

When Do Gradient-Based Explanations Identify Important Video Frames? A Global-Local Evaluation of Time-Series Explanations

Min Hun Lee

mhlee@smu.edu.sg

Singapore Management University
Singapore, Singapore

Abstract

Gradient-based methods are widely used to explain deep video models, yet their ability to identify temporally important evidence remains insufficiently understood. We study gradient-based video explanation as frame-level temporal evidence detection. Using a global-local spatiotemporal attention network trained with weak video-level labels, we convert VanillaGrad, SmoothGrad, and Grad-CAM outputs into frame-level importance scores and evaluate them against annotated activity frames across four medically relevant video classification tasks. The results show that combining full-frame global views with local region-of-interest views improves both video classification and important-frame detection. The global-local model achieves 96.49 F1 for classification, and its SmoothGrad explanations achieve 86.68 F1 for important-frame detection. At the same time, the evaluation reveals failure modes: Grad-CAM is less effective for the transformer-based model, and CNN-based gradients are more reliable for long videos. These findings suggest that video explanations should be evaluated by their ability to identify relevant frames, not only by visual inspection of saliency heatmaps.

CCS Concepts

• **Computing methodologies** → **Activity recognition and understanding**; **Machine learning**; • **Information systems** → **Multimedia information systems**; • **Applied computing** → **Health care information systems**.

Keywords

Gradient-based explanations; Important-frame detection; Video explanations; Spatiotemporal attention; Activity recognition

ACM Reference Format:

Min Hun Lee. 2026. When Do Gradient-Based Explanations Identify Important Video Frames? A Global-Local Evaluation of Time-Series Explanations. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26)*, November 07–11, 2026, Rome, Italy. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3799682.3839999>

1 Introduction

Deep video models are increasingly used to analyze sequential and multimodal data in health-related settings, including activity recognition and rehabilitation assessment [16, 23]. Although

these models can achieve strong predictive performance, it is often unclear which frames support their predictions. This challenge is especially important for video classification: a model may predict the correct activity label, but an explanation should also indicate *when* the relevant activity occurs. For video and time-series data, explanation is therefore not only a spatial localization problem, but also a temporal evidence detection problem.

Gradient-based methods are widely used as post-hoc explanations for deep models [6, 29, 30]. These methods estimate the contribution of input pixels or regions to a model prediction and are often visualized as saliency heatmaps. However, most saliency-based explanations have been studied in image settings, where the main question is which spatial regions are important. In videos, a frame-level heatmap may look visually plausible but still fail to identify the frames that contain the relevant temporal evidence. Prior work has explored saliency-based explanations for time-series and video models [10, 24, 25], while others have called for more rigorous temporal evaluation of XAI methods beyond visual interpretability [26]. However, it remains unclear when gradient-based explanations recover annotated important frames, and how this ability depends on model architecture, input representation, attribution method, and sequence length.

In this paper, we study gradient-based video explanation as *frame-level temporal evidence detection*. We instantiate this evaluation with a global-local spatiotemporal attention network (GLSTAN) trained using weak video-level labels. Frame-level annotations are used only for evaluation, allowing us to test whether explanations recover the moments when the target activity occurs. We compare global video frames, local region-of-interest views, and their combination, and convert VanillaGrad [29], SmoothGrad [30], and Grad-CAM [6, 27] outputs into frame-level importance scores. We evaluate these scores across four medically relevant classification tasks.

Our results show that combining full-frame global views with local region-of-interest views improves both video classification and important-frame detection. GLSTAN with global-local views achieves 96.49 F1 for classification, and its SmoothGrad explanations achieve 86.68 F1 for important-frame detection. At the same time, the evaluation reveals failure modes: Grad-CAM is weaker for the transformer-based model, and CNN-based gradients are more reliable for long videos. These findings suggest that video explanations should be evaluated by their ability to identify relevant frames, not only by visual inspection of saliency heatmaps.

This work makes three contributions. First, we formulate gradient-based video explanation as frame-level important-frame detection, using weak video-level labels for training and frame-level annotations only for evaluation. Second, we compare global, local,



This work is licensed under a Creative Commons Attribution 4.0 International License. *CIKM '26, Rome, Italy*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2539-5/2026/11

<https://doi.org/10.1145/3799682.3839999>

and global-local video representations for both video classification and temporal explanation quality. Third, we identify method- and sequence-dependent limits of gradient-based explanations, showing when global-local transformer explanations help and when CNN-based gradients remain preferable for long videos.

