2025 3rd International Conference on Computer Graphics and Image Processing (CGIP) | 979-8-3315-3790-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/CGIP66596.2025.00028

Fall Detection Using Vision Transformer

Nitu Saha
Department of Computer Science and Engineering (IoT, CS, BT), University of Engineering and Management, Kolkata, India
nitu.saha@uem.edu.in

Tanmoy Kundu
Atria Institute of Technology, India
tanmoykundu1024@gmail.com

Arunima Banerjee
Department of Computer Science and Engineering (IoT, CS, BT), University of Engineering and Management, Kolkata, India
banerji.oru@gmail.com

Ahona Ghosh*
Department of Computer Science and Engineering (IoT, CS, BT), University of Engineering and Management, Kolkata, India
ahonaghosh95@gmail.com
*Corresponding author

Aritrik Ghosh
Department of Computer Science and Engineering (IoT, CS, BT), University of Engineering and Management, Kolkata, India
aritrik.ghosh2005@gmail.com

Abstract*—One of the main reasons older people get hurt and end up in the hospital is falls, which can have long-term effects on their physical and mental health. With the widening ageing population worldwide, the demand for advanced and reliable fall detection systems becomes progressively important in supporting independent life and timely medical intervention. Traditional solutions, like wearable sensors or manual observation, tend to have low accuracy or are even characterized by usability challenges, particularly for elderly people who might be hesitant to wear or forget such devices. The paper describes an automated approach to human fall detection using a Vision Transformer (ViT) model. This deep learning model has recently proven to be a strong contender over convolutional neural networks (CNNs) for vision tasks. It is trained on images depicting fall and non-fall activities. All images are normalized and scaled prior to being fed into the ViT model, which classifies the image as either a fall or non-fall class. The ability of ViT to capture long-range interdependence and global spatial features is especially effective in understanding human movement between frames. Our findings suggest that the method puts high accuracy and generalization capacity to the test data, even when under different lighting conditions and at different locations. This model is a strong choice for real-time falling detection in smart surveillance systems and care homes for elderly.

Keywords*— Deep learning, human fall, elderly care, Vision Transformer

I. Introduction

Falling is one of the major causes of injury in the elderly, commonly putting their safety, health, and independence at risk. The World Health Organisation (WHO) quotes that millions of older people fall every year. When one falls, immediate help greatly minimises the effect of the injury. That is why automatic fall detection systems are becoming more and more valuable—they provide a means to track individuals in environments such as hospitals and nursing homes, ensuring timely help when help is most needed [3]. A typical fall detection system relies primarily on wearable sensors like accelerometers and gyroscopes [4]. The technologies tend to be wearable, which is uncomfortable and needs human intervention to operate successfully. Vision-based systems offer a non-intrusive approach to processing camera data, thus removing the need for human interaction. Further vision-based approaches utilize classic machine learning models, which extract features from silhouettes or motion patterns [5, 6]. With the emergence of deep learning, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have gained popularity due to their superior performance in capturing time-based and space-based features from video [7, 8]. However, these designs' local receptive fields and sequential structure may restrict their ability to describe sophisticated global interactions inside frames.

Transformer-based models, such as the Data-efficient Image Transformer (DeiT), Shifted Window Transformer (Swin), and Pyramid Vision Transformer (PVT), outperform in various vision tasks by successfully capturing global dependencies through self-attention processes. As they are not that old, their application in fall detection is still unexplored. Vision Transformers (ViTs) have attracted s lot of application areas for their impressive performance in image classification problems, and they're also showing promise in transfer learning. In contrast to conventional models, ViTs split an image into patches and process them through self-attention, that allows the model to learn the overall structure of the scene better. This capability of grasping long-range relationships throughout an image is particularly useful for fall detection, as distinguishing between the spatial location of various body parts is critical for effective detection. Furthermore, ViTs provide flexible and scalable structure, making them a suitable choice for enhancing recognition within niche domains with sparse data—something that is typically true in fall detection research [9].

In this work, we investigate the application of a pre-trained ViT model for fall detection. To make the model learn the task at hand, we fine-tuned it over a labelled dataset of both fall and non-fall images. We then benchmarked its performance against common classification metrics and compared it with those for other transformer-based models. This method not only analyzes how good ViTs are in tackling fall detection complexities but also showcases their applications in real-world scenarios [9]. Furthermore, to describe the VIT output, Gradient-weighted Class Activation Mapping (GradCAM) has been used.

