ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.1347

Fall detection for older adults: A comparative study of CNN-based deep learning and ViT architectures

Jorge L. Charco, Dr.[1,2], Angela Yanza-Montalván, Dra.[1],
Angélica Cruz-Chóez, Msc.[1], Verónica Mendoza-Morán, Msc.[1]
[1]Universidad de Guayaquil, Ecuador
[2]Artificial Intelligence Research Group, Universidad Bolivariana del Ecuador, Campus Durán, Ecuador
jorge.charcoa@ug.edu.ec, angela.yanzam@ug.edu.ec, angelica.cruzc@ug.edu.ec, veronica.mendozam@ug.edu.ec

Abstract*—This paper presents a comparative study on fall detection in older adults using four pre-trained convolutional neural network models (VGG16, VGG19, ResNet50V2 and ResNet101V2), and Vision Transformer for Image Classification (ViT) model. Falls among older adults remain one of the leading cause of injury and reduced quality of life, and real-time detection systems can allow timely intervention to minimize harm. The proposed models are evaluated on a publicly available dataset composed of RGB images categorized into falls and non-falls. Each image is pre-processed through cropping, resizing to 128×128 and 224×224 pixels for CNN-based and ViT models, respectively, and Min-Max normalization. Transfer learning is applied to fine-tune the models using ImageNet-initialized weights, modifying the final layers to address the binary classification task. Models are trained and tested under consistent conditions, and performance is evaluated using accuracy, precision, recall, and F1 score metrics, supported by confusion matrices and ROC curves. Among the models, ViT model achieves the highest classification accuracy (98%) and demonstrates a strong balance across performance metrics, particularly in detecting actual falls cases, which is critical in healthcare applications. This study confirms the effectiveness of ViT models for fall detection from single-camera input without the need for wearable sensors.

***Keywords*—fall detection, convolutional neural network, older adults, vision transformer, attention modules.**

I. Introduction

The increasing proportion of older adults in the global population poses significant challenges to healthcare systems, particularly with regard to safety, mobility, and autonomy. Among these, falls represent one of the leading causes of injury, disability, and mortality in this demographic. According to the Centers for Disease Control and Prevention (CDC), more than one in four older adults falls each year, and one in five falls cause serious injury such as broken bones or head trauma. Beyond physical consequences, falls also impose substantial psychological impacts, including fear of falling again, which can lead to decreased mobility, isolation, and deterioration in quality of life [1].

Fall detection systems, especially those that use artificial intelligence (AI), have emerged as key tools in both clinical and home settings. These systems are designed to identify fall events in real-time, enabling rapid response and intervention, which is crucial to reducing morbidity and mortality. According to [2], [3], traditional fall detection methods often rely on wearable sensors such as accelerometers and gyroscopes. However, these solutions are limited by user compliance and false alarm rates, which motivates the shift to vision-based and ambient approaches.

A growing body of research has explored machine learning (ML) and AI-driven models for falling detection. Early studies focused on threshold-based systems and classical ML classifiers, including support vector machines (SVM), decision trees, k-nearest neighbors (KNN), and artificial neural networks (ANN) [4], [5]. These methods achieved moderate success when applied to static and structured datasets but often failed to generalize in dynamic real-world conditions. To address these limitations, the authors in [6] have turned to deep learning architectures, such as convolutional neural networks (CNNs), which can learn spatio-temporal patterns directly from raw sensor data. In [7], the authors have also proposed hybrid systems that integrate edge AI and cloud computing, allowing real-time inference while preserving privacy and minimizing latency. The authors in [3] have proposed a mobile Internet of Things (IoT) approach with ML support for hybrid cloud-edge systems, demonstrating improved responsiveness and scalability. The authors in [8] have proposed a similar approach, which uses edge AI in a low-power embedded device for efficient fall detection using AI-based feature extraction. During recent years, different studies have been proposed to explore falls in older adults. The authors in [9] have proposed an analysis of scientific evidence of AI applications in the analysis of data related to postural control and risk of falls. The results identified certain risk factors that can cause a fall such as age, sex, stride length, and posture [10]–[13].

