Cloud Computing and Data Science
http://ojs.wiserpub.com/index.php/CCDS/

Research Article

A Comparative Study of Deep Learning Models for Human Activity Recognition

Mohammed Elnazer Abazar Elmamoon[1], **Ahmad Abubakar Mustapha**[2*]

[1]Department of Computer Science & Engineering, Vignan's Foundation for Science, Technology & Research (Deemed to be University), Vadlamudi, Guntur, A.P., India
[2]School of Computer Science and Engineering, VIT-AP University, Andhra Pradesh, India
E-mail: abubakar.22phd7073@vitap.ac.in

Received: 18 December 2024; **Revised:** 15 January 2025; **Accepted:** 17 January 2025

Abstract: Human Activity Recognition (HAR) is essential for real-time surveillance and security systems, enabling the detection and classification of human actions. This study evaluates five pre-trained Convolutional Neural Network (CNN) models, EfficientNetB7, DenseNet121, InceptionV3, MobileNetV2, and VGG19 on a dataset comprising 15 human activity classes. The models were compared based on accuracy, precision, recall, F1-score, loss, and Receiver Operating Characteristic Area Under the Curve (ROC AUC). InceptionV3 achieved the highest performance with a validation accuracy of 80.16%, precision of 80.20%, and ROC AUC of 0.81, demonstrating its effectiveness for HAR tasks. EfficientNetB7 and DenseNet121 also performed well, with ROC AUC scores of 0.74 and 0.80, respectively. VGG19, however, showed lower metrics, emphasizing its limitations for complex HAR applications. This work highlights the trade-offs between model performance and efficiency, offering guidance for selecting suitable architectures for real-time surveillance. The findings contribute to the optimization of HAR systems for applications in smart cities, healthcare, and security.

***Keywords*:** Human Activity Recognition (HAR), CNN, pretrained models, surveillance systems, performance evaluation

1. Introduction

Human Activity Recognition (HAR) has become a significant area of research, especially in the context of surveillance systems. HAR involves the automatic identification of physical activities based on data derived from various sensors [1], such as cameras [2], wearable devices [3], or other motion-capturing technologies [4]. In recent years, advancements in deep learning, particularly Convolutional Neural Networks (CNNs), have revolutionized the field, offering highly efficient and accurate methods for classifying human activities. CNN-based pretrained models, such as EfficientNet [5, 6], DenseNet [7], InceptionV3 [8], MobileNetV2 [9], and VGG19 [10], have emerged as state-of-the-art solutions for image classification tasks, demonstrating strong performance in various domains, including healthcare, security, and sports analytics. These models are particularly suited for surveillance applications where recognizing human actions in real-time can enhance monitoring, provide critical insights, and improve decision-making processes in security systems.

The application of HAR extends beyond surveillance and security. In healthcare, it plays a crucial role in

DOI: https://doi.org/10.37256/ccds.6120256264

monitoring patients' physical activities, helping with fall detection, elderly care, and rehabilitation. For example, in the domain of stroke rehabilitation, tracking the movements of patients during their recovery phase can provide valuable data for physiotherapists, ensuring that rehabilitation exercises are performed correctly and helping detect any irregularities in the movement patterns. In sports, HAR can be utilized to analyze players' movements for performance optimization [11]. Furthermore, the significance of HAR is heightened in modern smart cities where surveillance systems are essential for maintaining public safety, monitoring crowd behavior, or identifying suspicious activities.

One of the most pressing issues within HAR is the challenge of achieving high accuracy while maintaining low computational cost. Traditional machine learning models often fail to capture challenging patterns in human activities or require significant feature engineering. In contrast, deep learning models, particularly CNNs, have performed effectively in learning hierarchical features from raw images, making them suitable for HAR tasks. However, despite the success of deep learning in HAR, there remains a lack of comprehensive research comparing the performance of various CNN-based pre-trained models in the specific context of surveillance systems. The performance of these models can vary greatly based on factors such as model architecture, dataset complexity, and real-time processing constraints. Therefore, understanding how different models perform under varying conditions is vital for selecting the most suitable architecture for specific HAR applications, especially those in surveillance systems.

Despite significant advancements in human activity recognition, several challenges remain, particularly in the context of surveillance systems. The problem lies in identifying human activities in diverse and dynamic environments while maintaining high accuracy, real-time performance, and robustness. Surveillance systems, often deployed in uncontrolled environments, face issues such as varying lighting conditions, occlusions, diverse body postures, and complex interactions between individuals. These factors pose significant challenges to the effectiveness of HAR models, particularly when dealing with large-scale datasets and real-time processing requirements. Furthermore, while there are various CNN-based pretrained models available for image classification tasks, there is a lack of comprehensive studies that compare the performance of these models specifically for HAR tasks in surveillance applications. This gap in the literature makes it difficult to identify the most suitable model for real-world deployment, as performance metrics like accuracy, loss, computational efficiency, and generalization across diverse activity types are not always clearly understood.

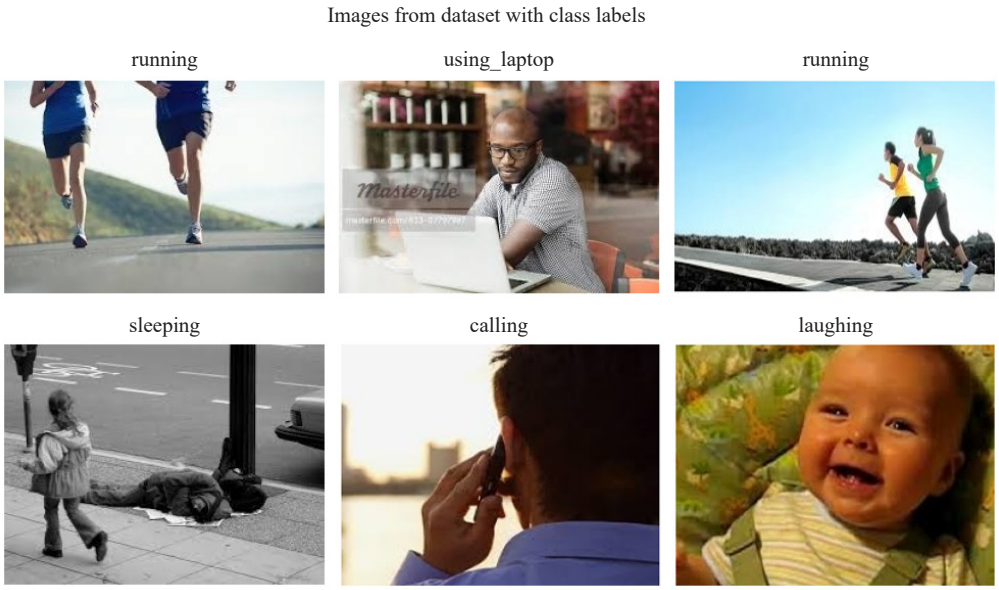


Figure 1. Sample of images from the datasets with their class labels

The primary aim of this research is to perform a comparative analysis of five CNN-based pretrained models

EfficientNetB7, DenseNet121, InceptionV3, MobileNetV2, and VGG19 on the task of human activity recognition within surveillance systems. This study aims to assess the performance of these models in terms of key metrics such as training accuracy, validation accuracy, training loss, and validation loss, using a publicly available Human Activity Recognition dataset, the sample of the images are shown in Figure 1. By evaluating the models' ability to generalize across various human activity classes, we aim to identify the strengths and limitations of each model in the context of real-time surveillance applications. Furthermore, the study seeks to explore how factors like model complexity, training time, and computational efficiency impact the deployment of HAR systems in practical scenarios. Ultimately, this research will contribute valuable insights into selecting the most suitable pretrained CNN model for enhancing the effectiveness and efficiency of surveillance systems, ensuring that they can operate optimally in diverse and dynamic environments.

