XAI for Transformer-based Fall Detection using Skeleton Time-series Data Analysis

Annaclaudia Bono[†1,2], Cosimo Patruno[1], Vito Renò [1], Grazia Cicirelli [1], Cataldo Guaragnella[2]

***Abstract*— Accurate fall detection is increasingly important for continuous monitoring in home and workplace environments. Recent advances have focused on vision-based systems and skeleton-based representations, exploiting Transformer-based architectures to model complex spatio-temporal relationships. However, the lack of transparency of these models limits trust and practical deployment. To address this challenge, this paper employs ShaTS, a variant of the SHAP technique, as a post-hoc, model-agnostic explainability method to interpret and understand the predictions of the Transformer-based model. ShaTS is used to quantify the global contributions of joints and temporal segments on the model's predictions. The results show that explainability can reveal meaningful motion cues, support domain-level validation, and guide future model refinement, ultimately enhancing confidence in the model's decision-making process.**

Keywords — Explainability, Transformer architecture, Fall detection, Human monitoring, ShaTS

I. INTRODUCTION

Research on fall detection has increased with the growing need of continuous monitoring in both domestic and workplace environments [1]. Vision-based systems have become dominant [2] as they exploit algorithms that capture both spatial postures and temporal dynamics [3]. However, misclassification can occur due to occlusions, partial visibility and ambiguous actions, such as bending or sitting. Skeleton-based representations mitigate these issues by improving robustness to lighting variations, background clutter, and privacy concerns. Recent advances of deep learning techniques, such as Transformer architectures, further enhance fall detection approaches by effectively modeling long-range temporal dependencies and joint interactions. Despite this progress, several challenges still remain. Despite these improvements, current methods are often trained on limited datasets that fail to capture real-world variability [1], reducing their reliability. Furthermore, the growing complexity of Transformer-based architectures makes them difficult to interpret [4]. This lack of transparency is a limit in safety-related applications such as fall detection, motivating the adoption of Explainable Artificial Intelligence (XAI) techniques to provide human-understandable insights about model behavior.

XAI methods can be categorized into intrinsic and post hoc approaches [5]. Intrinsic approaches rely on inherently interpretable models, whereas post hoc techniques aim to explain black-box systems without requiring access to their internal parameters. These methods can provide local or global explanations and may be model-specific or model-agnostic, depending on the application context. Given the architectural complexity of Transformer models and the lack of well-established architecture-specific explainability techniques, interpreting their decisions remains challenging. Although attention mechanisms are often intuitively associated with explanation, attention should not be identified with interpretability [6] since attention weights describe how the model distributes internal focus across inputs, but they do not necessarily reflect the features that truly determine the final prediction. For this reason, model-agnostic approaches such as SHAP (SHapley Additive exPlanations) [7] can offer a more reliable framework for analyzing model behavior.

However, even with model-agnostic techniques such as SHAP, additional challenges arise when explaining predictions for sequential data. These methods typically treat each time step as an independent feature, overlooking temporal dependencies and potentially producing incomplete or misleading explanations. In addition, the exact computation of Shapley values is computationally expensive, limiting their suitability for real-time applications. Standard Shapley-based approaches also provide importance scores at the level of individual data points, resulting in fragmented and less actionable explanations. These limitations can compromise the practical usability of explanations in operational contexts. Therefore, there is a need for methods specifically designed to preserve the temporal coherence and structural organization of time-series data. ShaTS (Shapley values for Time Series models) proposed in [8] is a model-agnostic, open-source XAI method developed specifically to address the limitations of conventional Shapley-based techniques for sequential data. Released as a fully documented, plug-and-play module, ShaTS is designed to improve both the precision and practical usefulness of explanations for time-series models. Its core innovation is an a priori feature-grouping strategy, in which features are grouped before importance values are computed. Unlike traditional approaches that first assign importance to every individual time point and later attempt to aggregate them, ShaTS performs grouping at the beginning, preserving temporal dependencies and logical relationships among features. This design leads to more coherent, context-aware explanations that are immediately interpretable and actionable for operators in real-world settings.

In this paper, ShaTS is employed to explain the predictions of a Transformer-based model for skeleton-based fall detection, providing a human-interpretable view of its decision process. Specifically, this XAI technique quantifies the contribution of joints and temporal segments to the final classification, acting as a post-hoc explainability layer that complements the Transformer's capacity to model complex spatio-temporal dependencies. While the Transformer learns fine-grained relationships across frames and joint coordinates, ShaTS operates a posteriori to summarize which body parts and motion phases most influence the predictions. This post-hoc analysis enables a comprehensive interpretation of model behavior without altering the underlying learning process. The results demonstrate how explainability can serve as an analytical tool to extract meaningful spatio-temporal insights, supporting domain-level validation, verifying reliance on relevant motion cues, and guiding future model refinement and feature design.