Recent developments incorporate Human Pose Estimation (HPE) as a method for inferring body kinematics and transitions, which mapping skeletal landmarks, these systems can better differentiate between fall events and benign movements such as sitting or lying down. However, the fusion of pose-based features with sequence-aware ML models such as LSTM [14], GRU [15], or transformers [16] remains underexplored. The authors in [17] have proposed the use of a smart gait analyzer with the help of deep learning techniques to assess the diagnostic accuracy of the risk of falling through vision. For this, the gait parameters of each participant are extracted using a computer vision algorithm such as AlphaPose. AlphaPose algorithm [18] is used by them to recognize human pose as a first step. This algorithm applies a symmetric spatial trans-

Fig. 1. Set of images that are used for training and testing process of proposed approach. (a) Images corresponding to non-falls. (b) Images corresponding to falls.

former network to extract accurate single-person region and employs parametric pose non-maximum suppression to reduce redundant detection. As a second step, three classical machine learning models, Naive Bayes [19], Logistic Regression [20], and Support Vector Machine [21] algorithms, were used to train binary classifiers to detect abnormal walking patterns. The Naive Bayes classifier achieved the best performance with a full data accuracy of 90.14%. In [22], the authors have proposed the use of a multi-depth camera human motion tracking system for fall risk assessment. The proposed model captured patients performing the well-known and validated Berg Balance Scale (BBS). The trained model was able to predict the patient's 14 BBS scores by extracting spatio-temporal features from the captured human motion record. The results obtained showed that the accuracy of the fall risk assessment was 97%.

In the current work, the use of pre-trained models, such as VGG, ResNet and ViT, are considered. These algorithms are able to detect objects into images, including persons into of a scene. These models only use images as input, which are captured from an one-view approach, i.e., images captured using only one camera. No sensors or additional devices will be used. This research has the following contributions:

- Use different Convolutional Neural Networks and ViT models to evaluate the problem of falls in older adults.
- Compare the experimental results of the proposed models in different scenarios, including different occlusions in the scene, to determine their performance.

The remainder of the paper is organized as follows. In Section II, the material and methods used in this article are presented, including dataset, proposed model, and the metrics used in this study. The experimental results are reported in Section III, the discussion of the results obtained is presented in Section IV. Finally, conclusions are given in Section V

II. Material and methods

A. Dataset Collection

The fall detection dataset was proposed by [23], and it is currently one of the publicly available fall detection benchmarks. This dataset contains different human activities such as walking, sitting, standing and falling. The dataset consists of around 485 RGB images of people, which are classified into two categories such as falls and non-falls (see Fig. 1), each category has 258 and 227 RGB images, respectively.

B. Images Pre-processing

As pre-processing dataset, the images were cropped and resized to 128×128 pixels for the training process of CNN-Based deep learning, and 224×224 pixels for ViT model. Likewise, the Min-Max normalization [24] was computed in each image as part of the normalization of the images. The pre-processing of images used during the training process was also used during the evaluation process. According to the obtained images, they were divided into two parts, 374 images for the training process and 111 images for the evaluation process.

C. Methods

To perform the fall detection in older adults, two architectures were proposed with the goal of detecting a possible fall in the area detected by the camera. The image dataset contains three channels (RGB). The architectures are described below.

1) VGG16 model: It is a convolutional neural network architecture proposed by [25] (see Fig. 2). It consists of 13 convolutional layers followed by 3 fully connected layers. The architecture uses small convolutional filters (3×3 kernel size) and max-pooling layers (2×2) to progressively reduce the spatial dimensions. The network uses ReLU as the activation function for non-linearity and a softmax classifier [26] at the output layer for final prediction. The feature maps generated by the convolutional blocks were flattened and passed through dense layers to perform classification. In this approach proposed, a pre-trained VGG16 model is

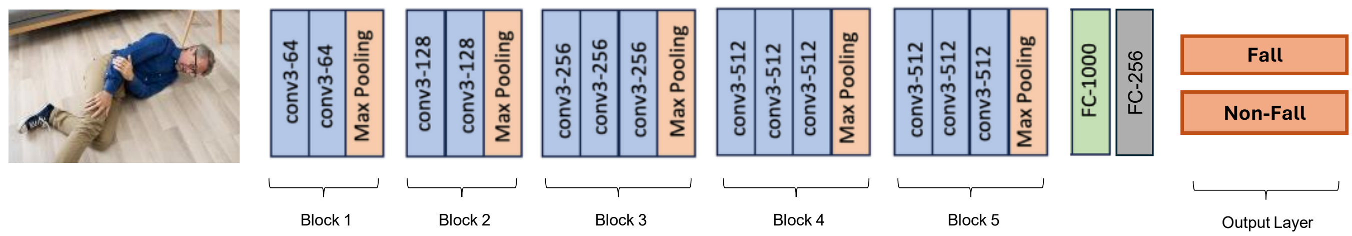

Fig. 2. VGG base architecture for classification problems.

