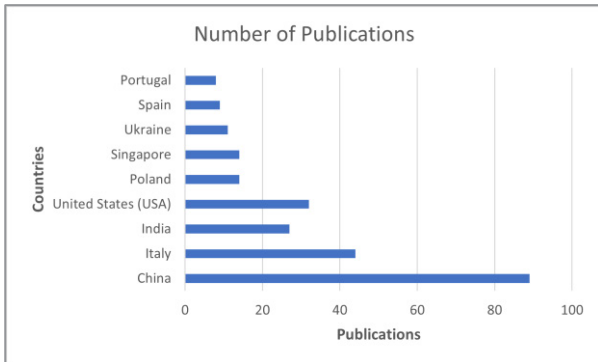


Figure 3. Country-wise Distribution of Publications

Figure 3 provides information on the country of publication of the papers. The majority of studies are from China, with Italy and the United States not far behind, and next to no publications from any of the other countries. This highlights the regional concentration of these large-scale experimental studies.

### 1.3.4 Journal-wise Distribution of Publications

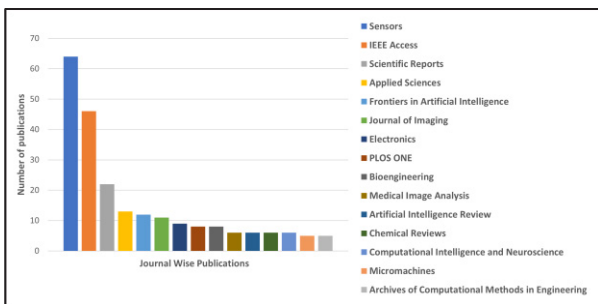


Figure 4. Journal-wise Distribution of Publications

The distribution of leading journals is illustrated in Figure 4. Sensors, IEEE Access, and Scientific Reports lead, showcasing the power of these outlets. There’s even representation of interdisciplinary journals, highlighting the cross-discipline of computer vision and deep learning.

### 1.3.5 Citation Impact Across Journals

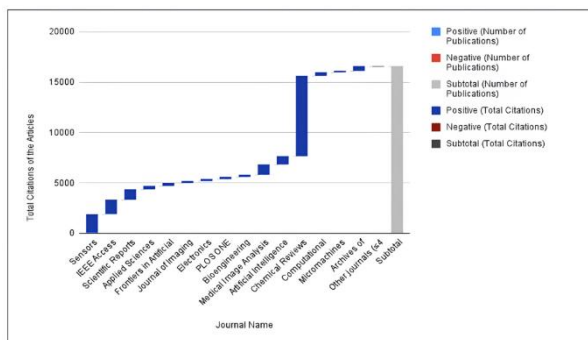


Figure 5. Citation Distribution Across Journals

Figure 5 depicts the distribution of citations received by the journals that were included in the analysis, showing that some journals are highly cited, indicating their significant impact. Moderately productive but highly cited journals are the focus of research. The distribution of citations received has a core-periphery topology.

## 2. Methodology

### 2.1 Dataset Description

This research employed a Stanford 40 Human Actions Dataset that included a total of 9,532 colour images that had been categorised into 40 different human action behaviours, including *applauding*, *climbing*, *rowing a boat*, and *waving hands*. The Stanford 40 Actions Dataset was collected from official Stanford Vision Lab repository. The dataset is publicly available and consists of real-world images depicting 40 different human action categories. All images were downloaded directly from the official source and used solely for research and academic purposes in accordance with the dataset’s usage guidelines.