II. MATERIALS AND METHODS

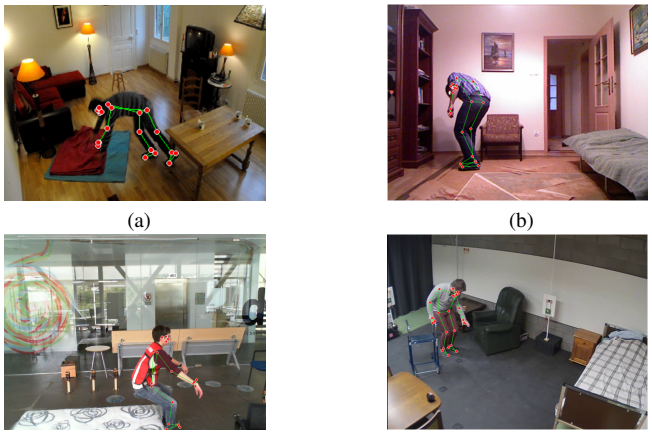
A. Dataset

Detecting falls from video sequences is challenging in complex environments, especially when obstacles partially occlude the subject. The extraction of 2D human body landmarks offers a good

[1] Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing (STIIMA) of the National Research Council (CNR), Bari, Italy.

[2]Polytechnic University of Bari, Department of Electrical and Information Engineering (DEI), Bari, Italy

[†]Corresponding author: annaclaudia.bono@cnr.it



(a) (b) (c) (d)

Fig. 1: Frame samples from (a) Le2i, (b) URFD, (c) FALL-UP and (d) High-Quality datasets, with extracted skeletons overlaid.

solution abstracting the problem from the raw images and reducing sensitivity to resolution, illumination, and other environmental variations [9], [10]. However, the effectiveness of landmark-based models still depends on the quality and diversity of the available data. Most existing fall-detection datasets are small and collected under controlled conditions, limiting their ability to represent real-world scenarios. Moreover, they provide videos usually annotated with only a single fall/no-fall label, overlooking the temporal dynamics of human motion. To address these limitations, this work constructs a dataset by integrating four publicly available and thematically relevant fall detection datasets. Specifically, the datasets are Le2i [11], UR Fall (URFD) [12], FALL-UP [13], and High-Quality Fall Simulation [14]. Figure 1 shows one sample frame from each dataset, with the extracted skeletons superimposed.

To enhance data expressiveness and handle the varying sequence lengths across datasets, each video sequence was segmented into overlapping sub-sequences of 90 frames using a sliding window with a 15-frame step size. For each segment, 2D body landmarks were extracted using the open-source Google MediaPipe framework [15] and annotated as either Fall or Activity of Daily Living (ADL). To reduce computational complexity, only 11 of the 33 detected joints (front face, shoulders, wrists, hips, knees, and ankles) were retained (Figure 2). From each frame, a feature vector is constructed containing 11 $< x, y >$ joint coordinates and an additional feature, the Body Aspect Ratio (BAR), defined as the ratio between maximum body height and maximum body width, which defines the overall body orientation. As a result, each sample is represented as a 23-dimensional multivariate time series of length 90, capturing the spatial configuration and posture dynamics of the body over about three seconds at 30fps. The final dataset consists of 7308 labeled samples, each represented as a 90×23 temporal sequence.

B. Transformer Architecture

The fall detection task is formulated as a multivariate time-series classification problem, where each input sequence captures the temporal evolution of 2D body poses. The Transformer architecture was adopted since it enables efficient parallel processing while capturing both local and global relationships in the temporal data. In detail, the input is first mapped to a higher-dimensional space and processed by two encoder blocks. A pooling operation aggregates the contextual information across the entire sequence, after which a series of linear layers followed by a sigmoid function produce the final classification output [16].

Transformers are well-suited for modeling time series, but they present significant interpretability challenges. Their distributed attention mechanisms and deep representation layers make it difficult to trace predictions back to specific temporal events or input features, leading to explanations that are difficult to interpret [17], [18]. This limitation is particularly critical in safety-sensitive applications such as fall detection, where understanding when and which body dynamics influence a decision is essential [6]. These considerations underline the need for explainability methods to make model predictions more transparent and accessible to end users.

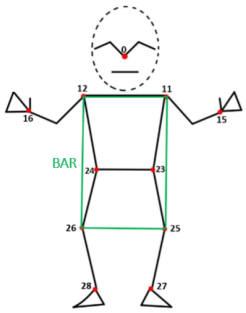


Fig. 2: Visualization of the selected subset of body landmarks extracted using MediaPipe and the BAR feature.