This study addresses a critical gap in existing research by systematically comparing the performance of five state-of-the-art pre-trained CNN architectures, EfficientNetB7, DenseNet121, InceptionV3, MobileNetV2, and VGG19, for Human Activity Recognition (HAR). Unlike previous studies, which often focus on a single model or limited comparisons, this research uniquely explores the impact of model complexity, varying architectural depths, and the number of layers on performance, computational efficiency, and generalization ability. By leveraging a standardized dataset and consistent evaluation metrics, the study provides a detailed analysis of how these factors influence the models' ability to classify human actions accurately. The novelty lies in its practical insights into balancing computational demands with accuracy, offering actionable recommendations for selecting optimal models for HAR systems in real-time surveillance applications, with implications for smart cities, healthcare, and security domains.

2. Literature survey

HAR has been an active area of research due to its applications in surveillance systems, healthcare, and smart environments. Several approaches have been explored, ranging from traditional machine learning techniques to DL-based methods, particularly convolutional neural networks.

2.1 *Traditional approaches*

Earlier methods in HAR relied heavily on handcrafted feature extraction techniques combined with machine learning classifiers such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Hidden Markov Models (HMM). These approaches often required domain expertise for feature engineering and struggled with generalizing to complex and dynamic activities.

2.2 *Deep learning techniques*

The enhancement of Deep Learning (DL) has significantly transformed HAR by automating feature extraction and achieving higher accuracy. CNN-based models have become a popular choice due to their ability to learn spatial features directly from images. Pretrained models such as VGG, ResNet, and Inception have been extensively used in transfer learning settings for HAR tasks.

Table 1 displays the literature review table presents a comprehensive summary of recent advancements in Human Activity Recognition (HAR) using various machine learning and deep learning models across diverse datasets. Studies employing models like CNN, ResNet, EfficientNet, and hybrid architectures (e.g., Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) and Convolutional Neural Networks-Gated Recurrent Unit (CNN-GRU)) demonstrate significant progress in accuracy and computational efficiency. For instance, EfficientNetB3 achieved notable accuracy for vision-based tasks, while a CNN-GRU hybrid excelled on the Wireless Sensor Data Mining (WISDM) dataset by effectively combining feature extraction and temporal pattern recognition. Despite these successes, limitations such as model complexity, overfitting, lack of temporal feature extraction, and dataset-specific validations highlight the need for improved generalization and robustness. Additionally, emerging techniques like transformers and data augmentation have shown promise in enhancing performance but require further exploration under diverse and complex scenarios. This table collectively underscores the gaps and challenges in HAR research, paving the way for our study's focus on comparing state-of-the-art CNN models for robust and efficient HAR applications.

Table 1. Literature survey of models on HAR

Authors (ref.)	Year	Model	Dataset used	Observation	Limitation
Xu et al. [12]	2018	CNN	WISDM	The CNN-based method for human activity recognition achieved 91.97% accuracy using raw accelerometer data, outperforming the SVM's 82.27% while requiring lower computational cost and no complex preprocessing.	Finding an efficient method of recognizing the physical activities becomes the pivotal, core and urgent issue.
Archana and Hareesh [13]	2021	ResNet & CNN	Kinetics 400	The study developed a real-time human activity recognition method using a modified ResNet combined with 3D CNN, achieving enhanced recognition accuracy and reduced overfitting by leveraging the extensive Kinetics dataset.	Excludes LSTM-attention mechanisms, potentially limiting temporal feature extraction.
Gill et al. [14]	2023	EfficientNetB3	HAR	The study proposed a model using EfficientNetB3 and the Adam optimizer, achieving 76% accuracy for vision-based action prediction and recognition tasks.	Model complexity and efficiency.
Luo et al. [15]	2022	EfficientNet	Light Detection and Ranging (LiDAR)	The study proposed an improved EfficientNet model optimized for 3D LiDAR data, achieving a remarkable accuracy of 99.69%, with superior environmental robustness and minimal invasiveness compared to traditional methods.	Limited analysis of performance under diverse environmental conditions and noise levels.
Fukace et al. [16]	2023	KNN, LR, MLP, NB, RF, and SVM	USC-HAD	The study integrates multiple publicly available datasets for HAR to create a unified dataset and evaluates various machine learning algorithms, with Random Forest (RF) achieving the best performance (accuracy and F-score of 0.969).	Integration of datasets may introduce inconsistencies or biases due to differences in data collection methods.
Saidani et al. [17]	2023	Transformer model	WISDM, HAR, UCI	The proposed HAR system using data enhancement and a transformer model achieved high accuracy: 98.2% for PAMAP2 (12 activities), 98.6% for UCI HAR (6 activities), and 97.3% for WISDM (6 activities), outperforming baseline methods and effectively capturing low- and high-level information.	Although various studies utilized manual features to identify human activities and obtained good accuracy. Nonetheless, the performance of such features degraded in complex situations.
Choudhury and Badal [18]	2023	CNN-LSTM	Inbuilt smartphone sensor-based dataset	The proposed lightweight CNN-LSTM model achieved 98% accuracy on raw sensor data for six daily activities in an uncontrolled environment, outperforming conventional models with minimal preprocessing and optimized computational time.	No comparison with advanced deep learning models beyond conventional approaches.
Gupta [19]	2021	CNN-GRU	WISDM	The proposed CNN-GRU hybrid model achieved superior accuracy on the WISDM dataset, outperforming state-of-the-art models like Inception Time and DeepConvLSTM, leveraging its ability to combine convolutional feature extraction with gated recurrent units for temporal pattern recognition.	Validation was limited to the WISDM dataset, potentially affecting generalizability to other datasets.
Pavliuk et al. [20]	2023	CNN	University of California Irvine-Human Activities and Postural Transitions (UCI-HAPT)	Transfer learning on the UCI-HAPT dataset showed performance improvements, particularly on smaller subsets, but posed challenges on large datasets due to potential negative transfer.	Negative transfer was observed when applying the pre-trained model to large datasets.
Sai Ramesh et al. [21]	2024	CNN	HAR	The proposed system uses transfer learning with three different deep learning models to reduce training time and computational costs for HAR.	Requires significant computational resources and access to large datasets for optimal performance.