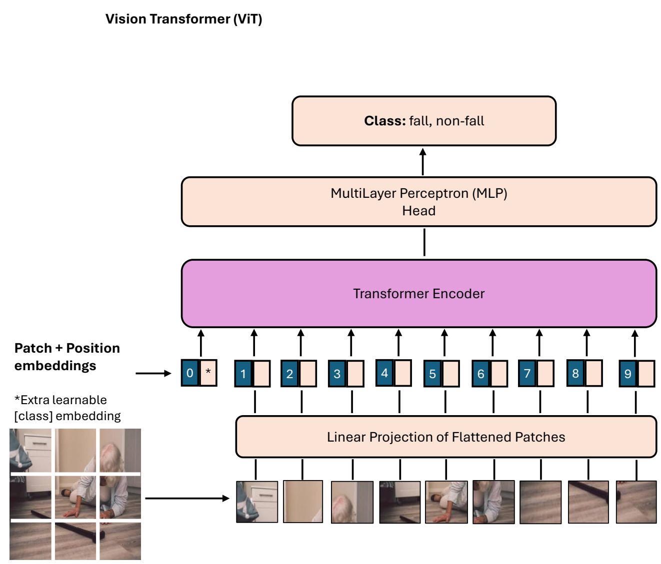

Fig. 3. ViT architecture for classification problems.

fine-tuned for the task of falls or non-falls classification in older adults. For this, the final layer was modified to estimate the class membership likelihood (falls, non falls), considering softwax as the activation function [27]. The weights of the convolutional layers are initialized from the ImageNet dataset to leverage learned feature representations while minimizing overfitting.

2) VGG19 model: It consists of 16 convolutional layers and 3 fully connected layers, totaling 19 weight layers. Similar to the VGG16 model [25] (see Fig. 2), it uses small convolutional filters of size 3×3, and uses max-pooling layers (2×2) to reduce spatial resolution. ReLu activation is used after each convolutional layer to introduce non-linearity. This architecture is initialized with pre-trained weights from the ImageNet dataset, and allows learning highly abstract features, making it effective for transfer learning tasks. Considering previously mentioned, the final layers of the model are modified to tackle the falls and non-falls classification problem in older adults.

3) ResNet50V2 model: It is a neural network based on the ResNet50 architecture proposed by [28] (see Fig. 4), which introduces modifications that improve gradient flow during training. The model comprises 50 layers organized into a series of residual blocks, where each block contains identity shortcuts to bypass one or more layers. The residual connections are modified to use pre-activation, where batch normalization and ReLU are applied before the convolutional layers, enhancing training stability and improving accuracy. The model uses a global average pooling layer followed by a fully connected layer and softmax activation for classification. A pre-trained ResNet50V2 model is used to tackle the problem of falls in older adults since allows extracting high-level features with reduced risk of vanishing gradients. However, the final layers of the model are modified to solve the new classification problem.

4) ResNet101V2 model: It is a deep convolutional neural network architecture with the depth to 101 layers, which is an extension of the ResNetV2 family (see Fig. 4), enhancing its capacity to learn complex feature hierarchies. Like ResNet50V2 [28], it employs a pre-activation residual structure, where batch normalization and ReLU activation precede the convolutional operations within each residual block. In order to help maintain gradient flow and enable the training of very deep networks without performance degradation, multiple bottleneck layers and identity shortcut connections are considered. A pre-trained ResNet101V2 model is used to solve the problem of falls in older adults; therefore, the final layers are modified to tackle this problem.

5) ViT for Image classification model: It is a vision transformer-based architecture designed for image recognition tasks (see Fig. 3), which processes image data by dividing each input image into a sequence of fixed-size patches. Each patch is linearly embedded and combined with positional embeddings to retain spatial information. These embedded patches are then fed into a transformer encoder composed of multiple layers of multi-head self-attention and feed-forward networks, being able to capture global dependencies across the image efficiently. A classification token is appended to the patch sequence, and its output embedding is passed through a multilayer perceptron (MLP) head with a softmax activation for final classification. It is initialized with pre-trained weights from ImageNet and fine-tuned to classify images into fall and non-fall categories, improving feature representation and generalization performance.