C. SHAP Tecnique

Understanding how a model makes its predictions is crucial in safety-critical applications such as fall detection. In this context, eXplainable Artificial Intelligence (XAI) techniques are useful for interpreting model behavior. Among XAI techniques, SHapley Additive exPlanations (SHAP) [7] quantify the contribution of individual input features to a model's prediction. This method is based on Shapley values [19], originally formulated within cooperative game theory. These values quantify the fair contribution of each player to a collective outcome. Similarly, in machine learning, features act as players and the model's prediction represents the payoff. Shapley values assign a numerical score to each feature, indicating how strongly it influenced the prediction. By considering all possible combinations of features, SHAP provides a comprehensive and theoretically sound measure of feature importance.

Let N denote the set of all input features, the Shapley value for feature i is defined as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left(\nu(S \cup \{i\}) - \nu(S)\right) \quad (1)$$

where S represents a coalition of features not containing i, and $\nu(S)$ denotes the value associated with the coalition S. The term $\nu(S \cup \{i\}) - \nu(S)$ captures the marginal contribution of feature i when added to the coalition S, while the weighting factor $\frac{|S|!\,(|N|-|S|-1)!}{|N|!}$ accounts for all possible permutations of feature.

A key advantage of Shapley values is that they satisfy a set of desirable theoretical properties:

1) *Local accuracy:* the sum of the feature attributions exactly accounts for the difference between the model prediction for a given input x and the baseline prediction. Formally,

$$\sum_{i=1} \phi_i = f(x) - \mathbb{E}[f(X)]$$

where $\mathbb{E}[f(X)]$ is the expected model output computed over a background dataset used as a reference distribution for Shapley value estimation.

2) *Missingness:* if a feature is not present in the input representation, it receives zero attribution. This guarantees that only observed features contribute to the explanation.

3) *Consistency:* if a model is modified such that the marginal contribution of a feature increases (or remains unchanged) for all coalitions, then the attribution assigned to that feature cannot decrease.

D. ShaTS formulation

SHAP is model-agnostic and therefore can be applied to a wide range of algorithms. However, it exhibits several limitations when used with time-series data, as it does not explicitly capture temporal dependencies, treating features independently across time. Moreover, SHAP lacks mechanisms to aggregate features associated with the same physical entity, such as body joints, which may limit the interpretability of resulting explanations. To address these limitations, in [8] a Shapley-based XAI method (ShaTS) is introduced, designed for time-series data. ShaTS includes an a priori feature grouping strategy, enabling explanations at multiple levels of abstraction. In detail, ShaTS supports: (i) Temporal Grouping, which assesses the relevance of specific time steps within a sequence; (ii) Feature Grouping, which evaluates the contribution of individual features across time; and (iii) Multi-Feature Grouping, which aggregates related features corresponding to the same physical component. This design yields explanations that are temporally coherent, physically meaningful, and easier to interpret. Moreover, by reducing the dimensionality of the attribution space through feature grouping, ShaTS significantly improves computational efficiency compared to traditional SHAP-based approaches. In this work, ShaTS is employed to generate both local and global explanations of the model's predictions, enabling a more reliable interpretation of the learned temporal patterns and strengthening trust in the proposed fall detection system.

Formally, the ShaTS formulation follows the classical definition of Shapley values, with the key difference that attributions are computed over groups of features rather than individual features. Let G denote the set of feature groups, where each group $G_i \in G$ replaces a single feature in the standard Shapley formulation. The ShaTS value ϕ_{G_i} quantifies the average marginal contribution of group G_i to the model's output across all possible subsets of the remaining groups and is defined as:

$$\phi_{G_i}(\nu) = \sum_{T \subseteq G \setminus \{i\}} \frac{|T|!\,(|G| - |T| - 1)!}{|G|!} \left[\nu(T \cup \{G_i\}) - \nu(T)\right] \tag{2}$$

In this work, ShaTS is preferred over standard SHAP because it enables temporally and semantically coherent feature aggregation. Since falls are defined by the evolution of the entire movement sequence rather than isolated time instants or individual joint coordinates, ShaTS' a priori grouping strategy better captures the structured spatio-temporal dynamics learned by the Transformer.

III. EXPERIMENTS

The experiments have been performed on a computer equipped with an Intel Core i7-9700K @ 3.60 GHz processor, with 16 GB RAM, and a 24 GB NVIDIA GeForce RTX 4090 GPU.

