Intelligent Monitoring for Elderly Well-being: Deep Learning-Based Activity Recognition for Fall Detection

Areeb Adnan Khan
Habib University

Syed Muhammad Mustafa
Habib University

Haania Siddiqui
Habib University

Syed Talal Wasim
MBZUAI

Syed Nouman Hasany
Normandie Universite

Muhammad Farhan
Habib University

***Abstract*—The aging population poses unique challenges in terms of healthcare and well-being, requiring innovative solutions to ensure the safety and quality of life for elderly individuals. This research focuses on applying deep learning techniques for activity recognition in tracking and monitoring the daily activities of elderly people. The current literature on the matter suggests that state-of-the-art models find it difficult to accurately distinguish between falling and laying down. The study proposes a unique solution by dividing the dataset into local and global tags with local tags representing labels that are based on the information of that particular frame while global tags represent labels of a frame that are the same for the entire video. By using this dataset structure, we employ two CNNs: EfficientNet, ResNet, and a Vision Transformer (ViT-B16). The method proposed achieved promising results, with EfficientNet successfully distinguishing between falling and non-falling events with high accuracy. The paper also discusses the remaining two models, their shortcomings, and potential solutions for future work. In conclusion, this research contributes to the field of elderly care by showcasing the potential of deep learning in real-time activity recognition, where intelligent monitoring systems adapt to the specific needs of elderly individuals, promoting their well-being and independence while providing essential support for caregivers and healthcare providers.**

***Index Terms*—Deep Learning, Human Activity Recognition, Convolutional Neural Network, Elderly Care, Healthcare**

I. Introduction

Falls are a significant health problem among the elderly, with about 36 million falls occurring among elderly people each year, resulting in over 32,000 deaths and approximately 3 million elderly individuals receiving medical care in emergency departments annually due to fall-related injuries [1]. According to the Centers for Disease Control and Prevention [2], the predominant cause of hip fractures, accounting for over 95% of cases, is attributed to falls. It can also cause other injuries such as broken bones and head injuries. Therefore, this is an important area of research. The current work in fall detection uses wearable sensors and IoT-based [3] [4] [5], radar-based [6], multi-modal and purely vision based [7] [8] [9] [10]. A limitation of wearable sensors is the need for frequent charging [11]. This becomes particularly problematic for older adults who may find it difficult to monitor battery levels consistently and recharge devices regularly. Moreover, wearable sensors can be uncomfortable and may have various side effects. Consequently, considering the discomfort, side effects, and inconvenience of frequent charging, these traditional options may not be the most suitable for elderly individuals [12]. Multi-modal approaches in fall detection have some drawbacks, as noted in [13]. Firstly, handling multiple sources of data from subjects and environments requires the need for efficient methods of extracting features, and handling different data from multiple sources, making the fall detection system computationally demanding and challenging. Secondly, placing multiple sensors on the body and environment can result in higher costs, discomfort (especially for the elderly), and issues related to deployment and implementation in real-world settings. In addition to that, the use of IoT-based solutions is computationally expensive because they require hardware. In light of this, non-intrusive vision-based sensors emerge as an appealing alternative for monitoring purposes. In purely vision-based solutions, machine learning-based techniques such as Support Vector Machines [14], Random Forests [15], Hidden Markov Models [16], and Gaussian Mixture Models [17] are used. However, most of the state-of-the-art work for fall detection are using Convolutional Neural Networks (CNNs) - both 2D and 3D CNNs, Recurrent Neural Networks such as Long Term Short Memory (LSTMs), and Autoencoders [18] which fall under the umbrella of deep learning. The existing body of work in this domain has employed images or videos as data inputs for fall detection. In a vision-based video classification system, the initial step involves preprocessing the video and extracting individual frames. Subsequently, feature extraction is conducted on each frame, followed by inference using a classifier to determine whether it corresponds to a fall or not. By analyzing frames, capturing spatial and temporal information, and leveraging classification models, the pipeline efficiently categorizes videos based on their content. This process empowers a range of applications including video surveillance, action recognition, and video

MBZUAI: Mohamed bin Zayed University of Artificial Intelligence

recommendation systems. However, video-based classification methods, although efficient, are also computationally expensive. For daily life activities, having such resources is inconvenient. Hence, there is a need for cheap and efficient solutions for detecting falls.