979-8-3315-3790-6/25/$31.00 ©2025 IEEE
DOI 10.1109/CGIP66596.2025.00028

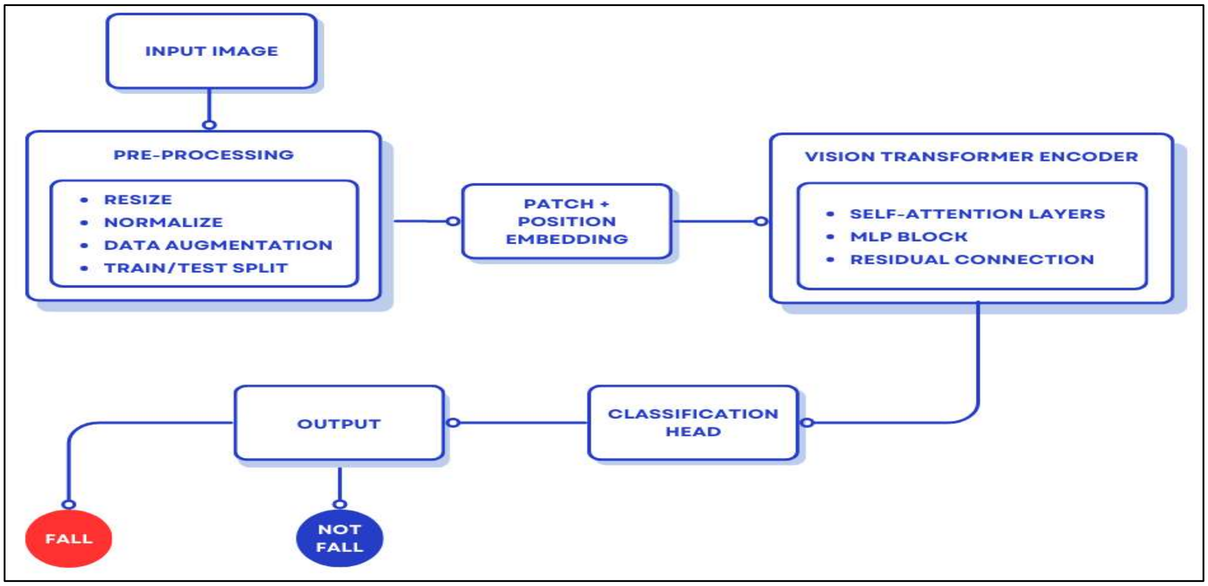

Fig. 1. Proposed architecture

Section II describes the suggested framework, Section III discusses the experimental outcomes and performance assessment, and Section IV recapitulates the research with potential future work directions.

II. Proposed Methodology

The model's suggested architecture identifies falls from a static image in two steps: data preprocessing and classification with a Vision Transformer. Subtle posture changes and body orientations must be recognized to detect a fall, and the model must understand local and spatial patterns. ViTs utilize self-attention to process image patches, allowing them to capture relationships across the entire image. This methodology enables the model to recognize fall-related cues in diverse visual contexts. In this work, we have trained and fine-tuned a labelled dataset of fall and non-fall images.

A. Preprocessing

The input dataset comprises images of various scenarios, each categorized into one of two classes: 'fall' or 'non-fall'. Each image is resized to 224✕224 pixels. The resizing is done to standardize the input size for the model. Each image undergoes several preprocessing steps, ensuring that the model receives consistently formatted input and improves generalization during training. The images are at first resized and cropped to 224×224 dimensions. Then those are normalized using ImageNet's mean and standard deviation values. Random horizontal flipping has been performed for data augmentation during training. Images are then split into training and testing sets, with batch processing handled via data loaders.