| Aspect            | Details   |
|-------------------|---|
| Dataset Name      | Stanford 40 Actions Dataset   |
| Task Type         | Human action recognition (still images)   |
| Number of Classes | 40 human action categories  |
| Total Images      | 9,532 images  |
| Images per Class  | 180–300 (varies by class)   |
| Standard Split    | 100 images per class for training (4,000 total); remaining ~5,532 for testing   |
| Annotation Type   | Action labels + human bounding boxes  |
| Bounding Boxes    | Yes (for person performing action)  |
| Image Resolution  | Varies (real-world, unconstrained images)                                       |
| Data Modality     | RGB images  |
| Scene Type        | Complex, cluttered real-world backgrounds                                       |
| Original Authors  | B. Yao, X. Jiang, A. Khosla, A.L. Lin, L.J. Guibas, and L. Fei-Fei              |
| Original Paper    | “Human Action Recognition by Learning Bases of Action Attributes and Parts[16]” |
| Common Use Cases  | Action recognition, object–action interaction, vision benchmarks                |
| License / Usage   | Research and educational purposes   |
| Official Source   | Stanford Vision Lab   |

Table 1. Stanford 40 Actions Dataset details.[16]

## 2.2 Data Preprocessing

The images were all resized to 224x224 pixels and normalised utilizing ImageNet mean and standard deviation values to fit pretrained model. It has used data augmentation methods when training to provide generalization, such as small-angle rotations, random horizontal flips, etc., to prevent overfitting and improve model robustness.

For enhancing generalization and mitigate overfitting, the following data augmentation methods were employed during training:

- Random horizontal flipping
- Random rotation
- Random cropping and scaling

## 2.3 Models

The diverse range of Convolutional Neural Network (CNN) models that we tested included modern and classic models. Following families with their respective versions were included:

- **ResNet:** ResNet-18, ResNet-50, ResNet-101, ResNet-152
- **Inception:** Inception-v1, Inception-v2, Inception-v3, Inception-v4
- **MobileNet:** MobileNet-V1, MobileNet-V2, MobileNet-V3 Small, MobileNet-V3 Large
- **DenseNet:** DenseNet-121, DenseNet-169, DenseNet-201
- **VGGNet:** VGG-16, VGG-19
- **EfficientNet:** B0 through B7
- **EfficientNetV2:** V2-S, V2-M, V2-L

This choice enables the overall performance comparison of architectures with different depths, number of parameters, and computational complexities, letting us have visions on efficiency-accuracy and accuracy-scalability trade-offs in HAR tasks.

## 2.4 Experimental Setup

To evaluate the models fairly and systematically, the following experimental procedures were used for all the models:

- **Procedure:** All models used the same training procedure: all models trained for 3 epochs, no early stopping or weight adjustment after training weight (to ensure fixed comparison with limited training resources)
- **Framework:** Model built in TensorFlow 2.x Keras API to develop a modifiable/reproducible development framework

- **Learning:** Most models used ImageNet pretrained weights, some were trained from scratch to evaluate convergence/variability in results
- **Weight Tuning:** High parameter models EfficientNet-B6 and B7 were weight tuned with pretrained weights to see the impact of different weights
- **Evaluation:** Accuracy, Precision, Recall, F1 Score were determined on test set for overall and class specific performance
- **Hardware:** Google Colab with NVIDIA T4 GPU (16 GB VRAM). A reproducible and available computing resource
- **Optimizer:** Adam Optimizer using TensorFlow default settings for adaptive learning rates and compatibility with most model architectures

## 3. Result Analysis

The results of the CNN models on the Stanford-40 dataset are compared in this section. All models were trained for three epochs to observe the learning outcomes of identical experiments.