Our research focuses on how to differentiate between falling and non-falling events using image models that are cost-effective, scalable, and can be deployed in real-time. Our objective is to investigate the effectiveness of a computationally efficient approach that utilizes still images for fall detection. Instead of processing entire videos, we will concentrate on analyzing *individual frames*. This study aims to determine the viability of a vision-based system that relies solely on still images to detect falls. The approach can be refined and implemented to enable *real-time fall detection*. By leveraging the use of still images, it has the potential to provide a practical solution for detecting falls in real-time scenarios without significant delays or processing constraints. We used two CNNs, ResNet [19] and EfficientNet [20] and a Vision Transformer ViT [21]. We train the models on individual frames extracted from a video dataset. We trained the models twice using two distinct labels or tags, 'Local' and 'Global,' derived from the same dataset. The 'Local' label is assigned to each frame according to the contents of that particular frame. This means if a subject is standing in a frame that belongs to a falling video. The frame would be labeled as standing. Global tags refer to labels of a frame that are the same for an entire sequence/video. For instance, if the video shows a subject falling, then even in the frames where the subject has not started falling and is standing up will be labeled as falling.

Our research contributes by:

- Introducing a unique approach to fall detection using two distinct labels (*global* and *local*).
- Conducting a comprehensive comparative analysis of the performance of state-of-art architectures for fall detection using video frames
- Exploring the possibility of classifying fall events accurately based on a single video frame, addressing potential real-time application scenarios.

II. Literature Review

Wan et al. [22] conducted a comparative analysis of five machine-learning models concerning human activity recognition: CNN, LSTM, BLSTM, MLP, and SVM. According to the study, CNN, a conventional neural network technique, is very useful for categorizing and identifying human behavior. The five models tested in this study used two datasets on human behavior, and the effectiveness of these models was evaluated by a variety of assessment markers. This paper still has some flaws as the authors suggested that the remaining four neural networks can still be further optimized, and more detailed comparison experiments can be conducted. Kothandapani et al. [23] aimed at recognizing and assisting human behaviors for elderly people using input surveillance cameras. The CNN network proposed by the authors, classified images with an accuracy of 79.94%. When compared to other models that have been pre-trained, the experimental portion shows that their model achieves respectable image categorization accuracy. Deep et al. [24] employed a VGG model to predict human activities from the Weizmann Dataset [21]. They used transfer learning and trained machine learning classifiers to obtain important features. The accuracy of VGG-16 with the deployment of transfer learning was 96.95%. In terms of feature extraction, results showed that VGG-16 outperformed other CNN models. Jalal et al. [25] describe a novel depth video-based HAR technique that employs robust multi-features and embedded Hidden Markov Models (HMMs) to recognize the daily life activities of elderly people. In terms of recognition performance, their results on three difficult depth datasets show that their features outperform state-of-the-art feature extraction techniques. The proposed system can be employed in any e-health monitoring system, such as monitoring healthcare issues for the elderly and sick, or investigating people's indoor activities at home or in hospitals. Kim et al. [26] suggest a powerful depth video-based Human Activity Recognition system for tracking older people's activities. The work tracks human silhouettes using depth maps and then creates a human skeleton out of each silhouette using 23 primary and secondary body joints. The nine tasks for which the HMM model is trained are: walking, eating, exercising, cooking, sitting down, standing up, stooping, reclined watching, and lying down. According to experimental data, the average recognition rate for nine daily routine activities was 84.33%. Ahad et al. [27] provided a summary of the obstacles and difficulties encountered in human activity recognition and suggested some solutions. Their survey paper looked at some relevant benchmark datasets, such as gestures, medical activities, sports and exercise movements, 3D actions, and so on. In addition, this paper compares previous research on specific benchmark datasets relevant to this field of study. The paper concluded that researchers' models should be sufficiently effective with quicker real-time processing. Care should be used while making the trade-off between processing speed and effectiveness. Adhikari et al. [28] present a video-based fall detection system in an indoor environment using a convolution neural network. This paper uses Convolutional Neural Networks (CNN) to recognize different poses. Using Kinect, the following image combinations are explored: RGB, Depth, RGB-D, and background subtracted RGB-D. The author also mentions why is it difficult to distinguish falling from laying down. First of all, datasets of falling or laying are very difficult to find, since most of the research in this field is privately funded. The other was the characteristics of the fall that they represent a sequential change in pose, for example, a sequence of poses that ends in laying can be considered as a fall or non-fall event depending upon how fast the action has occurred. The paper presented a solution based on a CNN model but was only able to achieve 74% accuracy. The paper also noted that the model was confused between bending as to whether it is a fall event or a non-fall event and was unable to classify it properly. In a more recent paper Ramirez at el. [29] proposed a solution for fall detection by using human skeleton estimation