3. Methodology

In this section, we detail the methodology employed to perform the comparative analysis of Human Activity Recognition (HAR) using CNN-based pre-trained models within a surveillance system. Our study focuses on evaluating five widely used pre-trained models: EfficientNetB7, DenseNet121, InceptionV3, MobileNetV2, and VGG19 on Human Action Recognition (HAR) Dataset [22]. Each of these models has been selected for its unique architectural characteristics and effectiveness in various image classification tasks. The following outlines the specific models used and their application in the context of HAR and the methodology workflow is been illustrated in Figure 2.

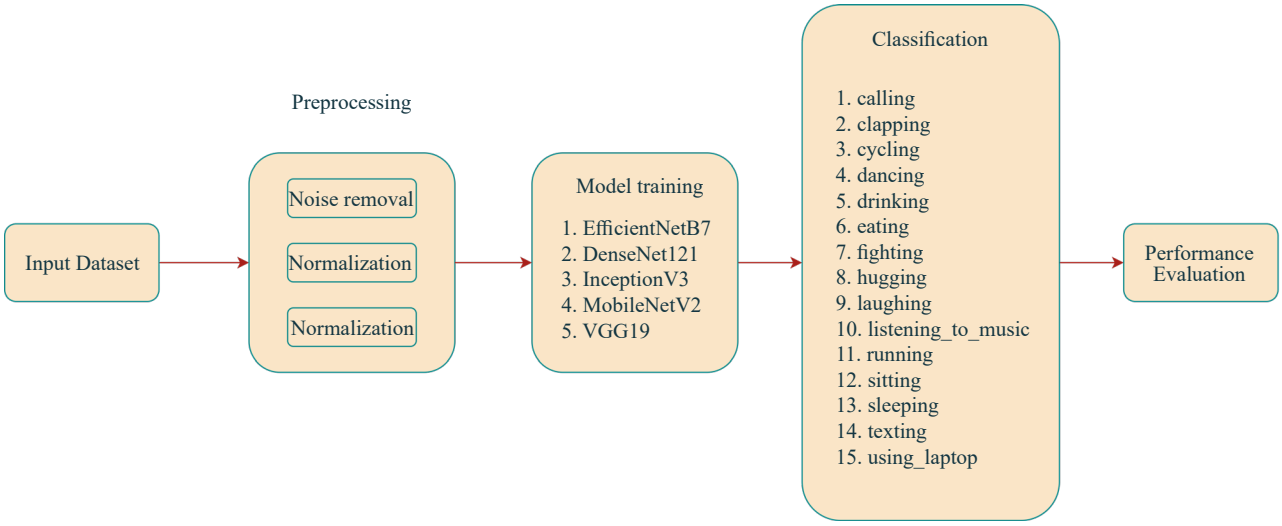


Figure 2. Workflow of this study

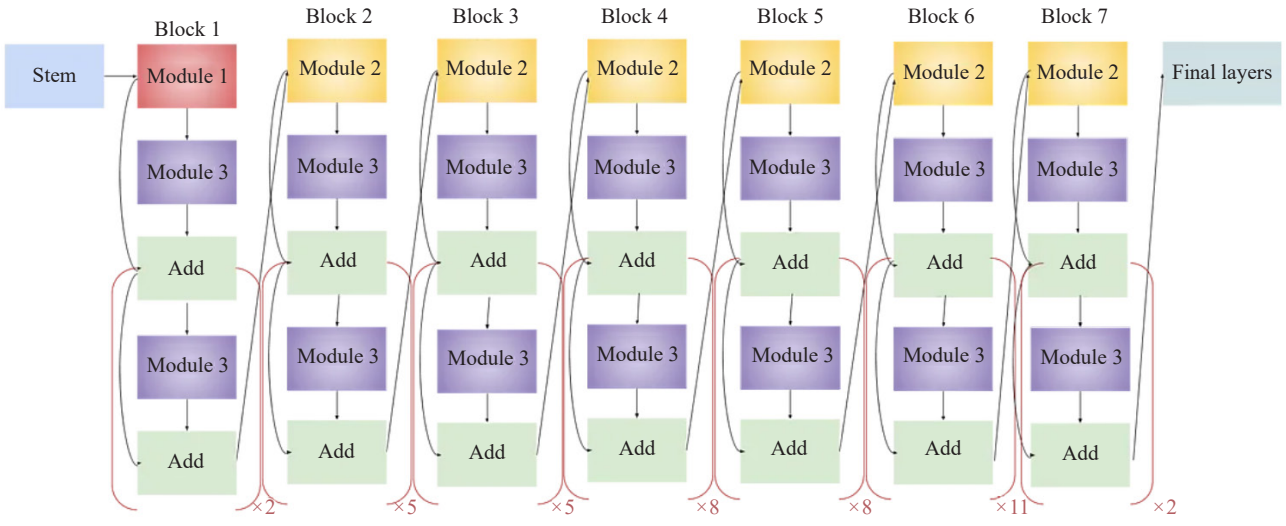


Figure 3. EfficientNetB7 Architecture [23]

3.1 *EfficientNetB7*

EfficientNetB7 is part of the EfficientNet family, which was introduced to optimize the balance between network depth, width, and resolution. The EfficientNet models, particularly B7, utilize a compound scaling method to achieve superior performance with fewer parameters compared to traditional CNN architectures. EfficientNetB7 is known for its high efficiency, where each layer has been scaled systematically to enhance both accuracy and performance as shown in the architecture in Figure 3. This model is ideal for the HAR task because of its ability to handle large datasets with high resolution while maintaining computational efficiency. We used this model's power by using it as a feature extractor for recognizing human activities in surveillance footage, benefiting from its state-of-the-art performance in image classification.

3.2 *DenseNet121*

DenseNet121 is a member of the Dense Convolutional Network (DenseNet) family, known for its novel approach to layer connectivity. Unlike traditional CNNs, DenseNet connects each layer to every other layer in a dense block. This results in more efficient parameter usage and improved gradient flow, addressing the vanishing gradient problem that can occur in deep networks. DenseNet121, with 121 layers, has proven to be highly effective in tasks requiring fine-grained visual recognition. In the context of HAR, this model is used to capture intricate features from human actions, especially in environments with varied lighting and backgrounds. DenseNet121 excels in utilizing all available features in the image by maintaining dense connections across layers, which is beneficial for detecting complex human activities. The architecture is depicted in Figure 4.