### 3.1 ResNet Family

| Model      | Accuracy      | Precision     | Recall        | F1-Score      |
|------------|---------------|---------------|---------------|---------------|
| ResNet-18  | 73.23%        | 70.22%        | 70.13%        | 69.67%        |
| ResNet-50  | <b>78.56%</b> | <b>77.16%</b> | <b>75.79%</b> | <b>75.58%</b> |
| ResNet-101 | 76.88%        | 76.26%        | 74.08%        | 73.67%        |
| ResNet-152 | 75.40%        | 76.43%        | 74.55%        | 73.02%        |

Table 2. ResNet variant results

The performance comparison of ResNet variants on the Stanford-40 dataset is summarised in Table 2 and illustrated in Figure 6. ResNet-50 achieved the best balance, with deeper variants showing diminishing returns under short training. Among the ResNets, ResNet-50 delivered the best results (accuracy 78.56%, F1-score 75.58%). Shallow ResNet-18 was worse (accuracy 73.23%), and neither deeper models ResNet-101 (accuracy 76.88%) nor ResNet-152 (accuracy 75.40%) helped either. ResNet-50, therefore, appears to be the optimal model with the right amount of depth, converging well with minimal training.

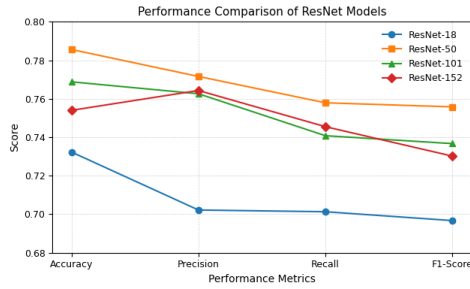


Figure 6. ResNet Metrics

### 3.2 Inception Family

| Model        | Accuracy      | Precision | Recall | F1-Score |
|--------------|---------------|-----------|--------|----------|
| Inception-v1 | 73.46%        | 70.33%    | 69.16% | 67.94%   |
| Inception-v3 | 81.69%        | 79.61%    | 79.00% | 79.01%   |
| Inception-v4 | <b>82.68%</b> | 73.94%    | 73.89% | 73.43%   |

Table 3. InceptionNet variant results

The comparative performance of Inception variants is summarised in Table 3 along with illustrated in Figure 7. Inception-v3 and v4 clearly outperform earlier variants. Inception showed consistent improvement across architectures. Inception-v3 yielded decent results at 81.69% accuracy and F1 of 79.01%, while Inception-v4 only marginally improved accuracy to 82.68% but at the cost of a balance between precision and recall (73.43% F1). Inception-v1 was much worse, highlighting the importance of model design for successful training.

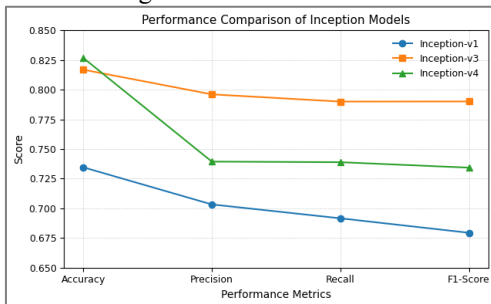


Figure 7. InceptionNet Metrics

### 3.3 MobileNet Family

| Model              | Accuracy | Precision | Recall | F1-Score |
|--------------------|----------|-----------|--------|----------|
| MobileNet V2       | 72.70%   | 71.28%    | 69.45% | 68.13%   |
| MobileNet V3-Small | 56.34%   | 51.71%    | 51.77% | 48.42%   |
| MobileNet V3-Large | 69.83%   | 66.25%    | 65.32% | 64.18%   |

Table 4. MobileNet variant results

The performance comparison of MobileNet variants is summarised in Table 4 and illustrated in Figure 8. MobileNetV3-Small significantly underperformed, likely due to its aggressive parameter reduction. MobileNet family models performed abysmally on the Stanford-40 dataset. MobileNetV2 gave the family's best finding: 72.70% (F1 68.13%). MobileNetV3-Large was not far behind, but MobileNetV3-Small suffered a dramatic performance crash (56.34% accuracy) due to drastic model compression. Such results show that lightweight models do not withstand short training and difficult problems.