TABLE I: Activities in UP-FALL Dataset

Activity ID	Description	Duration(s)
1	Falling forward using hands	10
2	Falling forward using knees	10
3	Falling backwards	10
4	Falling sideward	10
5	Falling sitting in empty chair	60
6	Walking	60
7	Standing	60
8	Sitting	60
9	Picking up an object	10
10	Jumping	30
11	Laying	60

for feature extraction. Their model was evaluated on the UP-FALL dataset. The paper claimed good accuracy but stated that the model was confused about the laying activity. It was not able to differentiate between falling and laying. Zin et al. [30] proposed a model for elderly action recognition that employs a stereo camera. The actions recognized in the experiment are "Outside the room", "Transition", "Seated in a wheelchair", "Sitting on the bed", "Laying in bed", and "Falling." The model showed some promise, but it misclassified "Laying in bed" as "Sitting in bed." The paper also proposed a solution for it, but it was reported to be time-consuming. When it comes to detecting elderly falls, time is of the essence, so a time-consuming solution is undesirable.

The analysis from the literature review highlights that there is a need for an efficient solution that is real-time, fast, and user-friendly, particularly in the context of elderly individuals.

III. Methodology

A. Method

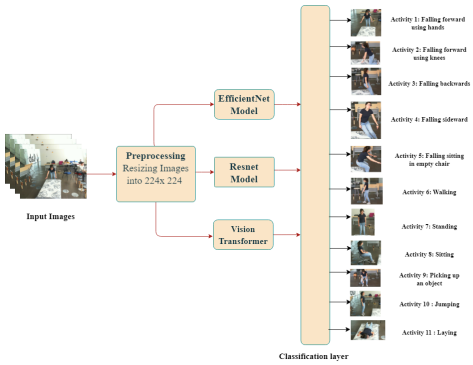


Fig. 1: Overview of the model

The methodology of the paper is summarized in a block diagram as shown in Figure 1.

B. Dataset

We have used UP-FALL dataset that contains both local and global labels for activities. The list of activities is provided in Table I.

C. Preprocessing

We conducted our research on 5 subjects out of 14 available in the UP-FALL dataset. We randomly chose subjects 9,10, and 12 for training and 1, and 3 for testing. The size of the train and test datasets was 54,000 and 33,000 images respectively. The images were cropped to 224 $\times$ 224 before being fed into the model.

D. Models

The research was conducted using two pre-trained Convolution Neural Networks-ResNet-18 [19] and EfficientNet-B0 [20] and a pre-trained Vision Transformer-ViT-B-16 [31].

*1) **EfficientNet-B0**:* EfficientNets are a family of models that uses neural architectural search and achieved better accuracy and used fewer parameters than ConvNets [32].The baseline architecture was developed by leveraging a multi-objective neural architecture search that optimizes both accuracy and FLOP. Figure 2 shows the architecture of EfficientNet-B0.

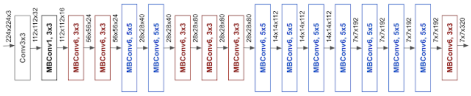


Fig. 2: EfficientNet-B0 baseline network architecture [33]

*2) **ResNet-18**:* ResNet-18 employs residual learning [19] to mitigate the vanishing gradient problem, which occurs when the gradients from the loss function become too small to update the weights of the network. This is achieved by adding shortcut (residual) connections between the input and output of each stack of layers (residual blocks).

The model architecture of ResNet18 is given in Figure 3.

*3) **Vision Transformer**:* Vision Transformer is a natural language processing-based image classification model. Similar to the original Transformer, ViT consists of an encoder network and uses self attention to for classification. The input image is split into visual tokens which are patches of fixed size. The architecture is illustrated in Figure 4.

All the models used were pre-trained on ImageNet. The batchsize was 128. The models were trained for 30 epochs, using a step scheduler (gamma =0.2). The learning rate was 0.001. The models were evaluated according to five evaluation metrics. AUC-ROC, Accuracy, F1 Score, Precision, Recall.

IV. Results and Discussion

The section provides an overview of the results obtained.Table II provides a summary of the results. We then provide a comparative analysis in subsection IV-D.