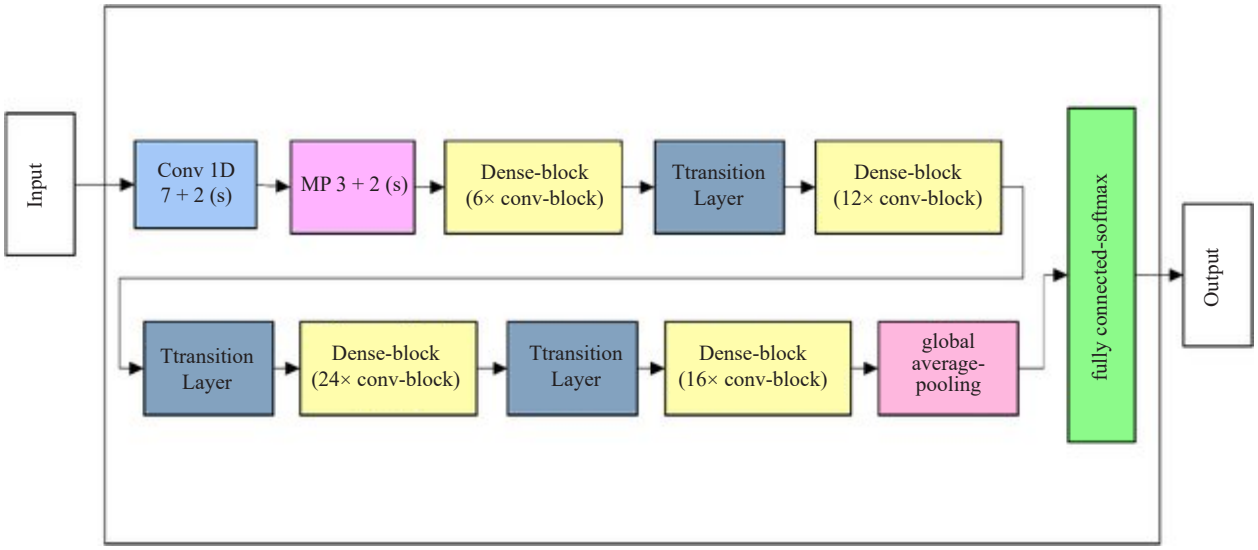


Figure 4. Architecture of DenseNet121 [24]

3.3 *InceptionV3*

InceptionV3 is a deep learning architecture developed by Google that employs the concept of "inception modules". These modules allow the network to perform convolutions of different sizes in parallel as the architecture in Figure 5, effectively capturing multi-scale information and improving the model's capacity to learn diverse patterns. InceptionV3 is designed to balance computational cost with model accuracy, making it a popular option for image recognition tasks. It incorporates advanced techniques such as batch normalization and auxiliary classifiers for better training efficiency and accuracy. For HAR, InceptionV3 is particularly useful due to its ability to capture complex human movements

across varying scales, making it well-suited for detecting human actions in dynamic surveillance environments.

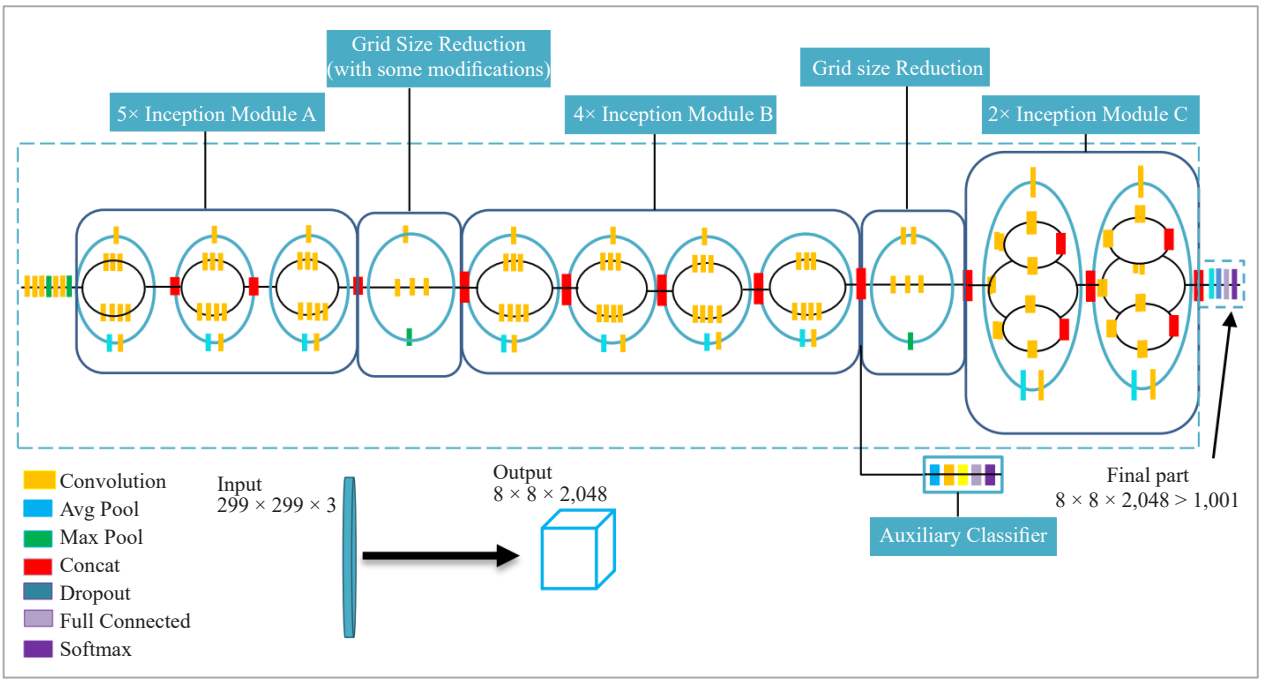


Figure 5. Architecture of IncetonV3 [25]

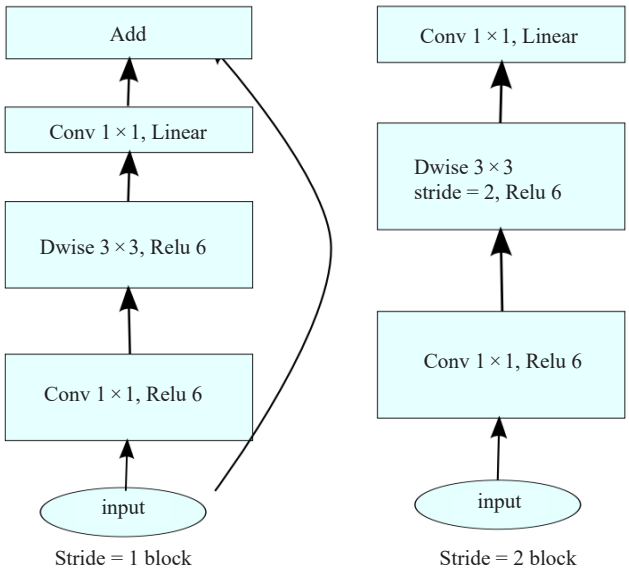


Figure 6. Convolutional Blocks of MobileNetV2 [26]

3.4 *MobileNetV2*

MobileNetV2 is a lightweight convolutional network designed for mobile and edge devices, focusing on reducing the computational cost without sacrificing too much performance. It uses depthwise separable convolutions as shown in

convolutional blocks of MobileNetV2 in Figure 6, which significantly reduce the number of parameters and operations needed, making it an efficient choice for real-time applications like human activity recognition in surveillance systems. MobileNetV2 introduces a novel linear bottleneck layer that improves feature representation and enhances model performance. While less computationally demanding, MobileNetV2 achieves competitive results in visual tasks, including human activity recognition. We used this model for scenarios where real-time processing and low-latency prediction are crucial, as it can run efficiently on devices with limited computational resources.