A. Transformer Results

The Transformer model was trained on 90-frame sub-sequences derived from the joint features extracted across all selected datasets. The model was trained using a low learning rate of 10^{-4} to ensure stable optimization and avoid abrupt parameter updates. To further enhance convergence, the learning rate was progressively reduced every 50 epochs, enabling finer optimization during the later stages of training. The training process was conducted for 500 epochs with a batch size of 32. To mitigate overfitting, an early stopping strategy, based on a patience parameter, was adopted. Specifically, the patience was set to 20, meaning that training was interrupted if the validation loss, computed using the BCE loss function, did not improve for 20 consecutive epochs. As shown in Table I, the model achieves an accuracy of 98.97%, demonstrating reliable classification on the considered dataset. Precision (98.27%) and recall (98.03%) are both high and well-balanced, indicating that the model effectively limits false positives while correctly identifying most fall events. The F1 score of 98.15% further confirms this balance, proving that the model does not favor one metric over the other, an important aspect in binary classification contexts.

TABLE I: Test results of the Transformer model after training on the four datasets jointly. The performance metrics evaluated during inference are reported.

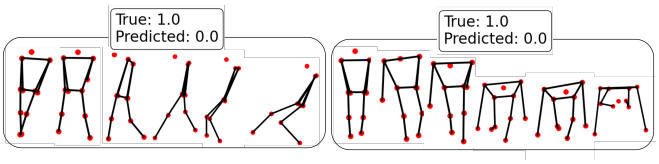
Accuracy ↑	Precision ↑	Recall ↑	F1-Score ↑
98.96%	98.27%	98.02%	98.15%

Figure 3 shows two examples of misclassifications. In the first example (Figure 3(a)), the model fails to detect the fall because the sequence ends just at the beginning of the fall. This leads to a discrepancy because while a human can intuitively infer the continuation of the movement and correctly label the sequence as a fall, the model that relies solely on the information contained within the provided frames interprets the initial motion as part of normal activity. A similar situation occurs in the second example (Figure 3(b)), where the subject stands and then bends forward without actually falling. In conclusion, the model in some cases fails because it does not always capture the full temporal evolution of the action within the analyzed sequence. However, when the subsequent frames are included and the movement continues, the fall is correctly detected.

B. ShaTS for Explainable Fall Detection

As widely explained in section II-D, ShaTS can handles time series data supporting feature grouping strategies and enabling different levels of abstraction. Since the data exhibit both temporal and spatial dependencies, three grouping strategies were adopted:

1) **Time Grouping**: each group corresponds to a single frame within the sequence. Since each sample contains 90 frames, this results in 90 groups, each including all the features observed at a given instant. This grouping strategy allows the evaluation of which frames contribute most to the model's final decision.
2) **Body-Parts Grouping**: joints are aggregated into five biomechanical regions: head, upper limbs, hips, and lower limbs, with an additional group dedicated to the BAR feature. This level of abstraction enables the assessment of which major body components most influence the model's decision.
3) **Joint Grouping**: features are grouped according to individual body joints. Each joint is represented by a pair of spatial coordinates $< x, y >$, which are then considered together within the same group. An additional group is defined for the BAR feature, resulting in a total of 12 groups. This strategy enables the evaluation of how the information associated with each joint contributes to the final prediction.



(a) (b)

Fig. 3: Examples of two false-negative predictions produced by the Transformer model: (a) and (b). For visualization purposes, only 6 out of the 90 frames per sequence are shown.

At the beginning of the experiments, a background dataset was defined to represent the model's average behavior and to compute the reference distribution (baseline) for the Shapley values as introduced in section II-C. It consists of 200 samples randomly selected from the training set, preserving the original fall/non-fall ratio. In our case, the baseline is $E[f(X)] = 0.28$, corresponding to the average predicted probability of a fall. For each test sequence x, the model provides a fall probability $f(x)$. The difference $f(x) - E[f(X)]$ represents how much the prediction deviates from the model's average behavior. ShaTS explains this deviation by decomposing it into contributions ϕ_i associated with the feature groups (e.g., body parts, joints, or temporal segments), such that their sum reconstructs the total deviation. A positive contribution $\phi_i > 0$ indicates that the corresponding group pushes the prediction above the baseline, increasing the likelihood of a fall, whereas a negative value $\phi_i < 0$ indicates evidence against a fall.

Figure 4 shows the distribution of predicted fall probabilities across the test set together with the baseline. Most predictions are concentrated near 0 or 1, indicating high confidence, while the explainability analysis focuses on how different feature groups contribute to both confident and borderline cases.

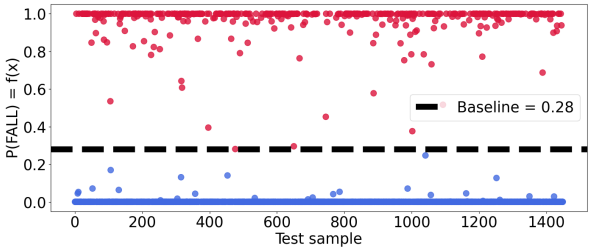


Fig. 4: Distribution of the model output $f(x) = P(\text{FALL} \mid x)$. The dashed line represents the baseline (0.28). Points are colored in red when $f(x) >$ baseline and in blue otherwise.