B. Classification using Vision Transformer

The proposed system has a core similar to the ViT-B/16 architectures pre-trained on the ImageNet dataset. A fully connected layer designed to output two logits, which significantly corresponds to the binary classification task, is how the original classification head is being replaced. The final classification head needs fine-tuning, and an optional step includes freezing the initial transformer layers. A one-dimensional array of token embeddings is input into the conventional Transformer model. The picture $x \in R^{H \times W \times C}$ has been restructured to accommodate 2D images by generating a series of flattened 2D patches $x_p \in R^{N \times (p^2 C)}$, where ($H$, W) denotes the original image's resolution, C represents the number of channels, (P, P) indicates the resolution of each image patch, and $N = \frac{HW}{P^2}$. The resultant quantity of patches also serves as the effective input sequence length of the transformer. The patches are flattened and mapped to D dimensions by a trainable linear projection, as the transformer maintains a consistent latent vector size D throughout all its layers. The outcome of this projection is referred to as the patch embeddings. We add a learnable embedding to the series of embedded patches ($z_0^0 = x_{\text{class}}$), with its state at the output of the Transformer encoder (z_0^L) serving as the image representation y. This is analogous to BERT's [class] token. Throughout pre-training and fine-tuning, z_0^L is linked to a classification head. The classification head employs a MultiLayer Perceptron (MLP) consisting of a single linear layer for fine-tuning and one hidden layer for pre-training. Position embeddings are added to the patch embeddings to maintain positional information. Given that advanced 2D-aware position embeddings have not demonstrated significant performance improvements, standard learnable 1D position embeddings are utilized. The encoder accepts the generated sequence of embedding vectors as input. The Transformer encoder is composed of alternating MLP and multi-headed self-attention (MSA) blocks. Every block is preceded by LayerNorm (LN) and succeeded by residual connections. Cross-Entropy Loss and Adam optimizer train the model.

Fig. 2. Samples of images considered

The model is highly suitable for tasks such as identifying postures and anomalies related to falls, due to the attention mechanism present within ViT. This gives the model a different kind of focus, which works only on spatial regions of the input.

C. Explanation using GradCAM

GradCAM understands and visualizes how a ViT makes decisions by highlighting an input image's important sections contributing most to a model's prediction. In CNNs, Grad-CAM computes the gradient of the output class score with respect to the feature maps of a convolutional layer. These gradients are global-average-pooled to get importance weights. The weighted combination of feature maps is passed through ReLU to produce a heatmap showing important image regions. But ViTs use self-attention over patch embeddings (image split into patches) and they don't have spatial feature maps like CNNs. In this case, a class token is output that aggregates information across patches. It chooses a transformer block (usually the last one) and focuses on attention or MLP outputs. Intermediate representations like the attention score maps or value vectors can also be picked. It focuses on the [class] token (or equivalent) used for classification.

Then it computes the gradient of the class score (e.g., softmax score) with respect to the selected layer's outputs and extracts the gradients of the class score with respect to patch tokens. Then it multiplies these gradients with the patch token embeddings or attention outputs. This gives an importance score for each patch. Since each patch corresponds to a region in the original image, the patch-level scores are reshaped into a grid (e.g., 14x14 for ViT-B/16). Then it upsamples to the original image size and applies ReLU and normalizes to get a final heatmap. The heatmap highlights the regions that were most influential in making the prediction (e.g., the fall or not fall). The heatmap is overlaid onto the original image, providing an easy-to-interpret visualization. Areas in warmer colors (e.g., red or yellow) show where the model "looked" when making the decision.

III. Experimental Results And Discussion

A series of experiments was conducted to outline the setup used for training and testing the model, as well as presenting the classification outcomes. This is done to evaluate the effectiveness of the proposed ViT-based fall detection framework. Lastly, it offers a comparative performance analysis against other transformer architectures. The primary goal is to assess the model's skill to generalize to unknown fall scenarios. This can be achieved through an experimental process that systematically evaluates its predictions, then fine-tunes the dataset and prepares it for processing. It also aims to validate its performance against state-of-the-art alternatives.

A. Dataset description

The images in the input dataset are taken from public repository Kaggle [24] and represent real-world scenarios. Each image is categorized into either fall or non-fall class. Each image is resized to 224 × 224 pixels and augmented using different techniques. The data is then split into test and train sets to ensure an unbiased evaluation.