3.5 *VGG19*

VGG19 is a deep CNN architecture that was developed by the Visual Geometry Group (VGG) at the University of Oxford. It is a deep network with 19 layers, characterized by its simplicity and uniformity, with all convolutions being 3x3 filters, and max-pooling layers following every convolutional layer as shown in the architecture in Figure 7. VGG19's main strength lies in its straightforward architecture, which has been proven to perform well on a variety of image classification tasks, albeit with a higher computational cost compared to more efficient models like MobileNetV2. While VGG19 is computationally expensive, its depth allows it to capture detailed spatial hierarchies of image features. In the context of HAR, VGG19 is utilized to extract high-level features of human movements, enabling accurate classification of human activities despite its relatively larger parameter size.

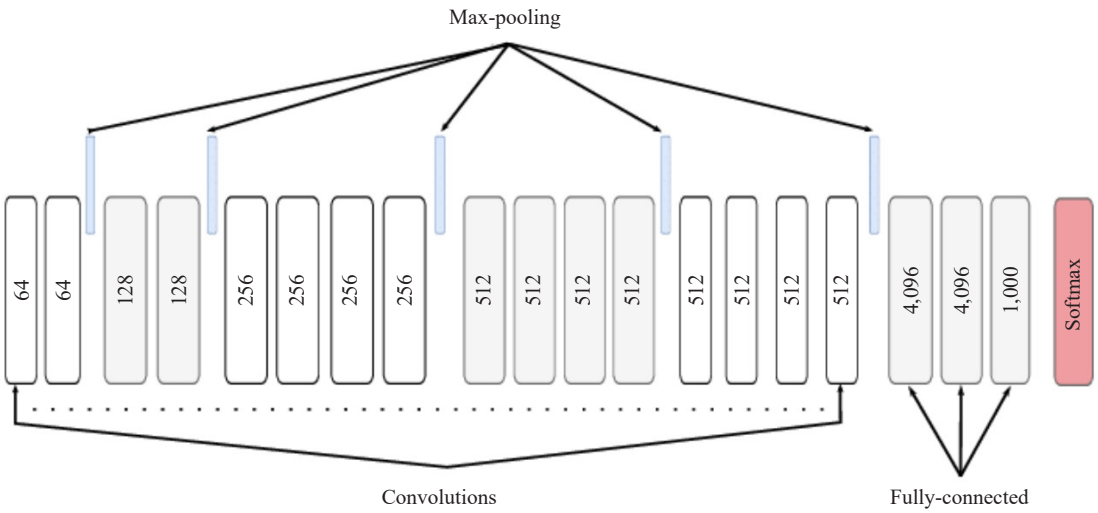


Figure 7. VGG-19 model architecture [27]

3.6 *Dataset overview*

The Human Action Recognition Dataset used in this study is a comprehensive collection of labeled images representing 15 different classes of human activities. These classes include diverse actions of human movements. The dataset is widely used in research related to activity recognition and behavior analysis due to its diverse and realistic scenarios, making it suitable for surveillance applications.

3.6.1 *Preprocessing*

Preprocessing included standardization and normalization of sensor readings, such as scaling accelerometer data to ensure consistent magnitude across different sensor types. Noise removal techniques were also applied to reduce artifacts, ensuring higher model performance.

Dataset Splitting

The original dataset was split into 80% : 20%. Post-augmentation, the dataset split remained the same, with the training data increased in size through augmentation techniques, ensuring a better model training phase. The augmentation techniques were applied to artificially expand the training data, preventing overfitting and enhancing generalization:

• Rotations: Applied to simulate different orientations of human activities and increase model robustness to changes in viewpoint.

• Zooms: Used to simulate changes in sensor distance, helping the model generalize well across variations in observation distances.

• Flipping: Applied to simulate mirrored movements, helping to balance the dataset and ensure the model's robustness to various activity directions.

The dataset consists of a total of 12,601 images, distributed across the 15 action classes. Each class contains approximately 840 images, ensuring a balanced dataset that facilitates unbiased training and evaluation of the models. Table 2 provides a detailed breakdown of the dataset distribution across the action classes.

Table 2. Dataset summary

Dataset split	Total images	Classes	Description
Train	7,201	'calling', 'clapping', 'cycling', 'dancing', 'drinking', 'eating', 'fighting', 'hugging', 'laughing', 'listening_to_music', 'running', 'sitting', 'sleeping', 'texting', 'using_laptop'.	Contains images for 15 activity classes, each containing label for training.
Test	5,400	Same as train.	Contains images for testing the model. Predictions are made for the same 15 class labels.

4. Results

4.1 *Model implementation and training*

For each of the five models, we used the pre-trained weights on ImageNet as the starting point and fine-tuned the models on the Human Activity Recognition dataset. The fine-tuning process involved replacing the top layers of each model with custom fully connected layers designed to predict one of the 15 action classes in the HAR dataset. The models were trained using categorical cross-entropy loss and optimized using the Adam optimizer with an initial learning rate of 0.0001. Data augmentation techniques, including random rotations, zooms, and horizontal flips, were applied to increase the robustness of the models. The training process was carried out for 20 epochs and callback was introduced to overcome overfitting with a batch size of 32, and the models' performance was evaluated on both training and validation datasets.

Table 3. Performance of the CNN-based models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Loss (%)	ROC AUC
EfficientNetB7	0.7361	0.74010	0.7398	0.7438	0.8868	0.74
DenseNet121	0.8000	**0.8011**	**0.8001**	**0.8000**	0.7869	0.80
InceptionV3	**0.8016**	**0.8020**	**0.8010**	**0.8000**	0.8125	**0.81**
MobileNetV2	0.7401	0.7401	0.7399	0.7398	1.1045	0.74
VGG19	0.7060	0.7101	0.7102	0.7100	1.1630	0.71

Table 3 presents the performance of five CNN architectures: EfficientNetB7, DenseNet121, InceptionV3, MobileNetV2, and VGG19, evaluated across accuracy, precision, recall, F1-score, loss, and ROC AUC. DenseNet121 and InceptionV3 exhibited the highest performance, achieving superior scores in accuracy, precision, recall, and F1-score, indicating their effectiveness in correctly classifying data while minimizing errors. In contrast, MobileNetV2 and VGG19 demonstrated lower performance across these metrics, suggesting potential limitations in their ability to effectively learn and generalize from the data.