Figure 5 reports the distribution of ShaTS values (ϕ) across the test set, considering the different grouping strategies. At the temporal level (Figure 5(a)), the highest contributions are observed in the final part of the sequence, approximately between frames 60 and 90, indicating that the model primarily relies on motion patterns occurring during the fall phase rather than on the initial postures. The presence of both positive and negative values reflects the fact that different samples contribute differently depending on whether the sequence corresponds to fall or non-fall event. Overall, the final frames exhibit the highest variability, confirming their discriminative role in the classification process. At the body-parts level (Figure 5(b)), the upper limbs show the widest dispersion and largest ShaTS values, while the hips and lower limbs exhibit moderate contributions. The BAR feature is approximately zero, indicating minimal influence on the predictions. This behavior can be justified by the fact that the BAR is a deterministic function of the shoulder and knee joint coordinates, which are already provided as input to the model. Consequently, when these joint coordinates are available, including or excluding BAR does not significantly affect the model output. Overall, these results indicate that the model identifies a fall primarily as an event characterized by the collapse of the head and pronounced movements of the upper body, which are the segments undergoing the changes during a real fall. This observation is further supported by the joint-level analysis (Figure 5(c)), which shows that the joints associated with the head and shoulders dominate the explanation. In conclusion, the Transformer detects a fall as a temporal event characterized by changes in the final frames and by pronounced movements of the head and upper limbs, which generate the largest and most variable Shapley attributions.

Considering the results of the temporal grouping, the Transformer model was retrained using only the last 30 frames, where

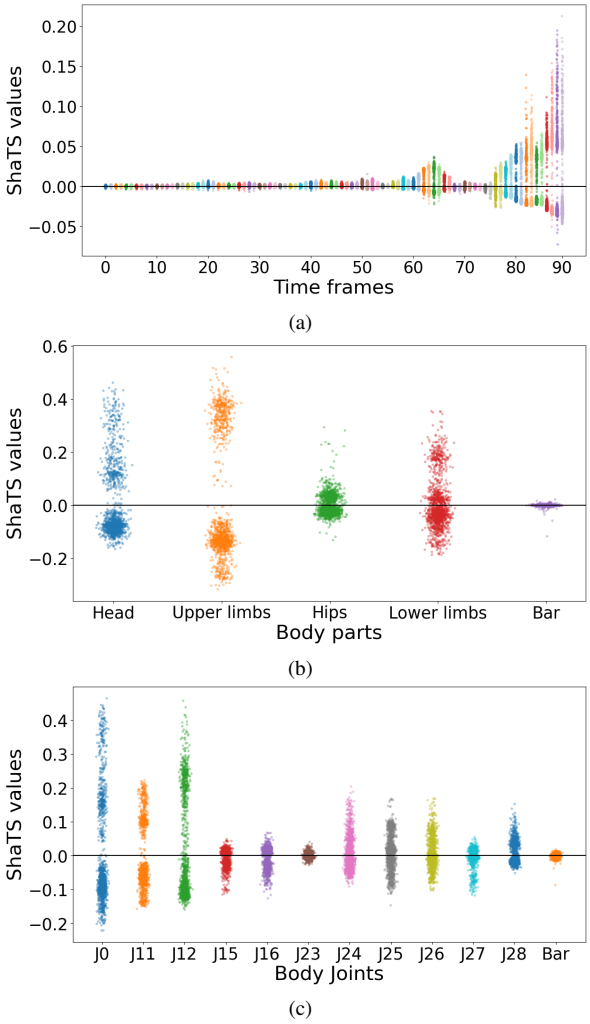


Fig. 5: Distribution of ShaTS values across test samples for each grouping strategy. Each point represents the Shapley contribution of each group, i.e. time frames (a), body part (b) and body joints (c), to the model's prediction.

attributions were highest, and the ShaTS values were recomputed on the updated model. This experiment aims to verify whether the final segment of the sequence is sufficient for accurate fall detection. Results show that the model trained on 30-frame sequences achieve performance comparable to the 90-frame configuration, with nearly identical accuracy and F1 score as shown in Figure 6. A slight increase in precision accompanied by a small decrease in recall indicates that the model reduces false positives at the cost of missing additional falls. This result demonstrates that, although the final frames contain most of the discriminative information, the 90-frame configuration remains preferable in practice. The initial motion dynamics contribute to higher recall, which is critical in fall detection scenarios where missing a true fall is more risky than generating a false alarm. The new ShaTS explanations confirm a consistent temporal pattern: attributions progressively increase toward the final frames (Figure 7), indicating that the decision is primarily driven by the last phase of the movement. Minor differences emerge as the 90-frame model shows sharper peaks in the final segment (Figure 8), suggesting stronger contrast with the preceding context, whereas the 30-frame model exhibits a more gradual increase due to the shorter temporal history. Spatial attribution patterns remain stable across configurations, indicating