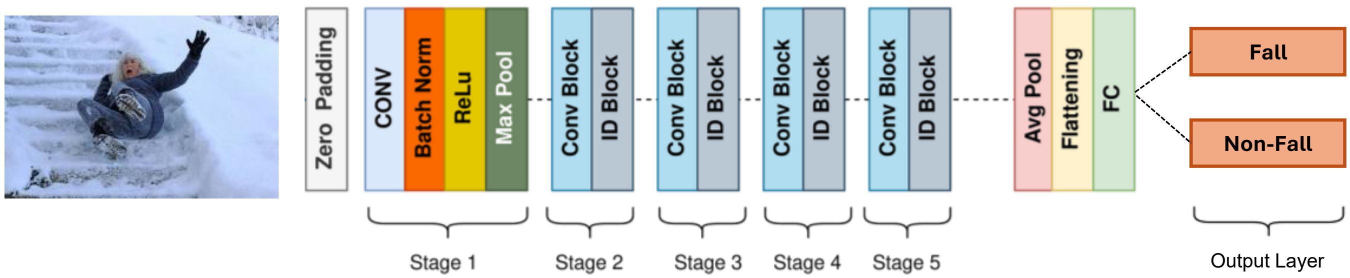

Fig. 4. ResNet base architecture for classification problems.

D. Metrics

In order to measure the performance of the used architectures, a set of metrics such as accuracy, precision, recall/sensitivity and F1 score are considered, including a confusion matrix and ROC curve. These metrics are used in classification problems. The values near to one indicate a higher accuracy. These metrics are formulated as follows:

$$accuracy = \frac{TN + TP}{TP + TN + FP + FN} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall(or)sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$F1Score = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

where TP (true positive), TN (true negative), FP (false positive), FN (false negative).

III. Experimental Results

As mentioned above, these models are used to solve the problem of falls in older adults. Four models are based on CNN (VGG and ResNet), and one of them is based on ViT, which are modified to tackle the new classification task. Keras (CNN-based models) and Pytorch (ViT model) library in python are used to implement these models. It has been trained using a 2.9 Ghz. Ryzen 7 with 16GB of memory and GeForce RTX 3050 GPU. The details of the experimental results of each proposed model, are presented in this section. The dataset considered for the training process of each model mentioned above was presented in Section II-A, which is divided into two stages (fall or non-fall), each stage had 208 and 166 RGB images. The images pre-processing were presented in Section II-B. The images were resized to 128×128 pixels for CNN models, and to 224×224 for ViT model, and then normalized considering Min-Max normalization [24]. For the training process, a set of 374 RGB images was used to feed the proposed models, which were trained until 50 epochs. In the evaluation a set of 111 RGB images was considered.

To protect pre-trained features during initial training on a new small dataset, transfer learning is used, which involves freezing all the weights of the base model, and leaving the final layers for new knowledge. Two layers with a width of 1000 and 256 output neurons, respectively, were used as the final layers of the CNN-based models proposed in this study, including dropout, which helps prevent overfitting during the training process. ReLu activation function was used in each layer mentioned above. Softmax activation is used by all models (CNN and ViT models) for the classification task. Likewise, Categorical CrossEntropy loss function [30] and Adam optimizer [31] are considered to train the different models. Additionally, the learning rate is set to 0.001.

The experimental results obtained with the proposed models are presented in Table I. The models were trained using different parameter settings. The results of proposed models were compared between them using the metric mentioned in Section II-D. The results of the proposed models were compared between them using metrics mentioned in Section II-D, being the results of ViT model better than the results obtained by other proposed models. Figure 5 shows the confusion matrix of the proposed models to tackle the problem of falls in older adults, considering the metrics mentioned previously. The ROC curves were used to compare the true-positive rate and the false-positive rate of falls and non-falls of older adults for each proposed model (see Fig 5).