Table 4. Results after introducing call back to overcome overfitting

Models	Training accuracy (%)	Testing accuracy (%)	Epoch stopped at
EfficientNetB7	0.8120	0.7361	13
DenseNet121	0.8855	**0.8000**	9
InceptionV3	0.8989	**0.8016**	**8**
MobileNetV2	0.8236	0.7401	11
VGG19	0.7998	0.7060	11

Table 4 illustrates the impact of a callback mechanism on the performance of five CNN architectures: EfficientNetB7, DenseNet121, InceptionV3, MobileNetV2, and VGG19. Overfitting, a common challenge in deep learning, occurs when a model excessively adapts to the training data, leading to poor generalization of unseen data. To mitigate this, a callback mechanism was employed, which dynamically monitors the model's performance on a validation set during training. When performance on the validation set begins to deteriorate, the callback mechanism interrupts the training process, preventing further overfitting. As expected, Table 4 demonstrates an increase in training accuracy for all models after the implementation of the callback. Crucially, the testing accuracy remained relatively stable for most models as visualized in Figure 8.

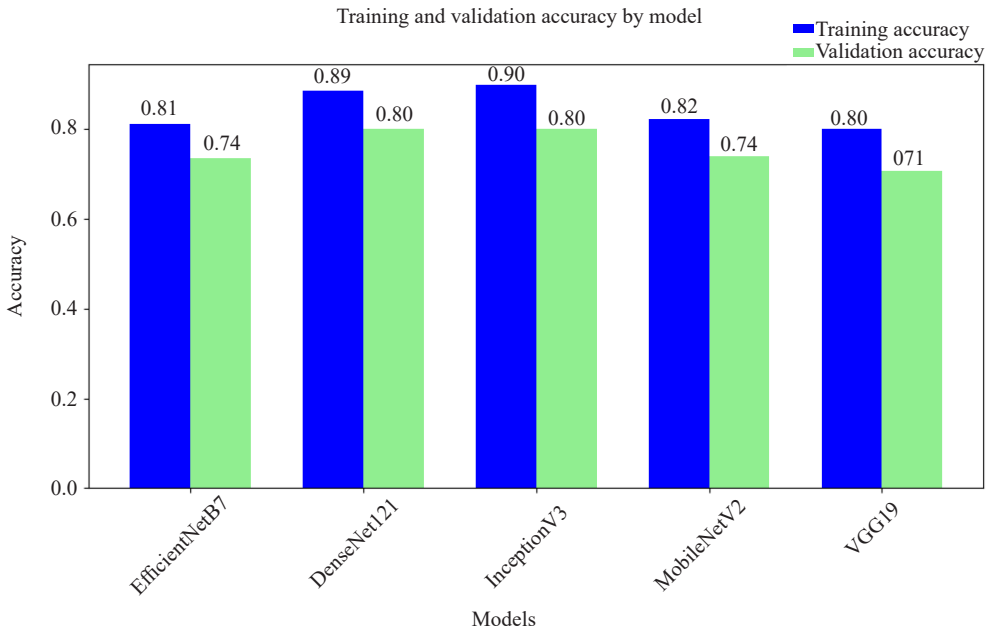


Figure 8. Training and validation accuracy of all models

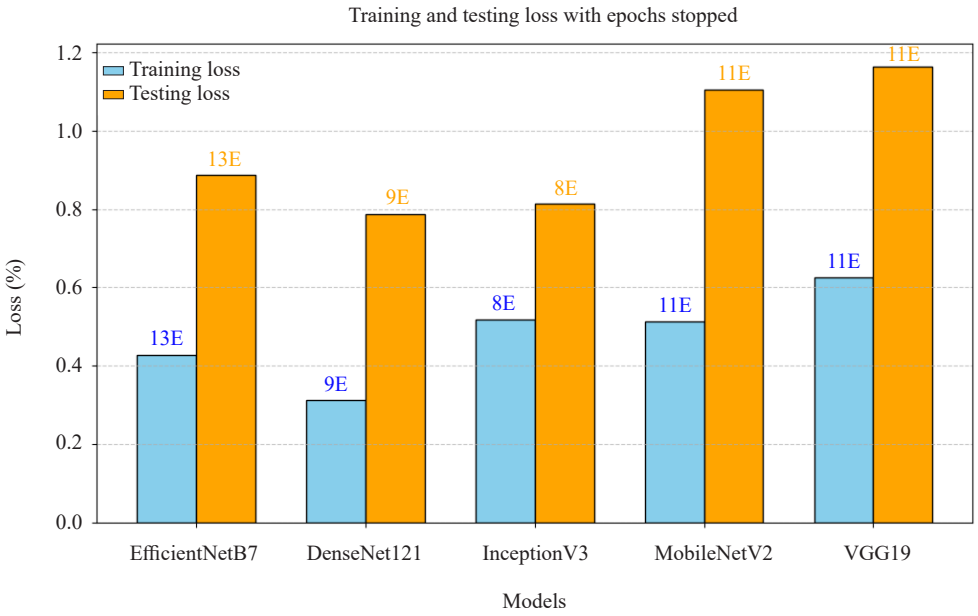


Figure 9. Training and validation loss of all models

Figure 9 illustrates the training and validation loss for all the CNN models in this study. By analyzing these losses, researchers can identify potential overfitting (increasing validation loss despite decreasing training loss), underfitting (high loss on both training and validation sets), and the optimal number of training epochs for each model.