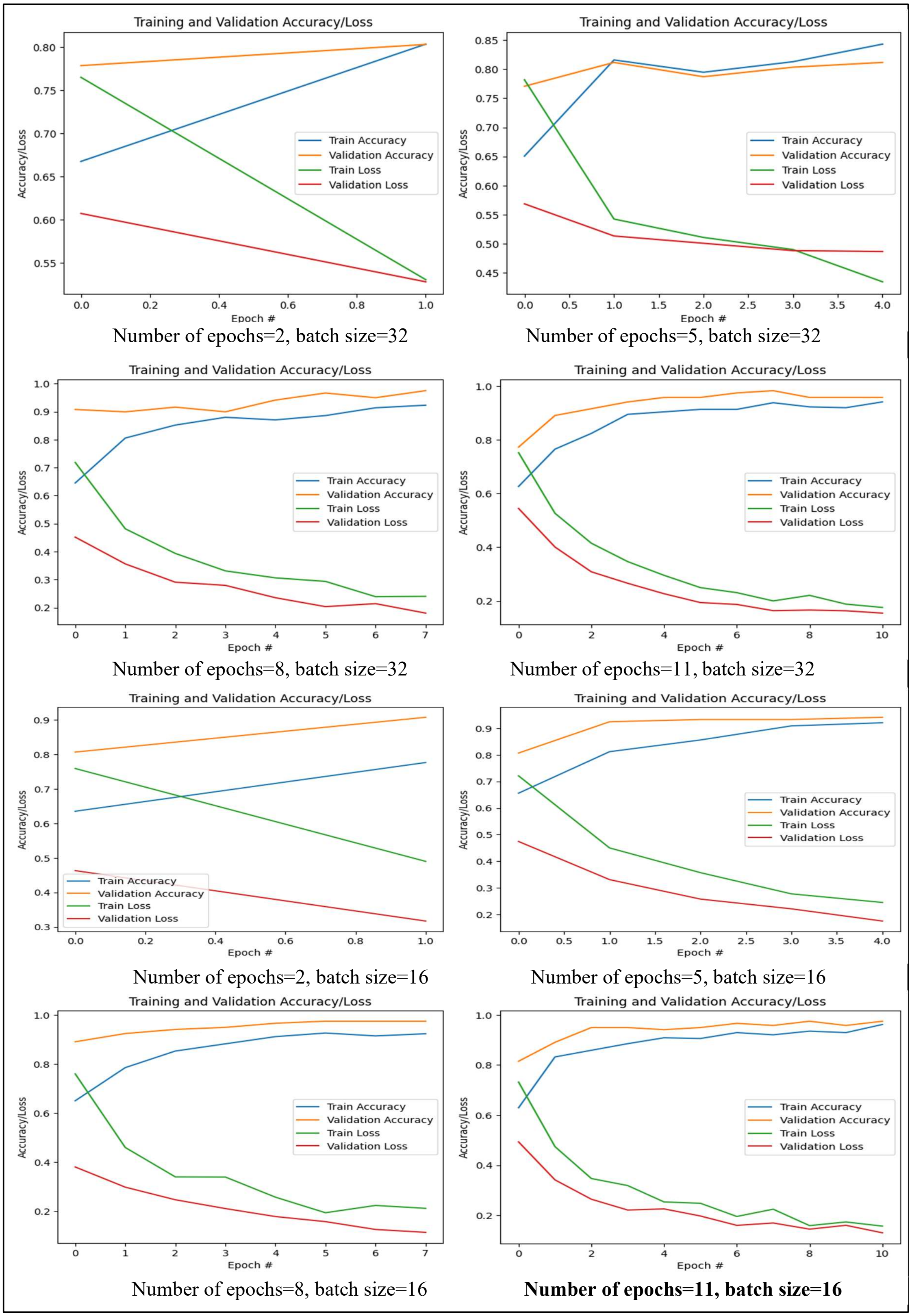

Fig. 3. Learning curves showing hyperparameter tuning results

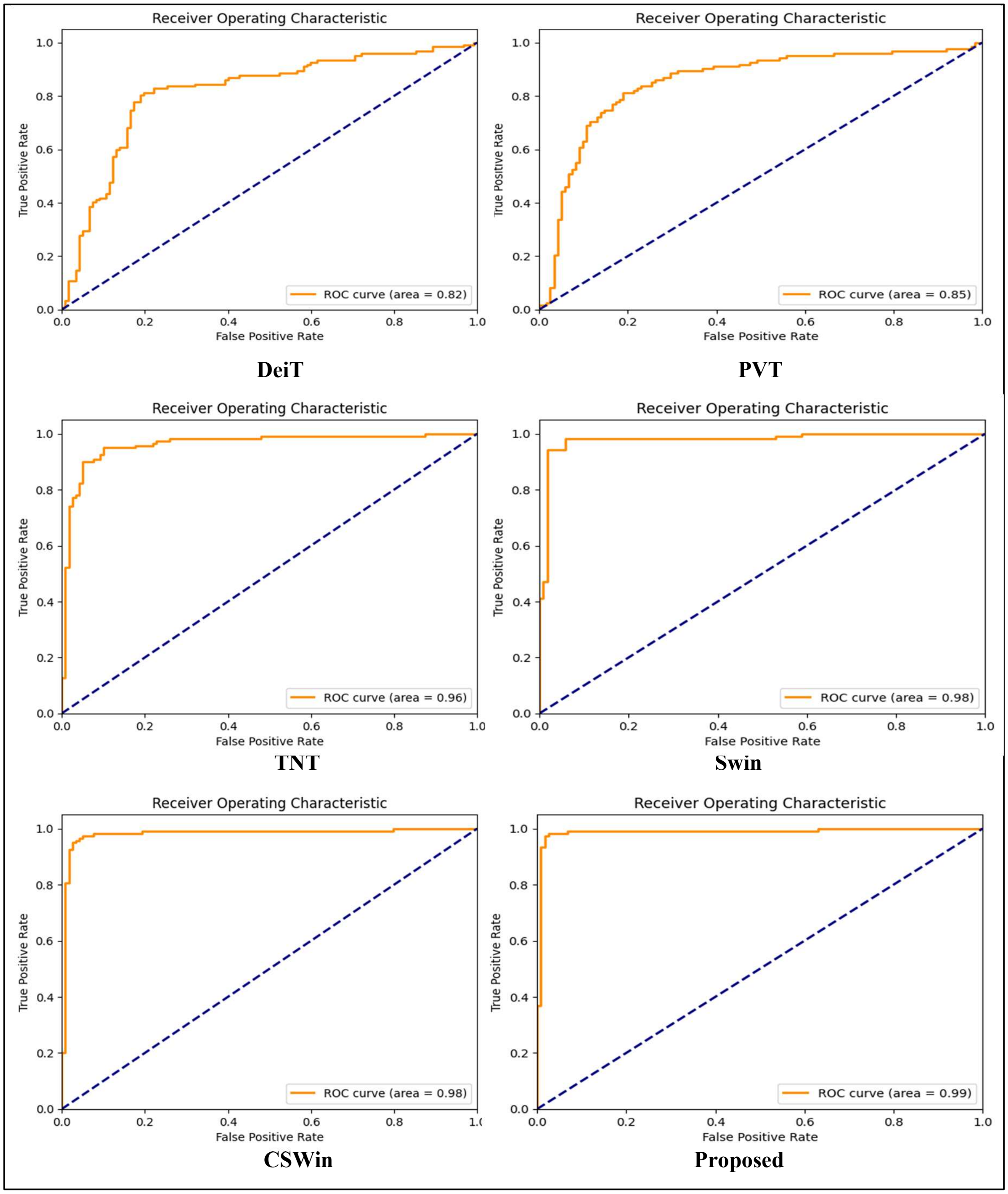

Fig. 4. ROC curves showing comparison with other transformer models

B. Classification results

The effectiveness of the proposed model's hyperparameters for fall detection using image processing has been evaluated first by varying the number of epochs and batch size in different combinations. The batch size, denoting the number of images processed together during training, has varied between 16 and 32, whereas the possible number of epochs has been considered 2, 5, 8, and 11. The learning curves shown in Fig. 3 indicate that as the number of epochs (iterations) increases, the recognition accuracy initially increases rapidly and then gradually slows down. Hence, the number of epochs reaches a specific value, the training and testing accuracy curves converge to the maximum value, and the training and testing loss converges to the minimum one and then stabilizes. The combination of 11 and 16 as the epoch number and batch size has outperformed the others and has been chosen for the proposed model.