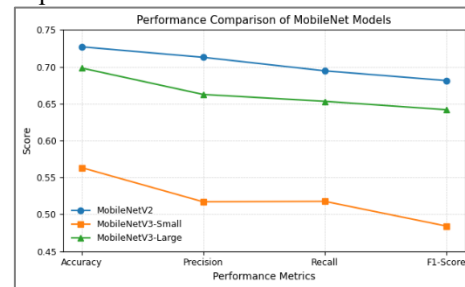


Figure 8. MobileNet Metrics

### 3.4 DenseNet Family

| Model        | Accuracy      | Precision     | Recall        | F1-Score      |
|--------------|---------------|---------------|---------------|---------------|
| DenseNet-121 | 76.95%        | 74.34%        | 73.76%        | 73.38%        |
| DenseNet-169 | 78.58%        | 77.52%        | 76.14%        | 75.83%        |
| DenseNet-201 | <b>80.82%</b> | <b>78.13%</b> | <b>78.35%</b> | <b>77.91%</b> |

Table 5. DenseNet variant results

The comparative performance of DenseNet variants is summarised in Table 5 and illustrated in Figure 9. DenseNet-201 had the best performance because of its dense connectivity. The DenseNet models all had similar gains in performance with regard to the number of layers, with DenseNet-201 achieving the best performance of 80.82% (F1 of 77.91%). Its dense connectivity allows for the retention of features and convergence in the challenging action recognition task.

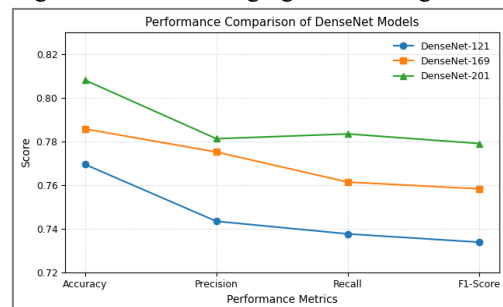


Figure 9. DenseNet Metrics

### 3.5 VGG Family

| Model  | Accuracy | Precision | Recall | F1-Score |
|--------|----------|-----------|--------|----------|
| VGG-16 | 68.89%   | 69.05%    | 66.39% | 65.38%   |
| VGG-19 | 69.70%   | 68.69%    | 66.75% | 65.71%   |

Table 6. VGG variant results

The performance comparison of VGG variants is summarised in Table 6 and illustrated in Figure 10. The VGG models exhibited low performance relative to state-of-the-art models under short training. VGG-16 and VGG-19 were comparable in accuracy (68.89% and 69.70%) but demonstrated low F1 scores ( $\approx 65\%$ ). Thus, these were ineffective models when generalizing to the test set under short training. These models likely lacked performance efficiency on short training sets due to missing features found in contemporary models, e.g. skip connections.

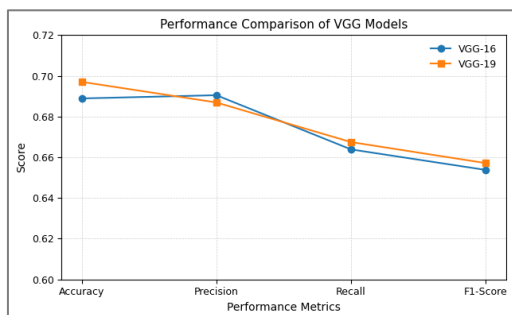


Figure 10. VGG-Net Metrics

### 3.6 EfficientNet Family

| Model       | Accuracy | Precision | Recall | F1-Score |
|-------------|----------|-----------|--------|----------|
| B0          | 70.77%   | 66.54%    | 67.08% | 66.29%   |
| B1          | 74.55%   | 71.03%    | 71.30% | 70.64%   |
| B2          | 72.27%   | 68.39%    | 68.76% | 68.22%   |
| B3          | 70.48%   | 67.09%    | 67.22% | 66.66%   |
| B4          | 64.62%   | 61.28%    | 60.61% | 60.34%   |
| B5          | 79.79%   | 77.42%    | 77.49% | 76.82%   |
| B6 (custom) | 78.78%   | 75.75%    | 76.45% | 75.62%   |
| B7 (custom) | 77.75%   | 74.67%    | 75.20% | 74.38%   |