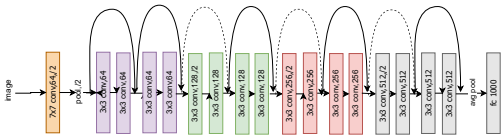


Fig. 3: ResNet-18 Model Architecture

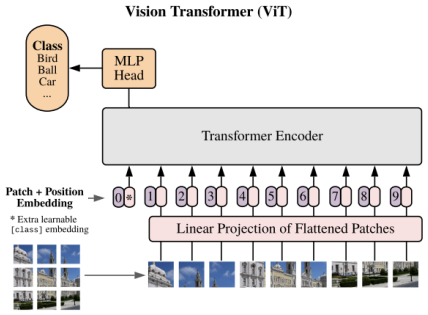


Fig. 4: Vision Transformer architecture

TABLE II: Overall accuracy of the models on all 11 activities combined. GFlops are calculated for a single input image of shape 224x224x3.

Method	Local Tag Accuracy	Global Tag Accuracy	GFlops
EfficientNet-B0	86%	68%	0.42
ResNet-18	83%	53%	1.83
VisionTransformer-B/16	84%	66%	17.58

A. *EfficientNet-B0*

EfficientNet-B0 demonstrated the highest accuracy for both the *local* and *global* tags, achieving 86% and 68% accuracy, respectively. Figure 5a illustrates that, concerning the *local* tags, falling activities 1 to 5 were detected with an average AUC of 94.81%, while non-falling activities gave an average AUC of 91.79%. Within the falling activities, EfficientNet performed the best in detecting activity 2, while it had the lowest AUC for activity 5. Overall, it performed the worst on activity 10 with an AUC of 79.15%. Figure 5b provides results of performance on *global tags*, presenting the AUC achieved for each class. It achieved an average AUC of 81.99% for falling activities 1 to 5, while non-falling activities demonstrated an average AUC of 90.08%. Among the falling activities, EfficientNet performed most effectively in detecting activity 2, while it had the lowest AUC for activity 5. It performed the worst on activity 5 with the lowest AUC of 63.36%.

B. *ResNet-18*

ResNet-18 performed the lowest on both the local and global tags with an accuracy of 83% and 53% respectively. Figure 6a illustrates that, concerning the *local* tags, falling activities 1 to 5 provided an average AUC of 88.072%, while non-falling activities gave an average AUC of 87.91%. Within the falling activities, ResNet performed the best in detecting activity 1, while it had the lowest AUC for activity 5. Overall, it performed the worst on activity 9 with an AUC of 78.65%. Figure 6b provides additional insights into the performance of ResNet on *global tags*. It achieved an average AUC of 79.91% for falling activities 1 to 5, while non-falling activities demonstrated an average AUC of 94.837%. Among the falling activities, ResNet performed most effectively in detecting activity 2, while it had the lowest AUC for activity 5. It had the worst performance on activity 5 with the lowest AUC of 62.46%.

C. *Vision Transformer-B/16*

ViT gave an accuracy of 84% and 66% on the local and global tags respectively. Figure 7a provides insights into the performance of ViT on *local tags*, presenting the AUC achieved for each class. Within falling activities, ViT performed most effectively in detecting activity 2, while it had the lowest AUC for activity 4. Overall, it performed the worst on activity 9. Figure 7b provides additional insights into the performance of ViT on *global tags*, presenting the AUC for each class. Among the falling activities, ViT performed most effectively in detecting activity 1, while it had the lowest AUC for activity 4. It performed the worst on activity 7 with the lowest AUC.