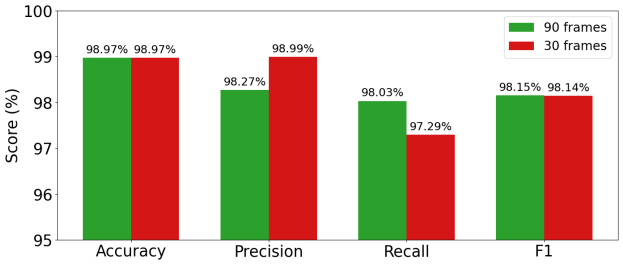


Fig. 6: Transformer performance comparison for 90-frame and 30-frame input sequences.

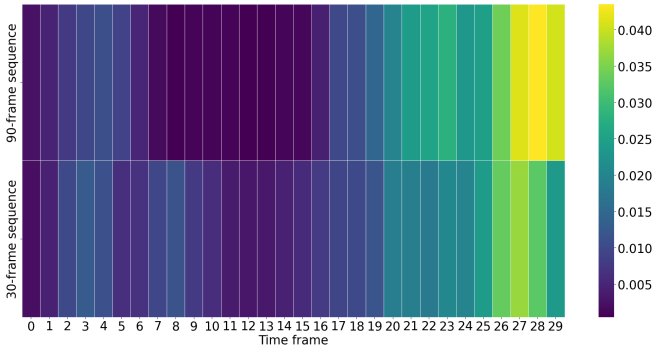


Fig. 8: Heatmap of mean absolute ShaTS values across the last 30 frames for the 90-frame model (top row) and the retrained 30-frame model (bottom row).

that reducing the temporal window does not alter the model's spatial strategy.

To further refine the spatial interpretation of the model, an additional analysis was conducted by separating the horizontal (x) and vertical (y) coordinates of each joint to reveal whether the model primarily relies on vertical displacements, horizontal movements, or a combination of both. In fall detection, vertical displacement is typically associated with loss of height, while horizontal movement may reflect imbalance or compensatory dynamics. The results, shown in Figure 9, reveal that vertical displacements dominate the decision process for most joints, particularly for those associated with the upper body, which is the part that most influences the prediction. This is consistent with the biomechanical characteristics of a fall, which involves a rapid loss of height. However, for other joints, the impact of the vertical component is less pronounced, and in some cases, the contributions of (x) and (y) are comparable. This suggests that, beyond vertical collapse dynamics, the model also captures lateral movements, integrating both spatial dimensions depending on the joint considered.

Finally, misclassified samples were analyzed to investigate how the model's decision process behaves under error conditions. False negatives (FN) were compared with true positives (TP) to understand what changes when a real fall is missed, while false positives (FP) were compared with true negatives (TN) to assess what differs when an ADL is incorrectly classified as a fall. In this setting, a negative FN−TP value indicates reduced feature contribution in missed falls, whereas a positive FP−TN value reflects increased contribution in false alarms. From a temporal perspective (Figure 10(a)), differences emerge mainly in the final frames: FN–TP values decrease toward the end of the sequence, while FP–TN values increase, confirming that errors are associated with attenuation or amplification of the same temporally localized evidence identified in the global analysis. At the spatial level (Figure 10(b)), the largest deviations involve upper-body regions, already recognized as the most influential components. In false negatives, their contributions are weaker than in correctly detected falls; in false positives, they are stronger than in correctly classified ADL samples. No new regions become dominant in error cases, and the overall importance hierarchy remains stable. Overall, the results indicate that the model's decision mechanism remains internally coherent. Errors do not arise from a shift in temporal or spatial focus, but from attenuation or amplification of the same decision-relevant features, suggesting issues related to the strength of the decision signal rather than instability in the learned representation.

IV. DISCUSSION AND CONCLUSION

This work aimed to investigate the decision-making process of a Transformer-based model for fall detection through explainability techniques, providing a human-interpretable understanding of its predictions. To this end, an alternative formulation of SHAP, namely ShaTS, was adopted. Unlike standard SHAP, which assumes feature independence, ShaTS is better suited for structured data characterized by temporal and spatial dependencies, as in this work. Given the intrinsic correlations between consecutive frames and between body joints, different grouping strategies were employed to account for both temporal and anatomical dependencies.