2 Related Work

2.1 Explainable AI for Time-Series and Video Classification

Attribution-based explanations, such as VanillaGrad, SmoothGrad, and Grad-CAM, are widely used to explain deep models by estimating which inputs contribute to a prediction [6, 27, 29, 30]. Extending these methods from images to time-series and videos is challenging because relevant evidence is distributed across both features and time [24, 31].

Prior work has studied instance-wise feature importance for temporal models [31], benchmarked interpretability methods for deep time-series prediction [13], and proposed perturbation-based explanations for time-series classifiers [22]. Recent studies further examine class-specific explanations [9], saliency robustness [2], interval-based counterfactual explanations [32], and time-frequency explanations [7]. Related work used saliency-derived pseudo-labels to train frame-level rehabilitation classifiers from weakly annotated videos [8]. Preliminary studies evaluated gradient-based saliency for recovering important frames in rehabilitation [16] and medically relevant activity videos [15]. Building on this work, we frame video explanation as important-frame detection and provide a consolidated analysis across model architectures, input views, attribution methods, and sequence lengths.

2.2 Spatiotemporal Video Models and Global-Local Evidence

Video classification models have advanced from convolutional and recurrent architectures to transformer-based spatiotemporal models that capture long-range dependencies across frames [1, 3, 19, 20]. These models improve activity recognition by modeling appearance, motion, and temporal context. However, stronger classification performance does not necessarily imply better explanation quality: a model may classify a video correctly while its attribution scores emphasize irrelevant frames or produce diffuse temporal evidence.

Prior video models have used spatiotemporal attention and local/global context to improve action recognition or temporal localization [21, 33]. Global frames preserve full-frame context, while local region-of-interest views can reduce background influence and focus the model on activity-relevant visual evidence. However, these representations have primarily been studied for recognition or localization performance, leaving their effect on gradient-based temporal explanations underexplored.

Our work addresses this gap by jointly evaluating video classification and important-frame detection across global, local, and global-local views. This allows us to test not only whether global-local transformer representations improve prediction, but also whether their gradient-based explanations recover annotated activity frames, and when attribution method or sequence length limits explanation reliability.

3 Methods

We study gradient-based video explanation as a frame-level temporal evidence detection problem (Figure 1). Given a video clip and a trained classifier, the goal is to test whether attribution scores identify the frames in which the target activity occurs. Our framework has two steps: training a weakly supervised video classifier with global, local, and global-local views, and converting gradient-based attribution maps into frame-level importance scores for evaluation.

3.1 Task Formulation

Let $\mathcal{X} \in \mathbb{R}^{3 \times T \times H \times W}$ denote a video clip with T frames, spatial resolution $H \times W$, and three color channels. Each clip has a video-level label $y \in \{1, \dots, C\}$, where C is the number of activity classes. A classifier f_θ predicts class probabilities and a video-level label:

$$\hat{\mathbf{p}} = f_\theta(\mathcal{X}), \quad \hat{y} = \arg \max_{c \in \{1, \dots, C\}} \hat{p}_c. \quad (1)$$

The classifier is trained using only video-level labels. Frame-level annotations are not used to train the model; they are reserved for evaluating whether explanations identify temporal evidence.

3.2 Global-Local Video Classification

We instantiate f_θ using a Global-Local SpatioTemporal Attention Network (GLSTAN), based on UniFormer blocks for efficient spatiotemporal representation learning [19]. GLSTAN uses local spatiotemporal relation aggregation in early stages and global relation aggregation in later stages, allowing the model to capture both short-range motion and long-range temporal context.

To examine how input evidence affects both prediction and explanation, we construct three views of each video. The *global* view uses the full frame. The *local* view focuses on the region of interest by applying a binary spatiotemporal mask $\mathcal{M} \in \{0, 1\}^{1 \times T \times H \times W}$ estimated from OpenPose keypoints [4]:

$$\mathcal{X}^L = \mathcal{X} \odot \mathcal{M}. \quad (2)$$

The *global-local* view concatenates the full-frame video with the ROI-masked video, allowing the model to use both scene context and localized activity evidence. We train separate classifiers using global, local, and global-local views, which allows us to test whether localized visual evidence improves video classification and whether it also changes the temporal quality of gradient-based explanations.