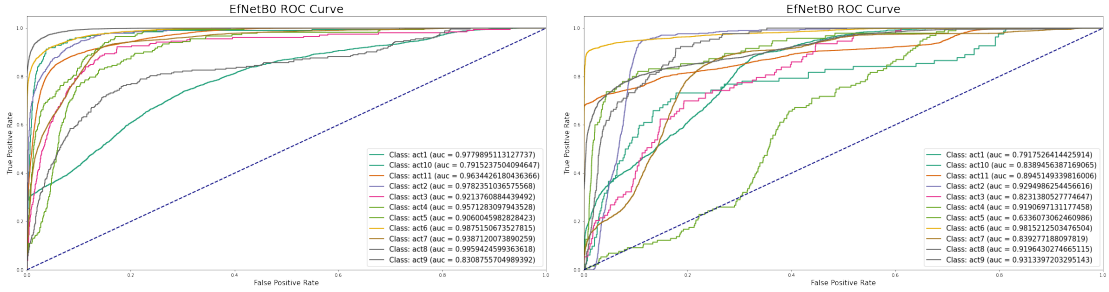
D. *Comparative Analysis*

As shown in Table II, EfficientNet demonstrates the best performance for both local and global tags, outperforming ResNet and ViT. Furthermore, EfficientNet is computationally efficient, requiring the least number of GFLOPs (0.42). In contrast, ResNet exhibits the worst performance. On the contrary, ViT produced comparable results to ResNet; however, its computational cost is notably higher due to its extensive usage of GFLOPs. In terms of fall detection, it is noteworthy that EfficientNet showed the best performance in classifying activity 2 (falling forwards using knees) for both local and global tags. On the other hand, ResNet performed the best on activity 2 within fall detection classes. When analyzing the ROC curves, it is evident that both ResNet and EfficientNet had lower performance in classifying activity 5 (falling forward using hands) for both local and global tags. However, EfficientNet achieved higher performance than ResNet for both local and global tags in this specific activity. This implies that EfficientNet is more effective in detecting falls, even in challenging scenarios like forward falls using knees. Based on the results, it is evident that the accuracy of local tags surpasses that of global tags. This supports our hypothesis that using still images (local tags) extracted from videos can be an effective approach to fall detection. It suggests that there is a more computationally efficient method for fall detection, where analyzing individual frames of a video is feasible, rather than processing the entire video sequence.

We have chosen not to compare our results with other techniques because as the literature review highlights, those methods rely on precomputed optical flow features computed. In contrast, our approach centers on the utilization of video frames for real-time processing. The application of optical flow to real-time video proves impractical within this context. Consequently, our focus shifts from benchmarking against previous methods to comparison amongst the state-of-the-art models of our approach.

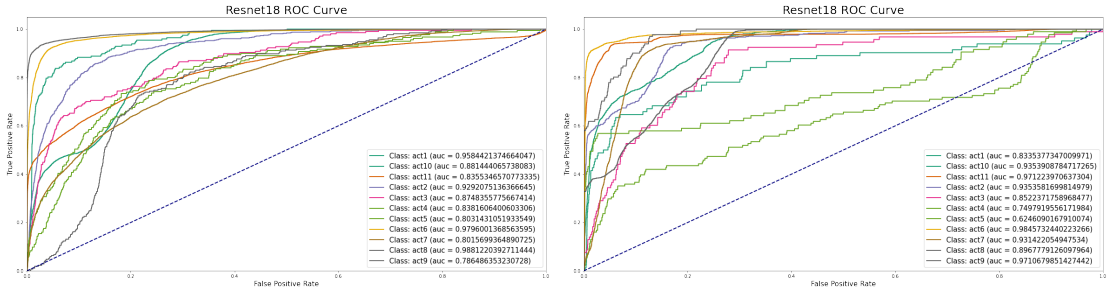
V. Limitations and Future Work

There are two main limitations in this work. The first limitation is the limited use of the UP-FALL [34] dataset.



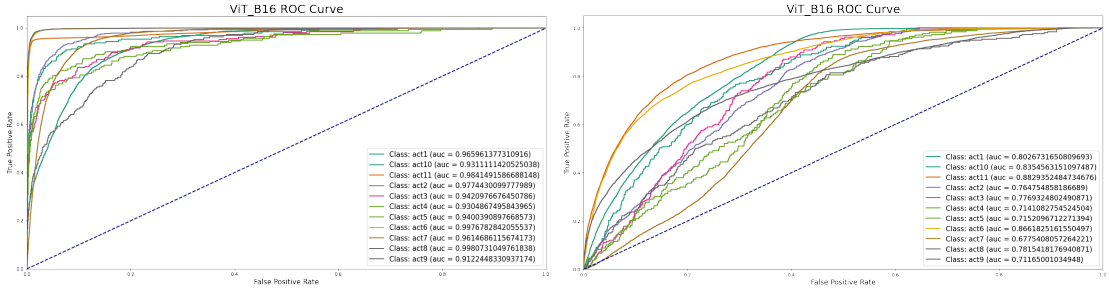
(a) EfficientNet-B0 Local Tag (b) EfficientNet-B0 Global Tag

Fig. 5: EfficientNet-B0 AUC-ROC curves.