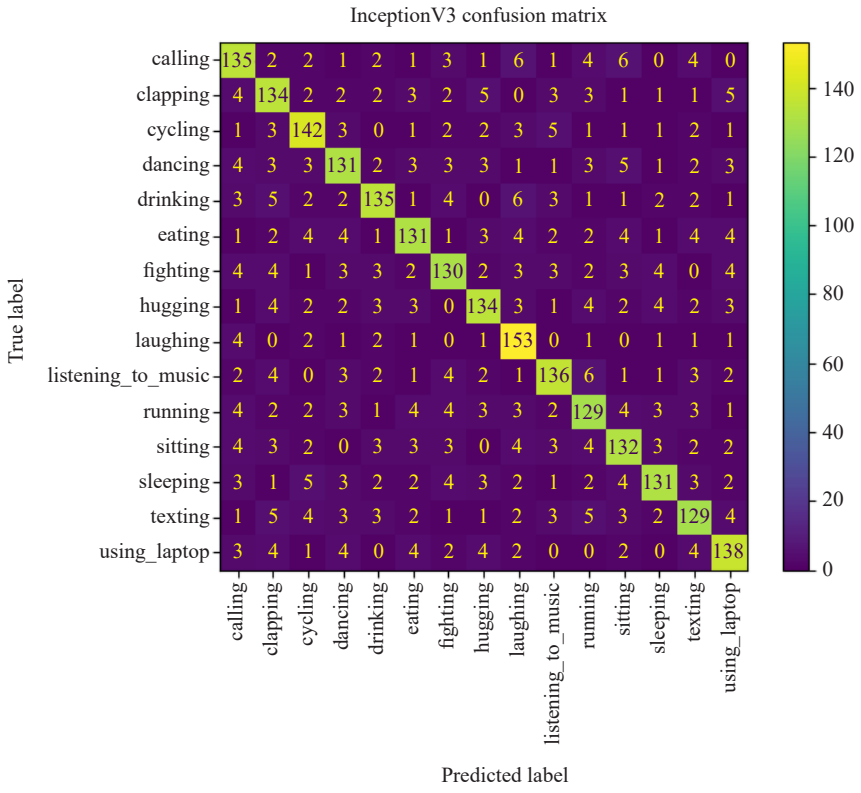


Figure 10. Confusion matrix of InceptionV3

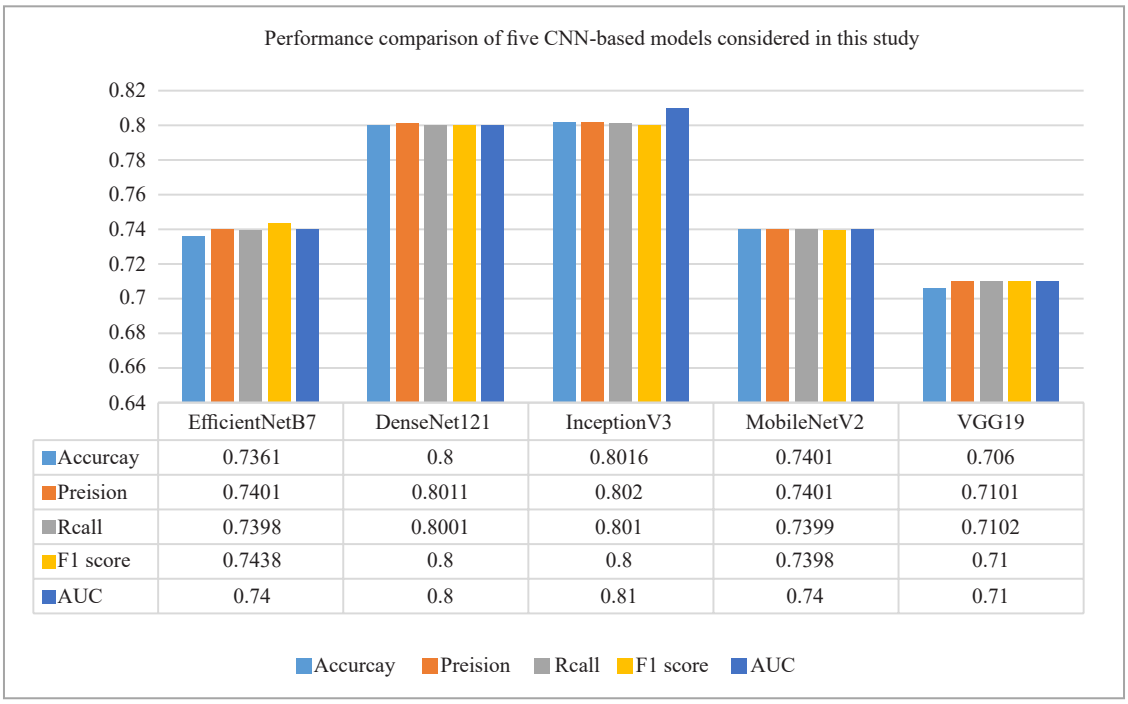
4.2 *Discussion*

The performance of the selected pre-trained Convolutional Neural Network (CNN) models for Human Activity Recognition (HAR) was comprehensively evaluated based on validation accuracy, loss metrics, precision, recall, F1-score, and ROC AUC. The results provide insightful conclusions regarding the suitability of these models for real-time HAR tasks and highlight the trade-offs involved in model selection.

4.2.1 *Performance comparison across models*

EfficientNetB7 exhibited strong performance with a validation accuracy of 73.61% and a ROC AUC of 0.74. Despite its relatively higher computational cost, the model demonstrated robust generalization capabilities, making it suitable for more complex HAR tasks. DenseNet121 and InceptionV3 achieved validation accuracies of 80.00% and 80.16%, respectively, with ROC AUC values of 0.80 and 0.81, showcasing their ability to balance efficiency and accuracy. These models are especially advantageous for applications requiring reliable activity recognition with moderate computational resources.

In contrast, MobileNetV2 and VGG19 demonstrated comparatively lower validation accuracies of 74.01% and 70.60%, respectively, and ROC AUC values of 0.74 and 0.71. The performance of these models underscores the challenges posed by less complex architectures when handling diverse activity classes, indicating limitations in feature extraction and discrimination. The confusion matrix for the top-performing model is shown in Figure 10, and Figure 11 shows the performance of every model.



	EfficientNetB7	DenseNet121	InceptionV3	MobileNetV2	VGG19
Accurcay	0.7361	0.8	0.8016	0.7401	0.706
Preision	0.7401	0.8011	0.802	0.7401	0.7101
Rcall	0.7398	0.8001	0.801	0.7399	0.7102
F1 score	0.7438	0.8	0.8	0.7398	0.71
AUC	0.74	0.8	0.81	0.74	0.71

Figure 11. Performance comparison of five CNN-based models considered in this study

4.2.2 *Model complexity vs. performance trade-offs*

EfficientNetB7 achieved superior loss metrics (0.89%) compared to the other models but required more epochs (13) for convergence. This suggests that while deeper architectures yield better generalization, they also come with higher training times and computational demands. Similarly, InceptionV3, with a loss metric of 0.81% across 8 epochs, emerged as a balance between accuracy and computational efficiency.

MobileNetV2 and VGG19 required 11 epochs each, with loss metrics of 1.10% and 1.16%, respectively, further emphasizing their suboptimal performance for HAR tasks. These findings illustrate that the computational simplicity of these models may not compensate for their inability to adequately classify intricate human activities.

4.2.3 *Implications for real-time applications*

The trade-offs between accuracy, computational efficiency, and loss highlight key considerations for deploying these models in real-time HAR systems. DenseNet121 and InceptionV3, with their high precision, recall, and ROC AUC, appear to be better candidates for applications such as surveillance and healthcare monitoring, where both accuracy and response time are critical. EfficientNetB7, despite its higher computational cost, is well-suited for environments where precision takes precedence over speed, such as detailed forensic analysis.

4.2.4 *Challenges and limitations*

The results reveal certain challenges in achieving optimal performance across all models. For instance, the relatively lower validation accuracy and loss metrics of VGG19 highlight its limitations in capturing complex patterns in HAR datasets. This could be attributed to its older architecture, which lacks the advancements of newer models like DenseNet and EfficientNet.

Moreover, the moderate ROC AUC values for all models indicate that while they perform well in distinguishing between activity classes, further optimization is required to achieve near-perfect classification. Incorporating techniques such as data augmentation, attention mechanisms, and transfer learning from more domain-specific datasets may improve their effectiveness.

5. Future directions

Future research could explore the integration of hybrid architectures, combining the strengths of EfficientNet and DenseNet, to enhance HAR performance. Additionally, extending the evaluation to real-world scenarios with more diverse datasets would provide a more comprehensive understanding of model performance. Incorporating temporal information using methods such as 3D CNNs or recurrent networks could further improve the robustness of HAR systems.