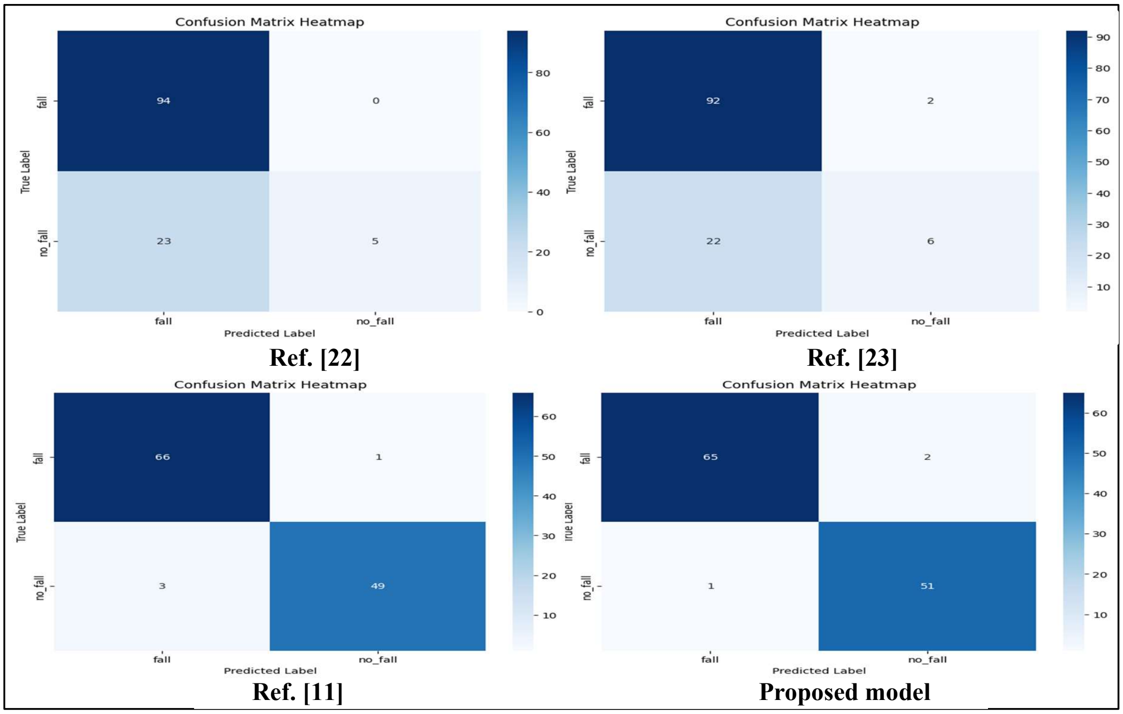

Fig. 5. Confusion matrices comparing the proposed work with existing related works