3.3 Gradient-Based Temporal Explanation

After training, we compute attribution maps for the predicted class $\hat{y}$ using VanillaGrad [29], SmoothGrad [30], and Grad-CAM [6, 27]. Let $\mathcal{A}^g \in \mathbb{R}^{D \times T \times H \times W}$ denote the attribution map produced by method g after resizing to the input resolution when needed, where D is the number of input channels for the corresponding view. To compare attribution methods at the frame level, we aggregate attribution values over channels and spatial dimensions:

$$r_t^g = \frac{1}{DHW} \sum_{c=1}^D \sum_{h=1}^H \sum_{w=1}^W |\mathcal{A}_{c,t,h,w}^g|, \quad t = 1, \dots, T. \quad (3)$$

The resulting frame scores are min-max normalized within each video to obtain $z_t^g \in [0, 1]$. A frame is predicted as important if its

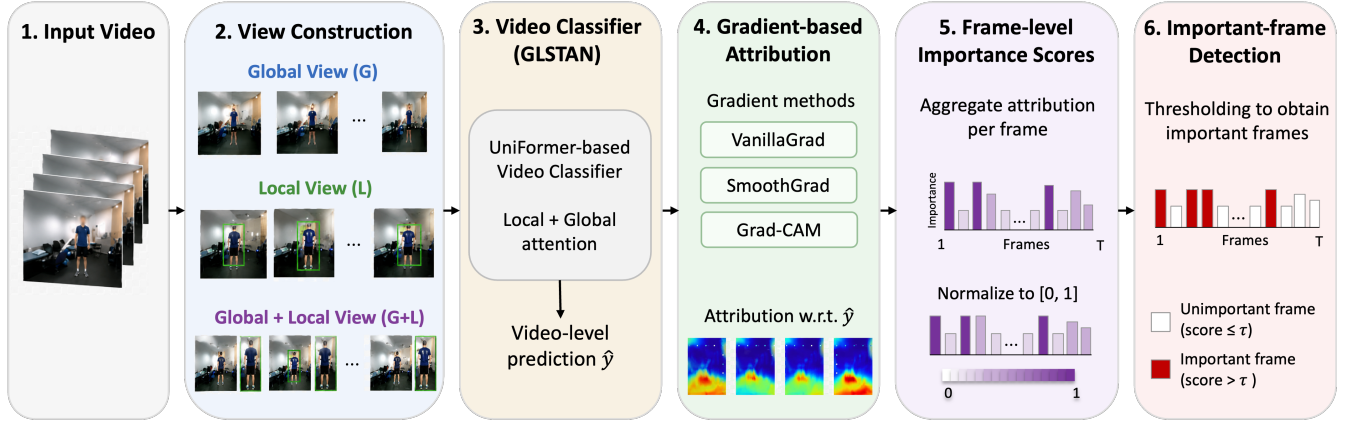


Figure 1: Overview of the proposed frame-level explanation pipeline. Given an input video, we construct global, local, and global-local views and train a GLSTAN classifier using weak video-level labels. We then apply gradient-based attribution methods, aggregate attribution scores over each frame, and threshold normalized scores to identify important frames as temporal evidence.

normalized score exceeds a threshold τ selected on the training split and then fixed for held-out evaluation:

$$\hat{b}_t^g = 1[z_t^g > \tau]. \quad (4)$$

This fixed-threshold protocol provides a controlled comparison across models and attribution methods, but it does not eliminate threshold sensitivity; we therefore interpret the resulting F1 scores as one operational measure of temporal evidence recovery rather than a complete temporal localization benchmark.

We evaluate explanation quality using important-frame detection F1 by comparing predicted important frames $\hat{b}_t^g$ with frame-level annotations b_t . This treats gradient-based video explanation as temporal evidence detection: a useful explanation should assign high importance to frames that contain the annotated activity and low importance to irrelevant frames.

4 Experiments

We evaluate whether gradient-based explanations recover temporally important evidence in video classification. Our experiments address two questions: (1) whether global-local spatiotemporal representations improve weakly supervised video classification, and (2) whether attribution scores from trained classifiers identify annotated important frames. We compare model architecture, input view, attribution method, and sequence length under the same frame-level evaluation protocol.

Datasets. We use two video datasets. From NTU RGB+D [28], which contains 114,480 videos across 60 activity classes, we select three medically relevant classes: falling down, headache, and chest pain. Each class contains 948 videos, yielding 2,844 videos for weakly supervised video classification. To evaluate temporal explanation quality, we annotate 900 videos with frame-level labels indicating whether the target activity occurs at each frame. We also use a stroke rehabilitation dataset from prior rehabilitation-assessment and personalized coaching work [17, 18], containing 300 videos from 15 post-stroke survivors performing rehabilitation exercises.