IV. Discussion

As mentioned previously, the obtained results for each proposed models (VGG, ResNet and ViT models) are shown in Table I. The results of ViT model show better performance in metrics such as accuracy, precision, recall and F1-score, if they are compared with other proposed models. Considering the context of the problem, the Recall metric could be decisive for the problem addressed in this study, specifically in falls cases of older adults. The number of parameters of proposed models (VGG16, VGG19, ResNet50V2, ResNet101V2 and ViT models), which allow the evaluation of the computational cost, were of 23M, 28M, 56M, 75M and 85M parameters, respectively. For the design of proposed CNN-based models, each set of convolution blocks, max-pooling, skip-connections, back-normalization, among others, have allowed extracting high-level features to detect falls in older adults. For the case of ViT model, the use of attention modules and positional embeddings to retain spatial information, has also helped to capture global dependencies in data sequence, allowing to

TABLE I
PERFORMANCE EVALUATION OF THE MODELS USED FOR CLASSIFICATION TASK OF FALLS IN OLDER ADULTS (VGG16, VGG19, RESNET50V2, RESNET101V2 AND VIT FOR IMAGE CLASSIFICATION MODELS)

Models	Params	Runtime	Metric	Non-Fall	Fall
VGG16 [25]	23M	3.3 ms	Accuracy	0.85	
			Precision	0.75	0.92
			Recall	0.88	0.83
			F1 Score	0.81	0.87
VGG19 [25]	28M	4.3 ms	Accuracy	0.87	
			Precision	0.83	0.90
			Recall	0.83	0.90
			F1 Score	0.83	0.90
ResNet50V2 [28]	56M	8.6 ms	Accuracy	0.90	
			Precision	0.88	0.92
			Recall	0.85	0.93
			F1 Score	0.86	0.92
ResNet101V2 [28]	75M	16.8 ms	Accuracy	0.86	
			Precision	0.77	0.94
			Recall	0.90	0.84
			F1 Score	0.83	0.89
ViT [29]	85M	14.3 ms	Accuracy	0.98	
			Precision	0.98	0.99
			Recall	0.98	0.99
			F1 Score	0.98	0.99

the model finds relation-ship/dependence between patches of images, getting better results. The accuracy of ViT model is 0.98, which shows that the model is able to obtain relevant features to tackle the problem of falls in older adults. This may be due to the configuration of models, capture global dependencies, and relations between patches of images, which allow getting relevant features of each image. Other important metric is Recall metric, which helps to measure the ability of model to predict all actual positives correctly. The results of this metric for non-falls and fall classes were 0.98 and 0.99, respectively, which show that the ViT model is able to predict falls cases better than other proposed models (VGG16=0.83, VGG19=0.90, ResNet50V2=0.93 and ResNet101V2=0.84), which is important in this study. Likewise, in the context of this problem, minimizing false positives is essential, as a high-accuracy score indicates that when the model correctly predicts the positive class, it is likely correct, which is important to correctly determine whether an older adult has fallen or not. Considering this, ViT precision metric for both cases, non-falls and falls cases, were 0.98 and 0.99, which were better than the results obtained for other proposed models. In order to measure how often the model correctly predicts the outcome, the accuracy metric is considered. In this sense, similar to the results previously obtained, the ViT accuracy metric is the best with 0.98 compared to the other proposed models (VGG16=0.85, VGG19=0.87, ResNet50V2=0.90 and ResNet101V2=0.86).