Table 7. EfficientNet variant results

The comparative performance of EfficientNet variants is summarised in Table 7 and illustrated in Figure 11. Custom pretrained weights improved performance for deeper models too. EfficientNet models were scaling-family sensitive. EfficientNet-B5 was the highest performer (79.79% (F1 76.82%)). Deeper EfficientNet models (B6, B7) also saw gains from the custom-pretrained weights, but did not surpass B5. This confirms that scaling is central to efficiency, even with the combination of optimal transfer learning and limited training time.

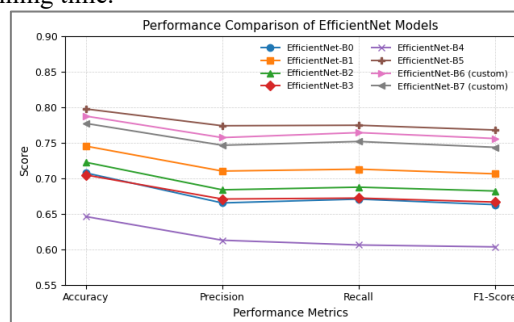


Figure 11. EfficientNet Metrics

### 3.7 EfficientNetV2 Family

| Model              | Accuracy | Precision | Recall | F1-Score |
|--------------------|----------|-----------|--------|----------|
| Efficient Net V2-S | 84.53%   | 83.45%    | 82.12% | 82.25%   |
| Efficient Net V2-M | 85.63%   | 85.27%    | 83.78% | 83.42%   |
| Efficient Net V2-L | 86.98%   | 85.94%    | 85.23% | 84.79%   |

Table 8. EfficientNet V2 variant results

The performance comparison of EfficientNetV2 variants is summarised in Table 8 and illustrated in Figure 12. EfficientNetV2 models were the best, with V2-L being the overall winner. EfficientNetV2 models were far superior to all other types of models. There was a "stacking" order, so to speak, from V2-S first, with 84.53% accuracy, then V2-M at 85.63%, and then the best of that model family, V2-L, the winner at 86.98% accuracy (F1 score 84.79%). These results show how much stronger the EfficientNetV2 family of models is for scaling.

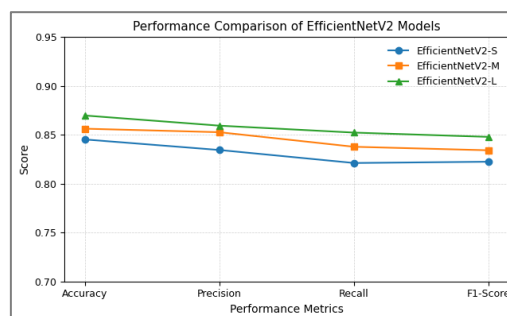


Figure 12 – EfficientNet-V2 Metrics

### 3.8 Overall Results

The cross-family evaluation shows different degrees of performances across architectures and depth. Compared to depth-sensitive architectures, moderate depth architectures like ResNet-50 and DenseNet families show consistent convergence and versatile metric performance. Depth does not enhance performance beyond a certain depth due to the limitations of training. EfficientNetV2-L displays best accuracy and F1 score with best precision-recall tradeoff, showing the best feature quality and convergence. Compared to earlier EfficientNet versions and baseline CNN architectures, EfficientNetV2-L shows markedly superior convergence for only a limited number of training epochs, highlighting the effectiveness of training-aware compound scaling. Overall, these findings indicate that depth is not as valuable as architectural efficiency and proper scaling for image-based human action recognition datasets like Stanford-40.