Each video has a video-level label indicating whether compensatory motion occurs and expert frame-level annotations indicating when compensation occurs. Across both datasets, video-level labels are used to train classifiers, while frame-level annotations are reserved only for evaluating explanation quality.

Models and Training. We compare GLSTAN with three baselines: a ResNet50 feature-based fully connected classifier (FCL) [11], a CNN over patched frame features [14], and an LSTM-based recurrent classifier [12]. GLSTAN is initialized from a UniFormer model pre-trained on Kinetics-400 [5, 19]. For each model, we evaluate global, local, and global-local input views. The global view uses full frames, the local view masks non-ROI pixels using OpenPose-estimated body regions [4], and the global-local view concatenates full-frame and ROI-masked views. For NTU RGB+D tasks, we use 10-fold cross-validation over 2,844 videos. For the rehabilitation task, we use leave-one-patient-out cross-validation over 15 participants. GLSTAN is trained with AdamW, a cosine learning-rate schedule, and a batch size of 64.

Evaluation. We report video classification F1 and important-frame detection F1. For explanation evaluation, we generate attribution maps using VanillaGrad [29], SmoothGrad [30], and Grad-CAM [6, 27]. Attribution maps are converted into frame-level importance scores by aggregating attribution values over spatial dimensions and normalizing scores within each video. The threshold for important-frame prediction is selected on the training split and fixed for held-out evaluation. We evaluate short 20-frame clips and long 394-frame rehabilitation videos to test whether explanation reliability changes with sequence length.

5 Results and Discussion

Table 1 reports weakly supervised video classification F1. GLSTAN with global-local views achieves the best performance, with 96.49 F1. Compared with GSTAN using global views only and LSTAN using local views only, GLSTAN improves F1 from 95.43 and 95.61 to 96.49, suggesting that full-frame context and ROI-focused evidence

provide complementary information. ResNet-based baselines also benefit from global-local views, but remain substantially below the UniFormer-based models.

Table 1: Weakly supervised video classification F1 for global, local, and global-local views.

Model	Backbone	Pretrain	Global	Local	Global+Local
FCL	ResNet50	–	80.87	–	–
FCL	ResNet50	Kinetics	81.48	74.20	90.63
CNN	ResNet50	Kinetics	82.43	74.20	90.39
RNN/LSTM	ResNet50	Kinetics	83.61	–	–
STAN	UniFormer	Kinetics	95.43	95.61	96.49

Table 2 reports important-frame detection F1 across attribution methods. GLSTAN with global-local views performs best with VanillaGrad and SmoothGrad, reaching 86.43 and 86.68 F1, respectively. Compared with CNN global-local using SmoothGrad, GLSTAN improves important-frame detection from 83.36 to 86.68. This shows that the global-local representation improves not only video-level classification, but also the ability of attribution scores to recover temporally important evidence.

Table 2: Important-frame detection F1 across four activities, input views, and attribution methods.

Model	View	VanillaGrad	SmoothGrad	Grad-CAM
CNN	Global	82.09	82.04	81.30
CNN	Local	82.33	82.33	82.08
CNN	Global+Local	82.89	83.36	82.22
GLSTAN	Global	85.87	86.19	80.47
GLSTAN	Local	85.82	86.01	79.04
GLSTAN	Global+Local	86.43	86.68	80.60

The gap between attribution methods highlights why temporal explanation should be evaluated separately from video classification. VanillaGrad and SmoothGrad benefit from GLSTAN’s global-local representation, suggesting that this representation yields attribution patterns that better align with temporally relevant activity evidence. In contrast, Grad-CAM does not show the same benefit with GLSTAN and performs best with the CNN global-local model, reaching 82.22 F1 compared with 80.60 for GLSTAN global-local. This indicates that attribution quality depends not only on the predictive model, but also on how the explanation method extracts and aggregates evidence from that model. One possible reason is that Grad-CAM relies on coarser intermediate activation maps and subsequent resizing, which can make frame-level evidence more spatially diffuse than input-gradient methods such as VanillaGrad and SmoothGrad. This finding cautions against treating saliency quality as a direct byproduct of classification accuracy.

The qualitative comparison in Figure 2 illustrates the method-dependent nature of GLSTAN explanations. In this example, VanillaGrad produces more localized saliency around the relevant body region, while Grad-CAM with GLSTAN produces more diffuse activations across frames. This pattern is consistent with Table 2, where

GLSTAN performs strongly with VanillaGrad and SmoothGrad but less effectively with Grad-CAM.