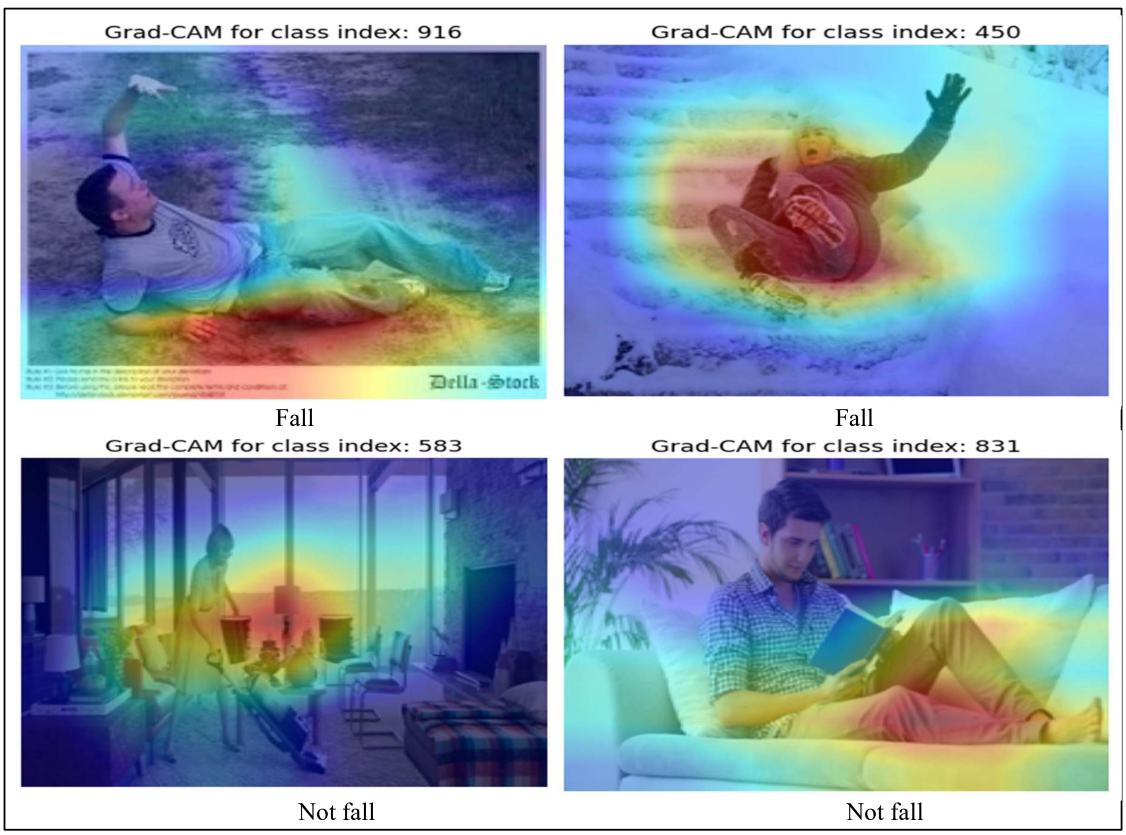

Fig. 6. Decision areas indicated by GradCAM

C. Performance analysis

Comparison of the proposed ViT with other transformer models like DeIT, PVT, Transformer in Transformer (TNT), Swin, and Cross-Shaped Window Transformer (CSWin) has been presented in Fig. 4 using the Receiver Operating Characteristic (ROC) curve. The curve plots the True Positive Rate on the x-axis, representing the number of actual positives properly classified, and the False Positive Rate on the y-axis, denoting the number of actual negatives wrongly classified. The proposed model, with an Area Under the Curve value of 0.99, has outperformed its competitors.

The proposed work has been compared to existing works [9], [10], and [11], and the results have been shown in terms of the confusion matrix in Fig. 5. The confusion matrix illustrates the model's performance, with the diagonal representing the count of accurately anticipated observations; a denser distribution of predicted values along this diagonal signifies superior model performance. The results of the proposed classifier are almost perfect in the confusion matrix, with only 3 points being misclassified and 116 correct observations.

D. Explanation using GradCAM

The heatmap overlaid on the image uses **colors** (e.g., red, yellow, blue) to show the regions that are important for the prediction. The red and yellow sections are the most significant in the model's prediction. In this case, they highlight the person's body, particularly the lower body (which might be relevant to the fall or slipping event). **Blue s**ections are the least significant in the predictions. The heatmap primarily highlights the person's **body** and the **staircase** around where they are falling. This indicates that the model focuses on the key region (the person) involved in the predicted class (e.g., a fall or accident).

IV. Conclusion

One major health problem for the elderly is falls, leading to serious damages and increased dependence. Traditional fall detection methods, like wearable sensors, have usability challenges and are often impractical in real-life scenarios. This paper offers a non-intrusive, vision-based fall recognition system utilizing a Vision Transformer (ViT) model, fine-tuned on a dataset comprising images of falls and non-falls. The ViT architecture features a self-attention mechanism that enables the effective modelling of spatial relationships, allowing it to identify fall-related postures across varied conditions accurately. In our experiments, the model demonstrates high performance, particularly at 11 epochs and a batch size of 16, with ViT in classification accuracy and robustness, outperforming other transformer models such as DeiT, PVT, Swin, and CSWin. In this work, the potential of Vision Transformers is highlighted in safety-critical applications, such as fall detection, which offers a reliable alternative to conventional approaches. The current system is limited to static images. However, it lays the groundwork for future extensions, including video-based detection, lightweight deployment on edge devices, and enhanced datasets to improve generalization. In conclusion, this model, built on ViTs, offers a powerful solution for real-time fall detection, aiming to improve elderly care through intelligent and non-intrusive monitoring.