| Model Family    | Best Variant     | Accuracy (%) | Precision (%) | Recall (%)   | F1-Score (%) |
|-----------------|------------------|--------------|---------------|--------------|--------------|
| ResNet          | ResNet-50        | 78.56        | 77.16         | 75.79        | 75.58        |
| Inception       | Inception-v3     | 81.69        | 79.61         | 79.00        | 79.01        |
| Mobile Net      | MobileNet V2     | 72.70        | 71.28         | 69.45        | 68.13        |
| Dense Net       | DenseNet-201     | 80.82        | 78.13         | 78.35        | 77.91        |
| VGG Net         | VGG-19           | 69.70        | 68.69         | 66.75        | 65.71        |
| Efficient Net   | EfficientNet-B5  | 79.79        | 77.42         | 77.49        | 76.82        |
| Efficient NetV2 | EfficientNetV2-L | <b>86.98</b> | <b>85.94</b>  | <b>85.23</b> | <b>84.79</b> |

Table 9. Overall models results

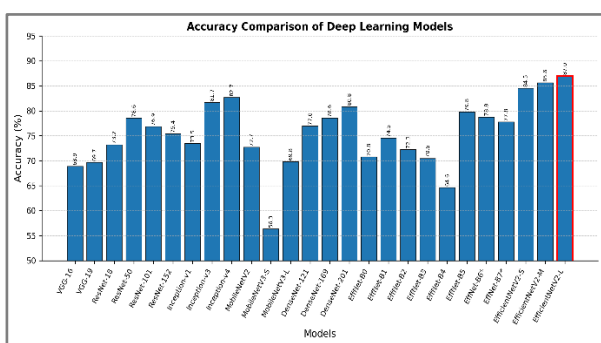


Figure 13. Accuracy Comparison

Among all the tested CNN architectures, EfficientNetV2-L achieved the highest performance, winning the most in accuracy and F1 score, thus being the overall best within the short three-epoch training duration.

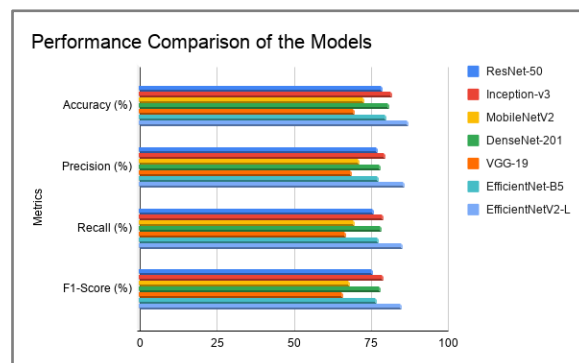


Figure 14 – Overall Performance Comparison

Figure 14 represents the performance comparison of selected best-performing models across accuracy, precision, recall, as well as F1-score. EfficientNetV2-L reliably achieves the highest values across all metrics, indicating strong and balanced classification performance. Other architectures show competitive accuracy but comparatively lower consistency across precision and recall.

### 4. Discussion

This comparative study corroborates and builds upon previous research in the domain of deep learning-based human action recognition. The consistent finding that increased depth does not provide a benefit, particularly in conditions of limited training data or limited epochs [2], holds true; ResNet-50 once again outperforms deeper versions of the model, such as ResNet-101 and ResNet-152. This finding adds support to the claim that models with moderate depth provide the best balance of representational power and optimization in non-ideal conditions [2].

The significance of transfer learning in this instance is also well established in the research literature. The difficulty of parameter-heavy models, including state-of-the-art variants of EfficientNet, to be effectively optimized when trained from scratch on moderately-sized datasets is widely documented [7], [8]. The improved performance of EfficientNet-B6 and EfficientNet-B7 when using weights pretrained on the ImageNet dataset supports claim that transfer learning increases the robustness of convergence and performance in deep learning models [8].