6. Conclusion

This study evaluates the performance of five pre-trained convolutional neural network models, EfficientNetB7, DenseNet121, InceptionV3, MobileNetV2, and VGG19, for Human Activity Recognition across 15 activity classes. Results reveal that InceptionV3 achieved the highest accuracy (80.16%), precision (80.20%), and ROC AUC (81%), demonstrating its robustness and suitability for activity classification tasks. DenseNet121 followed closely with comparable performance metrics, confirming its effectiveness for HAR applications. EfficientNetB7, despite achieving relatively lower accuracy (73.61%), exhibited balanced performance across metrics, indicating its potential in scenarios prioritizing computational efficiency and moderate accuracy. MobileNetV2 and VGG19, while exhibiting lower accuracies (74.01% and 70.60%, respectively), highlighted the limitations of lightweight and older architectures for handling complex HAR tasks. The findings emphasize the trade-offs between accuracy, computational complexity, and model selection for real-world applications. The insights from this study contribute to optimizing HAR systems, with implications for enhancing surveillance, healthcare monitoring, and smart city infrastructures. Future research can explore custom architectures and hybrid approaches to further improve classification accuracy and efficiency.

Availability of supporting data

Dataset used in this study is publicly available in [22].

Conflict of interest

The authors declare that there are no competing interests.

References

[1] Demrozi F, Pravadelli G, Bihorac A, Rashidi P. Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey. *IEEE Access*. 2020; 8: 210816-210836.

[2] Jang Y, Jeong I, Younesi Heravi M, Sarkar S, Shin H, Ahn Y. Multi-camera-based human activity recognition for human-robot collaboration in construction. *Sensors*. 2023; 23(15): 6997.

[3] Ramanujam E, Perumal T, Padmavathi S. Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review. *IEEE Sensors Journal*. 2021; 21(12): 13029-13040.

[4] Das MR, Ram R. A review on human motion capture. In: *Proceedings of the International Conference on Systems, Energy & Environment (ICSEE)*. India: Government College of Engineering Kannur; 2021.

[5] Chiranjeevi VR, Murugan BS, Dhanasekaran S, Senthil Pandi S. Human activity recognition using EfficientNet for wearable sensor data. In: *2024 International Conference on Communication, Computing and Internet of Things (IC3IoT)*. Chennai, India: IEEE; 2024. p.1-5.

[6] Beaulieu A, Thullier F, Bouchard K, Maître J, Gaboury S. Ultra-wideband data as input of a combined EfficientNet and LSTM architecture for human activity recognition. *Journal of Ambient Intelligence and Smart Environments*. 2022; 14(3): 157-172.

[7] Huillcen Baca HA, Gutierrez Caceres JC, de Luz Palomino Valdivia F. Efficiency in human actions recognition in video surveillance using 3D CNN and DenseNet. In: *Future of Information and Communication Conference*. Cham: Springer International Publishing; 2022. p.342-355.

[8] Sarah J, Danny AM, Deen JM. Performance enhancement of action recognition system using inception V3 model. In: *International Conference on Soft Computing and Pattern Recognition*. Cham: Springer International Publishing; 2021. p.3-22.

[9] Zhou XL, Tian J, Hao D. A lightweight network model for human activity classifiction based on pre-trained mobilenetv2. In: *IET Conference Proceedings CP779*. Stevenage, UK: The Institution of Engineering and Technology; 2020. p.1483-1487.

[10] Vaghela R, Labana D, Modi K. Efficient I3D-VGG19-based architecture for human activity recognition. *The Scientific Temper*. 2023; 14(04): 1185-1191.

[11] Host K, Ivašić-Kos M. An overview of human action recognition in sports based on computer vision. *Heliyon*. 2022; 8(6): e09633.

[12] Xu W, Pang Y, Yang Y, Liu Y. Human activity recognition based on convolutional neural network. In: *2018 24th International Conference on Pattern Recognition (ICPR)*. Beijing, China: IEEE; 2018. p.165-170.

[13] Archana N, Hareesh K. Real-time human activity recognition using ResNet and 3D convolutional neural networks. In: *2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*. Ernakulam, India: IEEE; 2021. p.173-177.

[14] Gill KS, Sharma A, Anand V, Sharma K, Gupta R. Human action detection using EfficientNetB3 model. In: *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*. Erode, India: IEEE; 2023. p.745-750.

[15] Luo CY, Cheng SY, Xu H, Li P. Human behavior recognition model based on improved EfficientNet. *Procedia Computer Science*. 2022; 199: 369-376.

[16] Fukace VK, Costa YMG, Natal IDP. Integrating multiple public datasets for human activity recognition using machine learning. In: *2023 30th International Conference on Systems, Signals and Image Processing (IWSSIP)*. Ohrid, North Macedonia: IEEE; 2023. p.1-5.

[17] Saidani O, Alsafyani M, Alroobaea R, Alturki N, Jahangir R, Jamel L. An efficient human activity recognition using hybrid features and transformer model. *IEEE Access.* 2023; 11: 101373-101386.

[18] Choudhury NA, Soni B. An efficient and lightweight deep learning model for human activity recognition on raw sensor data in uncontrolled environment. *IEEE Sensors Journal*. 2023; 23(20): 25579-25586.
[19 Gupta S. Deep learning based human activity recognition (HAR) using wearable sensor data. *International Journal of Information Management Data Insights.* 2021; 1(2): 100046.
[20] Pavliuk O, Mishchuk M, Strauss C. Transfer learning approach for human activity recognition based on continuous wavelet transform. *Algorithms*. 2023; 16(2): 77.
[21] SaiRamesh L, Dhanalakshmi B, Selvakumar K. Human activity recognition through images using a deep learning approach. *Research Square*. 2024. Available from: https://doi.org/10.21203/rs.3.rs-4443695/v1.
[22] *Human Action Recognition (HAR) Dataset*. kaggle. Available from: https://www.kaggle.com/datasets/meetnagadia/human-action-recognition-har-dataset [Accessed 10th January 2025].
[23] Agarwal V. *Complete Architectural Details of all EfficientNet Models*. 2020. Available from: https://towardsdatascience.com/complete-architectural-details-of-all-efficientnet-models-5fd5b736142 [Accessed 15th December 2024].
[24] Tareq I, Elbagoury BM, El-Regaily S, El-Horbaty El-SM. Analysis of ton-iot, unw-nb15, and edge-iiot datasets using dl in cybersecurity for iot. *Applied Sciences*. 2022; 12(19): 9572.
[25] Iparraguirre-Villanueva O, Guevara-Ponce V, Roque Paredes O, Sierra-Liñan F, Zapata-Paulini J, Cabanillas-Carbonell M. Convolutional neural networks with transfer learning for pneumonia detection. *International Journal of Advanced Computer Science and Applications*. 2022; 13(9): 544-551.
[26] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. Mobilenetv2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. San Francisco, CA, USA: IEEE; 2018. p.4510-4520.
[27] Mascarenhas S, Agarwal M. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. In: *2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON)*. Bengaluru, India: IEEE. 2021. p.96-99.