References

1. Xu, Q., Ou, X. and Li, J., 2022. The risk of falls among the ageing population: a systematic review and meta-analysis. *Frontiers in public health*, *10*, p.902599.
2. Montero-Odasso, M., Van Der Velde, N., Martin, F.C., Petrovic, M., Tan, M.P., Ryg, J., Aguilar-Navarro, S., Alexander, N.B., Becker, C., Blain, H. and Bourke, R., 2022. World guidelines for falls prevention and management for older adults: a global initiative. *Age and ageing*, *51*(9), p.afac205.
3. Alharbi, H.A., Alharbi, K.K. and Hassan, C.A.U., 2023. Enhancing elderly fall detection through iot-enabled smart flooring and ai for independent living sustainability. *Sustainability*, *15*(22), p.15695.
4. Nooruddin, S., Islam, M.M., Sharna, F.A., Alhetari, H. and Kabir, M.N., 2022. Sensor-based fall detection systems: a review. *Journal of Ambient Intelligence and Humanized Computing*, *13*(5), pp.2735-2751.
5. H. Nait-Charif and S. J. McKenna, "Activity summarisation and fall detection in a supportive home environment," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, Cambridge, UK, 2004, pp. 323–326.
6. D. Anderson, R. H. Luke, J. M. Keller, M. Skubic, M. Rantz, and M. Aud, "Linguistic summarisation of video for fall detection using voxel person and fuzzy logic," *Comput. Vis. Image Underst.*, vol. 110, no. 2, pp. 160–176, May 2008.
7. J. Cheng, Y. Wang, and G. Zhou, "A novel fall detection method based on deep convolutional neural network," *IEEE Access*, vol. 6, pp. 61025–61033, 2018.
8. Y. Li, S. Mikami, and Y. Otani, "Fall detection using LSTM neural network with deep features extracted from skeleton images," *Sensors*, vol. 20, no. 16, p. 4491, Aug. 2020.
9. Yunusa, H., Qin, S., Chukkol, A.H.A., Yusuf, A.A., Bello, I. and Lawan, A., 2024. Exploring the synergies of hybrid CNNs and ViTs architectures for computer vision: A survey. *arXiv preprint arXiv:2402.02941*.
10. Mubashir, M., Shao, L., & Seed, L. (2013). A survey on fall detection: Principles and approaches. Neurocomputing, 100, 144–152.
11. Rougier, C., Meunier, J., St-Arnaud, A., & Rousseau, J. (2007). Fall detection from human shape and motion history using video surveillance. International Conference on Advanced Information Networking and Applications Workshops.
12. Kwolek, B., & Kepski, M. (2014). Human fall detection on embedded platform using depth maps and wireless accelerometer. Computer Methods and Programs in Biomedicine, 117(3), 489–501.
13. Li, Y., Li, W., & Zhang, T. (2019). Fall detection for elderly by depth images using convolutional neural networks. Computer Methods and Programs in Biomedicine, 171, 31–39.

14. Nweke, H. F., et al. (2018). Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. Expert Systems with Applications, 105, 233–261.
15. Cheng, C., et al. (2020). Human fall detection using hybrid deep neural network architecture. Sensors, 20(4), 1214.
16. Touvron, H., et al. (2021). Training data-efficient image transformers & distillation through attention. ICML.
17. Wang, W., et al. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. ICCV.
18. Liu, Z., et al. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. ICCV.
19. Han, K., et al. (2021). Transformer in Transformer. NeurIPS.
20. Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? ICML.
21. Li, Y., et al. (2022). Video Swin Transformer. CVPR.
22. Chhetri, S., Alsadoon, A., Al-Dala'in, T., Prasad, P.W.C., Rashid, T.A. and Maag, A., 2021. Deep learning for vision-based fall detection system: Enhanced optical dynamic flow. *Computational Intelligence*, *37*(1), pp.578-595.
23. Kumar, V., Badal, N. and Mishra, R., 2021. Elderly fall detection using IoT and image processing. *Journal of Discrete Mathematical Sciences and Cryptography*, *24*(3), pp.681-695.
24. Kumar kandagatla (2021), *Fall Detection Dataset*, available at https://www.kaggle.com/datasets/uttejkumarkandagatla/fall-detection-dataset (Accessed: 09 July 2025).