The capabilities of EfficientNet and its successor have been well established in previous research, notably in relation to compound scaling and advanced training techniques [7]. The finding of superior performance from EfficientNetV2, as established by this study, still holds true given that the training and architecture modifications allow effective training even at lowered epochs [8]. Its improved generalisation performance relative to outdated CNN models also holds true compared to previous benchmarking results [8]. The limitations of lightweight models also hold true with established findings. MobileNet, along with shallow versions of the VGG model, struggle with complex,

spatially-dense tasks such as human action recognition even with their efficiency [1], [5], [6]. Overall, then, the findings support the recommendations to practitioners to use models of medium size.

| Study  | Model / Method                   | Training Strategy                     | Reported Metric  |
|--|----------------------------------|---------------------------------------|--|
| The study of human action for scene [17]                                 | DenseNet-201 + WCO               | Transfer learning with optimization   | Accuracy = 94.7%                                       |
| The study on convolutional layer that uses Vision Transformer (ViT) [18] | ConViT (Hybrid CNN-Transformer)  | Data augmentation + extended training | mAP = 95.2% (ConViT), 95.5% (ConViT+HB)                |
| The study on Distilling Knowledge from CNN-Transformer [19]              | ConvNeXt (Teacher) / PVT-v2 (KD) | Knowledge distillation (CNN→ViT)      | Accuracy = 89.08% (Teacher), 88.59% (KD); mAP = 90.49% |
| Present Work   | EfficientNet V2-L                | Uniform 3-epoch constraint            | Accuracy = 86.98%                                      |

Table 10 Comparison with Recent State-of-the-Art Results on Stanford-40

Table 10 shows results for a comparison of the proposed benchmark to recent Stanford-40 studies, achieving high accuracy for DenseNet-201 with optimization in a transfer learning scenario[17]. Hybrid convolution-transformers such as ConViT achieve over 95% mAP with global attention and extended training [18]. Transformers with knowledge distillation, which take a CNN inductive bias on backbones, also improve Stanford-40 discriminative accuracy [19]. The present study, by contrast, evaluates multiple families of CNNs with strict three epoch training to enable family comparison. Absolute levels of accuracy are lower than with many tailored state-of-the-art approaches. The proposed framework, however, emphasizes convergence, efficiency, and structured benchmarking even in limited training scenarios.

• **Scalability and Robustness Considerations:**

Lightweight architectures are not robust for fine-grained action recognition. EfficientNetV2-L has better precision and recall, indicating improved robustness. Scaling and training-aware properties make it both scalable and efficient.

## 5. Future Research Directions

In order to verify model stability, future research should investigate the scalability of deep learning models trained on the more intricate video datasets HMDB51 and UCF101. The action recognition task’s performance can be improved through temporal modelling approaches, like 3D CNN, two-stream networks, and transformers. The models can also be made more efficient for real-time and edge usage by applying compression methods, like pruning, quantization and knowledge distillation to decrease inference costs. The need for large training datasets can be avoided using cross-domain transfer learning and self-supervised learning. Explainable AI methods can be applied to enhance users’ trust in model predictions.

## 6. Conclusion

To sum up, we provided a detailed benchmark of significant CNN models on the Stanford 40 Human Actions dataset with a similar and limited training schedule, providing an unambiguous and equal comparison of eight popular model families. Although DenseNet-201 and ResNet-50 have shown good and consistent results, EfficientNetV2-L has been the most accurate and has surpassed all the other models in terms of precision, recall, and F1-score, and has become the new standard of image-based HAR. Our findings underscore importance of the choice of architecture, pretrained initialization, and computational budget on model effectiveness, especially when resources are limited. Additionally, findings demonstrate the benefits of the modern and highly optimised designs, such as EfficientNetV2-L, and confirm the importance of the widely adopted architectures in particular applications. These observations have offered a practical guide to the system designers as well as a starting point for future exploration of the ability to balance accuracy, efficiency, and scalability of HAR to utilization in real world.