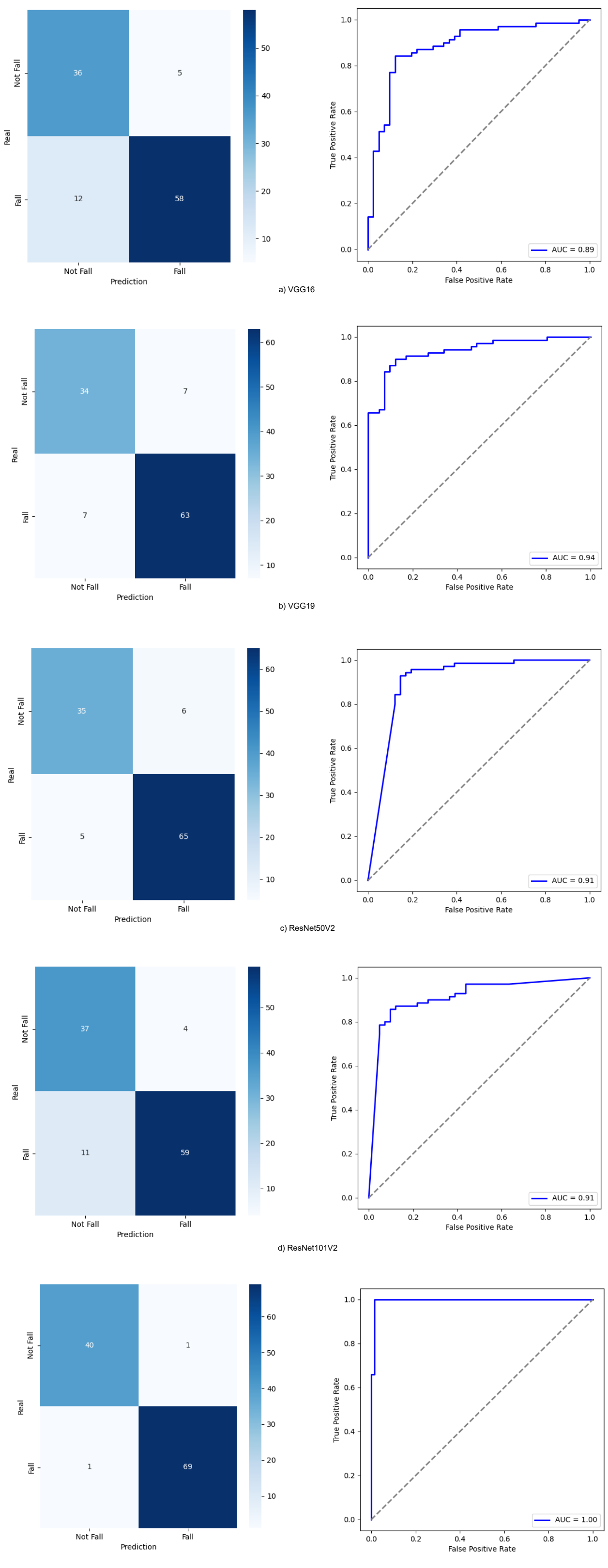

Fig. 5. Confusion Matrix of results obtained by proposed models (VGG, ResNet and ViT models).

V. Conclusions

This study addresses the important challenge of detecting falls in older adults using pre-trained Convolutional Neural Network and Vision Transformer models. Falls remain one of the most significant causes of injury and loss of independence among older adults. In this work, different models (VGG16, VGG19, ResNet50V2, ResNet101V2 and ViT) were evaluated on a publicly available fall detection dataset. The obtained experimental results show that the ViT model achieved the highest overall accuracy (0.98) and demonstrated superior performance in both precision and recall metrics for fall detection. This model balances depth and computational efficiency, benefiting from attention modules and positional embeddings that improve training. Although other models such as VGG19 and ResNet50V2 also achieved competitive results, the second model required higher computational resources with a slight improvement in fall detection performance. In the context of fall detection, false negatives, i.e., failing to identify an actual fall, are critical to minimize. The preliminary results show that ViT model is particularly effective in identifying such high-risk cases, although they are limited to a small benchmark dataset. Future work will aim to extend this approach to multi-view environments and integrate pose estimation or temporal modeling techniques such as recurrent networks to enhance robustness and generalization in real-world settings.

References

[1] N. Thakur and C. Y. Han, "A study of fall detection in assisted living: Identifying and improving the optimal machine learning method," *Journal of Sensor and Actuator Networks*, vol. 10, no. 3, 2021.

[2] S. Usmani, A. Saboor, M. Haris, M. A. Khan, and H. Park, "Latest research trends in fall detection and prevention using machine learning: A systematic review," *Sensors*, vol. 21, no. 15, 2021.

[3] D. Mrozek, A. Koczur, and B. Małysiak-Mrozek, "Fall detection in older adults with mobile iot devices and machine learning in the cloud and on the edge," *Information Sciences*, vol. 537, pp. 132–147, 2020.

[4] P. Vallabh, R. Malekian, N. Ye, and D. C. Bogatinoska, "Fall detection using machine learning algorithms," in *2016 24th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 2016, pp. 1–9.

[5] D. Yacchirema, J. S. de Puga, C. Palau, and M. Esteve, "Fall detection system for elderly people using iot and ensemble machine learning algorithm," *Personal Ubiquitous Comput.*, vol. 23, p. 801–817, 2019.

[6] S. Badgujar and A. S. Pillai, "Fall detection for elderly people using machine learning," in *International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, pp. 1–4.

[7] A. Chelli and M. Pätzold, "A machine learning approach for fall detection and daily living activity recognition," *IEEE Access*, pp. 38 670–38 687, 2019.