(a) ResNet-18 Local Tag (b) ResNet-18 Global Tag

Fig. 6: ResNet-18 AUC-ROC curves.



(a) VisionTransformer-B/16 Local Tag (b) VisionTransformer-B/16 Global Tag

Fig. 7: VisionTransformer-B/16 AUC-ROC curves.

Specifically, we were only able to utilize five out of the seventeen subjects in the dataset due to resource constraints. This limited sample size may impact the generalizability of our findings. The second limitation is the lack of utilizing architectures of different scales in our methodology. Although we employed various types of architectures, incorporating an analysis of different variants within each architecture could have enhanced the analysis and potentially yielded better results. For example, in addition to using ResNet-18, we could have also included ResNet-50 [35] in our analysis. In our future work, we aim to evaluate our fall detection methodology more effectively. To achieve this, we will test our approach on a larger dataset, allowing for a more comprehensive analysis. Additionally, we plan to incorporate computationally efficient sequential data and compare our results on standard benchmark datasets to assess how well our approach performs as compare to existing state-of-the-art methods. This comparative analysis will provide valuable insights, further contributing to the advancement of fall detection techniques.

VI. Conclusion

State-of-the-art deep learning models present us with immense opportunities to make our lives easier for our loved ones. The method proposed in this study divides the existing UP-FALL dataset into local and global tags. Through the utilization of state-of-the-art CNN and Vision Transformer models, the study demonstrates the effectiveness and accuracy of deep learning algorithms in recognizing and distinguishing between various falling and non-falling events. The results indicate that EfficientNet is the most appropriate model for fall detection using images. Additionally, the analysis demonstrates that local tags yield favorable accuracy in detecting

falls. However, it is important to note that there is potential for further refinement and improvement in future research to enhance the system's performance and achieve more robust results. The findings of this research indicate that deep learning-based activity recognition systems have the ability to improve the quality of care provided to the elderly population. Furthermore, employing only video frames specifically for fall detection purposes can enhance overall efficiency and reduce computational requirements. This knowledge can be leveraged to ensure the safety, well-being, and independence of the elderly, while also enabling timely intervention in case of emergencies or abnormal activities.

VII. Acknowledgements

This research was conducted under Dr. Muhammad Farhan as part of the Summer Tehqiq Research Program, an initiative by Habib University. We acknowledge the valuable support and funding provided by this program for the research.

References

[1] "Keep on your feet—preventing older adult falls," Mar 2023. 1

[2] " Centers for Disease Control and Prevention." 1

[3] Y. Feng, Y. Wei, K. Li, Y. Feng, and Z. Gan, "Improved pedestrian fall detection model based on yolov5," in *2022 IEEE 6th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2022, pp. 410–413. 1

[4] A. Singh, P. Koshy, and B. S. Manoj, "Multi-person fall detection in complex iot-assisted living environments," in *2022 IEEE 19th India Council International Conference (INDICON)*, 2022, pp. 1–7. 1

[5] S. Renake, N. Singh, A. Singh, P. Adke, K. Bhangale, and R. Mapari, "Iot based fall detection system," in *2022 6th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, 2022, pp. 1–5. 1

[6] B. Wang and Y. Guo, "Soft fall detection using frequency modulated continuous wave radar and regional power burst curve," in *2022 Asia-Pacific Microwave Conference (APMC)*, 2022, pp. 240–242. 1

[7] J. Li, Q. Zhao, T. Yang, and C. Fan, "An algorithm of fall detection based on vision," in *2021 6th International Symposium on Computer and Information Processing Technology (ISCIPT)*, 2021, pp. 133–136. 1

[8] T.-H. Tsai, C.-W. Hsu, and W.-C. Wan, "Live demonstration: Vision-based real-time fall detection system on embedded system," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2020, pp. 1–1. 1

[9] G. Chen and X. Duan, "Vision-based elderly fall detection algorithm for mobile robot," in *2021 IEEE 4th International Conference on Electronics Technology (ICET)*, 2021, pp. 1197–1202. 1

[10] R. Espinosa, H. Ponce, S. Gutiérrez, L. Martínez-Villaseñor, J. Brieva, and E. Moya-Albor, "A vision-based approach for fall detection using multiple cameras and convolutional neural networks: A case study using the up-fall detection dataset," *Comput. Biol. Med.*, vol. 115, no. C, dec 2019. 1