## References

- [1] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Apr. 10, 2015, *arXiv*: arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556
- [2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Dec. 10, 2015, *arXiv*: arXiv:1512.03385. doi: 10.48550/arXiv.1512.03385.
- [3] C. Szegedy *et al.*, “Going Deeper with Convolutions,” Sep. 17, 2014, *arXiv*: arXiv:1409.4842. doi: 10.48550/arXiv.1409.4842.
- [4] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” Jan. 28, 2018, *arXiv*: arXiv:1608.06993. doi: 10.48550/arXiv.1608.06993.
- [5] A. G. Howard *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision

- Applications,” Apr. 17, 2017, *arXiv*: arXiv:1704.04861. doi: 10.48550/arXiv.1704.04861.
- [6] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” Mar. 21, 2019, *arXiv*: arXiv:1801.04381. doi: 10.48550/arXiv.1801.04381.
- [7] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” Sep. 11, 2020, *arXiv*: arXiv:1905.11946. doi: 10.48550/arXiv.1905.11946.
- [8] M. Tan and Q. V. Le, “EfficientNetV2: Smaller Models and Faster Training”.
- [9] M. Kaseris, I. Kostavelis, and S. Malassiotis, “A Comprehensive Survey on Deep Learning Methods in Human Activity Recognition,” *Mach. Learn. Knowl. Extr.*, vol. 6, no. 2, pp. 842–876, Apr. 2024, doi: 10.3390/make6020040.
- [10] R. R. Dokkar, F. Chaieb, H. Drira, and A. Aberkane, “ConViViT -- A Deep Neural Network Combining Convolutions and Factorized Self-Attention for Human Activity Recognition,” Oct. 22, 2023, *arXiv*: arXiv:2310.14416. doi: 10.48550/arXiv.2310.14416.
- [11] X. Lu, H. Xing, C. Ye, X. Xie, and Z. Liu, “A key-points-assisted network with transfer learning for precision human action recognition in still images,” *Signal Image Video Process.*, vol. 18, no. 2, pp. 1561–1575, Mar. 2024, doi: 10.1007/s11760-023-02862-y.
- [12] Y. Zhang and Y. Wang, “A comprehensive survey on RGB-D-based human action recognition: algorithms, datasets, and popular applications,” *EURASIP J. Image Video Process.*, vol. 2025, no. 1, p. 15, Aug. 2025, doi: 10.1186/s13640-025-00677-0.
- [13] A. Dhatarwal, S. Ratnoo, A. Bajaj, and A. Abraham, “Ensemble Transfer Learning for Robust Human Activity Recognition from Images”.
- [14] C. Hu *et al.*, “Teacher-Student Architecture for Knowledge Distillation: A Survey,” Aug. 08, 2023, *arXiv*: arXiv:2308.04268. doi: 10.48550/arXiv.2308.04268.
- [15] M. J. Page *et al.*, “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews,” *BMJ*, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.
- [16] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, “Human action recognition by learning bases of action attributes and parts,” in *2011 International Conference on Computer Vision*, Barcelona, Spain: IEEE, Nov. 2011, pp. 1331–1338. doi: 10.1109/ICCV.2011.6126386.
- [17] R. Surendran, A. J., and J. D. Hemanth, “Recognition of human action for scene understanding using world cup optimization and transfer learning approach,” *PeerJ Comput. Sci.*, vol. 9, p. e1396, May 2023, doi: 10.7717/peerj-cs.1396.
- [18] S. R. Hosseini, S. Seyedin, and H. Taheri, “Human Action Recognition in Still Images Using ConViT,” Jan. 11, 2024, *arXiv*: arXiv:2307.08994. doi: 10.48550/arXiv.2307.08994.
- [19] H. Ahmadabadi, O. N. Manzari, and A. Ayatollahi, “Distilling Knowledge from CNN-Transformer Models for Enhanced Human Action Recognition,” Nov. 02, 2023, *arXiv*: arXiv:2311.01283. doi: 10.48550/arXiv.2311.01283.