[8] B.-S. Lin, T. Yu, C.-W. Peng, C.-H. Lin, H.-K. Hsu, I.-J. Lee, and Z. Zhang, "Fall detection system with artificial intelligence-based edge computing," *IEEE Access*, vol. 10, pp. 4328–4339, 2022.

[9] A. González-Castro, R. Leirós-Rodríguez, C. Prada-García, and J. A. Benítez-Andrades, "The applications of artificial intelligence for assessing fall risk: Systematic review," *J Med Internet Res*, vol. 26, 2024.

[10] S. Gillain, M. Boutaayamou, C. Schwartz, O. Brüls, O. Bruyère, J.-L. Croisier, E. Salmon, J.-Y. Reginster, G. Garraux, and J. Petermans, "Using supervised learning machine algorithm to identify future fallers based on gait patterns: A two-year longitudinal study," *Experimental Gerontology*, vol. 127, 2019.

[11] Y. Lo, S. F. Lynch, R. J. Urbanowicz, R. S. Olson, A. Z. Ritter, C. R. Whitehouse, M. O'Connor, S. K. Keim, M. McDonald, J. H. Moore *et al.*, "Using machine learning on home health care assessments to predict fall risk," in *MEDINFO 2019: Health and Wellbeing e-Networks for All*. IOS Press, 2019, pp. 684–688.

[12] B. Noh, C. Youm, E. Goh, M. Lee, H. Park, H. Jeon, and O. Y. Kim, "Xgboost based machine learning approach to predict the risk of fall in older adults using gait outcomes," *Scientific reports*, vol. 11, no. 1, p. 12183, 2021.
[13] A. Dubois, T. Bihl, and J.-P. Bresciani, "Identifying fall risk predictors by monitoring daily activities at home using a depth sensor coupled to machine learning algorithms," *Sensors*, vol. 21, no. 6, 2021.
[14] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.
[15] T. R. Mim, M. Amatullah, S. Afreen, M. A. Yousuf, S. Uddin, S. A. Alyami, K. F. Hasan, and M. A. Moni, "Gru-inc: An inception-attention based approach using gru for human activity recognition," *Expert Systems with Applications*, vol. 216, p. 119419, 2023.
[16] W. Mao, Y. Ge, C. Shen, Z. Tian, X. Wang, and Z. Wang, "Tfpose: Direct human pose estimation with transformers," *arXiv preprint arXiv:2103.15320*, 2021.
[17] Y. min Tang, Y. hong Wang, X. yu Feng, Q. sha Zou, Q. Wang, J. Ding, R. C. jin Shi, and X. Wang, "Diagnostic value of a vision-based intelligent gait analyzer in screening for gait abnormalities," *Gait & Posture*, vol. 91, pp. 205–211, 2022.
[18] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2334–2343.
[19] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve bayes algorithm," *Knowledge-Based Systems*, p. 105361, 2020.
[20] M. P. LaValley, "Logistic regression," *Circulation*, no. 18, pp. 2395–2399, 2008.
[21] S. Suthaharan, "Support vector machine," in *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*. Springer, 2016, pp. 207–235.
[22] N. Eichler, S. Raz, A. Toledano-Shubi, D. Livne, I. Shimshoni, and H. Hel-Or, "Automatic and efficient fall risk assessment based on machine learning," *Sensors*, vol. 22, no. 4, p. 1557, 2022.
[23] U. K. Kandagatla, "Fall detection dataset," 2021.
[24] S. Patro and K. K. Sahu, "Normalization: A preprocessing stage," *arXiv preprint arXiv:1503.06462*, 2015.
[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[26] S. Sharma, S. Sharma, and A. Athaiya, "Activation functions in neural networks," *Towards Data Sci*, vol. 6, no. 12, pp. 310–316, 2017.
[27] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," *arXiv preprint arXiv:1612.02295*, 2016.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[29] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, "Vision transformers for image classification: A comparative survey," *Technologies*, vol. 13, no. 1, 2025.
[30] A. Mao, M. Mohri, and Y. Zhong, "Cross-entropy loss functions: Theoretical analysis and applications," in *International conference on Machine learning*. PMLR, 2023, pp. 23 803–23 828.
[31] Z. Zhang, "Improved adam optimizer for deep neural networks," in *International symposium on quality of service (IWQoS)*. Ieee, 2018, pp. 1–2.