At the temporal level, ShaTS showed that the model concentrates most of its discriminative power in the final portion of the sequence. Contributions increase from approximately frame 60 onward, indicating that the Transformer primarily identifies a fall through motion patterns occurring in the last phase of the event, while earlier frames provide mainly contextual support. Based on this finding, the model was retrained using only the last 30 frames of each sequence. The reduced model achieved performance comparable to the full 90-frame configuration, and the corresponding ShaTS explanations preserved the same temporal structure, confirming that the final segment contains most of the discriminative information required for fall detection. However, the 30-frame configuration

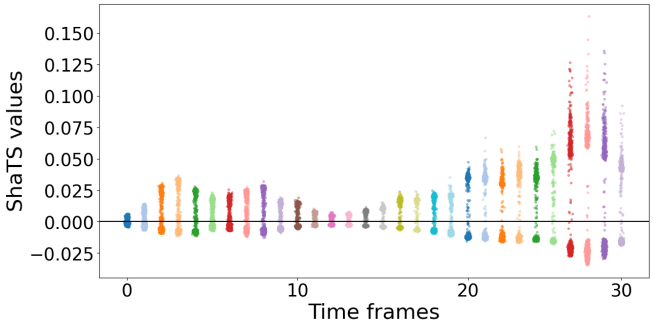


Fig. 7: Distribution of ShaTS values for the retrained Transformer model over 30 frames.

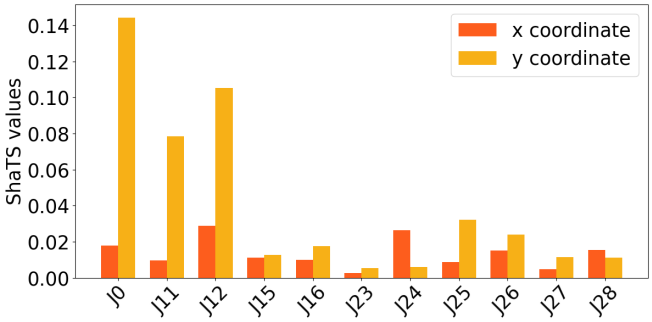


Fig. 9: Comparison of ShaTS values for horizontal (x) and vertical (y) coordinates across individual joints.

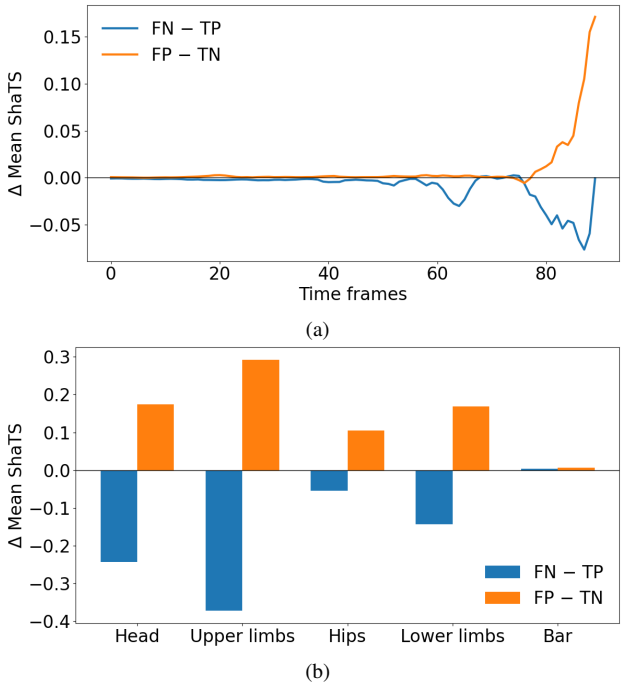


Fig. 10: Differences in mean ShaTS values between misclassified and correctly classified samples within the same ground-truth class, computed as FN−TP (missed falls vs. correctly detected falls) and FP−TN (false alarms vs. correctly classified ADL). (a) Temporal differences across frames. (b) Differences aggregated by body region.

exhibited a slight reduction in recall, indicating that some true falls are missed when the initial motion dynamics are excluded. Although the shorter window remains informative, the full 90-frame model is preferable in safety-critical contexts, where maximizing recall is more important than minimizing false alarms.

At the spatial level, both body-part and joint analyses revealed that the upper-body regions, particularly the head and shoulder joints, exhibited the highest contributions. Lower limbs and hips showed moderate influence, while the BAR feature had a negligible impact. These findings indicate that the model detects a fall mainly as a dynamic event characterized by upper-body collapse and pronounced motion in the final frames. Furthermore, the analysis of spatial coordinates showed that vertical (y) displacements dominate the decision process for most upper-body joints, consistent with the biomechanical nature of a fall, which involves rapid loss of height. However, horizontal (x) contributions were also non-negligible for several joints, indicating that the model integrates both vertical collapse and lateral motion patterns.