[11] O. Kerdjidj, N. Ramzan, k. Ghanem, A. Amira, and F. Chouireb, "Fall detection and human activity classification using wearable sensors and compressed sensing," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, 01 2020. 1

[12] A. Sufian, C. You, and M. Dong, "A deep transfer learning-based edge computing method for home health monitoring," in *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, 2021, pp. 1–6. 1

[13] G. Koshmak, M. Lindén, and A. Loutfi, "Challenges and issues in multisensor fusion approach for fall detection: Review paper," *Journal of Sensors*, vol. 2016, 12 2015. 1

[14] I. Charfi, J. Miteran, J. Dubois, M. Atri, and R. Tourki, "Definition and performance evaluation of a robust svm based fall detection solution," in *2012 Eighth International Conference on Signal Image Technology and Internet Based Systems*, 2012, pp. 218–224. 1

[15] S. Kozina, H. Gjoreski, M. Gams, and M. Luštrek, "Efficient activity recognition and fall detection using accelerometers," in *Evaluating AAL Systems Through Competitive Benchmarking*, J. A. Botía, J. A. Álvarez-García, K. Fujinami, P. Barsocchi, and T. Riedel, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 13–23. 1

[16] N. Zerrouki and A. Houacine, "Combined curvelets and hidden markov models for human fall detection," *Multimedia Tools and Applications*, vol. 77, 03 2018. 1

[17] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "Robust video surveillance for fall detection based on human shape deformation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 5, pp. 611–622, 2011. 1

[18] E. Alam, A. Sufian, P. Dutta, and M. Leo, "Vision-based human fall detection systems using deep learning: A review," *Computers in Biology and Medicine*, vol. 146, p. 105626, 2022. 1

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. 2, 3

[20] M. M. Hasana, M. Ibrahim, and M. S. Ali, "Speeding up efficientnet: Selecting update blocks of convolutional neural networks using genetic algorithm in transfer learning," 2023. 2, 3

[21] C. Thurau and V. Hlavac, "Pose primitive based human action recognition in videos or still images," 06 2008. 2

[22] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, "Deep learning models for real-time human activity recognition with smartphones," *Mobile Networks and Applications*, vol. 25, no. 2, p. 743–755, 2019. 2

[23] V. K., S. K., and P. M., "Video-based human activity recognition for elderly using convolutional neural network," *International Journal of Security and Privacy in Pervasive Computing*, vol. 12, no. 1, p. 36–48, 2020. 2

[24] S. Deep and X. Zheng, "Leveraging cnn and transfer learning for vision-based human activity recognition," *2019 29th International Telecommunication Networks and Applications Conference (ITNAC)*, 2019. 2

[25] A. Jalal, S. Kamal, and D. Kim, "A depth video-based human detection and activity recognition using multi-features and embedded hidden markov models for health care monitoring systems," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 4, p. 54, 2017. 2

[26] K. Kim, A. Jalal, and M. Mahmood, "Vision-based human activity recognition system using depth silhouettes: A smart home system for monitoring the residents," *Journal of Electrical Engineering and Technology*, vol. 14, no. 6, p. 2567–2573, 2019. 2

[27] M. A. R. Ahad, A. D. Antar, and O. Shahid, "Vision-based action understanding for assistive healthcare: A short review," in *CVPR Workshops*, 2019. 2

[28] K. Adhikari, H. Bouchachia, and H. Nait-Charif, "Activity recognition for indoor fall detection using convolutional neural network," *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, 2017. 2

[29] H. Ramirez, S. A. Velastin, I. Meza, E. Fabregas, D. Makris, and G. Farias, "Fall detection and activity recognition using human skeleton features," *IEEE Access*, vol. 9, p. 33532–33542, 2021. 2

[30] T. T. Zin, Y. Htet, Y. Akagi, H. Tamura, K. Kondo, S. Araki, and E. Chosa, "Real-time action recognition system for elderly people using stereo depth camera," *Sensors*, vol. 21, no. 17, p. 5895, 2021. 3

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. 3

[32] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *CoRR*, vol. abs/1905.11946, 2019. 3

[33] M. Tan, "Efficientnet: Improving accuracy and efficiency through automl and model scaling," May 2019. 3

[34] L. Martínez-Villaseñor, H. Ponce, J. Brieva, E. Moya-Albor, J. Núñez-Martínez, and C. Peñafort-Asturiano, "Up-fall detection dataset: A multimodal approach," *Sensors*, vol. 19, no. 9, 2019. 4

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. 5