Finally, the error analysis demonstrated that misclassifications do not occur from a shift in temporal or spatial focus. Instead, false negatives and false positives arise from weaker or stronger activation of the same decision-relevant features identified in correct predictions. This suggests that the model's decision strategy remains internally coherent, and that errors are more likely related to variations in the strength of the decision signal rather than instability in the learned representation. Based on these findings, future work could investigate strategies to improve robustness in borderline cases to reduce false negatives without substantially increasing false positives. Given that most discriminative information is concentrated in the final frames, further research could explore reduced or adaptive temporal windows tailored to real-time applications, dynamically adjusting the observation length according to the evolution of the motion to balance latency and reliability. Finally, a systematic comparison between ShaTS and attention-based interpretability approaches could provide additional insights into the relationship between internal attention distributions and post-hoc attribution methods, helping to better understand the complementarity and limitations of different explainability techniques in Transformer-based fall detection models.

References

[1] E. Rassekh and L. Snidaro, "Survey on data fusion approaches for fall-detection," *Information Fusion*, vol. 114, p. 102696, 2025.

[2] K. W. Park, M. S. Mirian, and M. J. McKeown, "Artificial intelligence-based video monitoring of movement disorders in the elderly: a review on current and future landscapes," *Singapore Medical Journal*, vol. 65, no. 3, pp. 141–149, 2024.

[3] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognition*, vol. 108, p. 107561, 2020.

[4] P. Fantozzi and M. Naldi, "The explainability of transformers: Current status and directions," *Computers*, vol. 13, no. 4, p. 92, 2024.

[5] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[6] S. Jain and B. C. Wallace, "Attention is not explanation," *arXiv preprint arXiv:1902.10186*, 2019.

[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[8] M. Franco De La Peña, Ángel Luis Perales Gómez, and L. Fernández Maimó, "ShaTS: a Shapley-based explainability method for time series artificial intelligence models," *Future Generation Computer Systems*, vol. 176, p. 108178, 2026.

[9] C. Patruno, G. Cicirelli, L. Romeo, and T. D'Orazio, "Analysis of Input Data Configurations in CNN-based Human Action Recognition for Assembly Task," in *CoDIT 2025 - 11th International Conference on Control, Decision and Information Technologies*, vol. 1, 2025, pp. 22–27.

[10] C. Patruno, L. Romeo, A. Bono, T. D'Orazio, and G. Cicirelli, "Skeleton-based human action recognition for manufacturing in assembly task by deep learning," in *Multimodal Sensing and Artificial Intelligence for Sustainable Future*, vol. 13570, International Society for Optics and Photonics. SPIE, 2025, p. 135701D.

[11] I. Charfi, J. Miteran, J. Dubois, M. Atri, and R. Tourki, "Optimised spatio-temporal descriptors for real-time fall detection: comparison of SVM and Adaboost based classification," *Journal of Electronic Imaging (JEI)*, vol. 22, no. 4, p. 17, 2013.

[12] B. Kwolek and M. Kepski, "Human fall detection on embedded platform using depth maps and wireless accelerometer," *Computer methods and programs in biomedicine*, vol. 117, no. 3, pp. 489–501, 2014.

[13] 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, p. 1988, 2019.

[14] G. Baldewijns, G. Debard, G. Mertes, B. Vanrumste, and T. Croonenborghs, "Bridging the gap between real-life data and simulated data by providing a highly realistic fall dataset for evaluating camera-based fall detection algorithms," *Healthcare Technology Letters*, vol. 3, no. 1, pp. 6–11, 2016.

[15] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[16] A. Longo, A. Bono, G. Guaragnella, P. Boccadoro, A. Rana, A. Petitti, and T. D'Orazio, "WATCHDOG: Autonomous Elderly Assistance via Attention-based Fall Detection and Trajectory Prediction," in *Proceedings of IEEE International Conference on Robotics & Automation (ICRA)*, 2026.

[17] A. Ali, T. Schnake, O. Eberle, G. Montavon, K.-R. Müller, and L. Wolf, "XAI for transformers: Better explanations through conservative propagation," in *International conference on machine learning*. PMLR, 2022, pp. 435–451.

[18] D. Bhusal, R. Shin, A. A. Shewale, M. K. M. Veerabhadran, M. Clifford, S. Rampazzi, and N. Rastogi, "Sok: Modeling explainability in security analytics for interpretability, trustworthiness, and usability," in *Proceedings of the 18th International Conference on Availability, Reliability and Security*, 2023, pp. 1–12.

[19] L. S. Shapley *et al.*, "A value for n-person games," 1953.