Sequence length further changes the relative advantage of model architectures. On the rehabilitation task, GLSTAN global-local performs best on short 20-frame clips, reaching 83.69 F1 with SmoothGrad compared with 79.89 for CNN global-local. On long 394-frame videos, however, CNN global-local reaches 95.60 F1, while GLSTAN global-local reaches 87.80. This suggests that transformer-based explanations are effective for short sampled clips, whereas CNN-based gradients may be more reliable for dense long-sequence attribution.

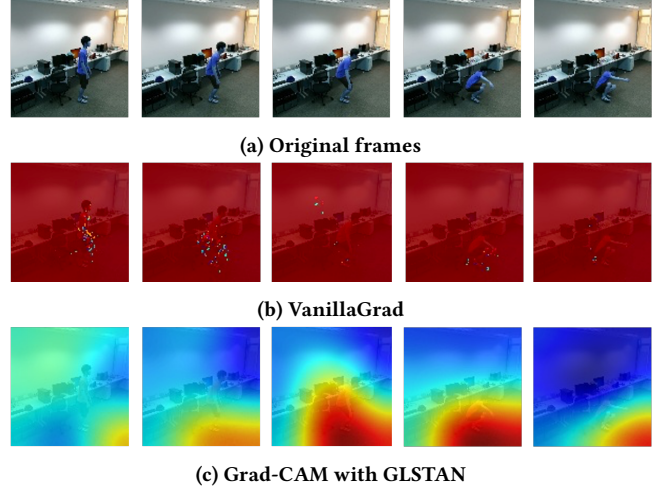


Figure 2: Qualitative comparison of gradient-based explanations across representative video frames from the NTU RGB+D Action Recognition Dataset. In this example, VanillaGrad highlights localized body-region evidence, while Grad-CAM with GLSTAN produces more diffuse activations. © ROSE Lab, Nanyang Technological University, Singapore.

These results position important-frame detection as a useful evaluation protocol for video explanations. Unlike qualitative saliency inspection, frame-level evaluation directly tests whether explanations recover the temporal evidence associated with a prediction. This is especially important for video and time-series settings, where explanations must identify not only which visual evidence is relevant, but also when that evidence occurs. The proposed evaluation therefore provides a compact way to compare model architectures, input representations, and attribution methods under the same temporal evidence detection objective.

Overall, our findings show that global-local views improve both classification and important-frame detection, but explanation reliability remains method- and sequence-dependent. Our study is limited to one transformer-based architecture, a small set of medically relevant activity tasks, ROI construction from a pose estimator, and threshold-based F1 evaluation. In particular, the local view depends on OpenPose-derived body regions, which may fail under occlusion, poor pose estimation, or non-human activity settings. Future work should examine additional architectures, ranking-based temporal metrics, and broader video or time-series domains.

6 Conclusion

We studied gradient-based video explanation as frame-level temporal evidence detection. Across four medically relevant activity tasks, global-local video representations improved both weakly supervised classification and important-frame detection, with GLSTAN and SmoothGrad achieving the strongest temporal explanation performance. The results also reveal important failure modes: Grad-CAM is less effective with GLSTAN, and CNN-based gradients are stronger for long videos. These findings show that video explanations should be evaluated by whether they identify relevant frames, not only by visual inspection of saliency heatmaps.

Acknowledgments

This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (MOE-T2EP20223-0007). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not reflect the views of the Ministry of Education, Singapore. Portions of this work used the NTU RGB+D Action Recognition Dataset made available by the ROSE Lab at Nanyang Technological University, Singapore.

GenAI Usage Disclosure

The author used GPT-based tools for grammar and spelling checks, minor rewording, and structural editing support. These tools were not used to generate code, create or modify datasets, design experiments, perform analyses, produce results, or contribute to the intellectual content or scientific conclusions of the work. All content was subsequently reviewed and edited by the author, who assumes full responsibility for the publication.

References

- [1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. 2021. ViViT: A Video Vision Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6836–6846.
- [2] Chiara Balestra, Bin Li, and Emmanuel M. Muller. 2023. On the Consistency and Robustness of Saliency Explanations for Time Series Classification. *arXiv preprint arXiv:2309.01457* (2023).
- [3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding?. In *Proceedings of the 38th International Conference on Machine Learning*. 813–824.
- [4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.
- [5] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.
- [6] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 839–847.
- [7] Hyunseung Chung, Sumin Jo, Yeonsu Kwon, and Edward Choi. 2024. Time is Not Enough: Time-Frequency based Explanation for Time-Series Black-Box Models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 394–403. doi:10.1145/3627673.3679844
- [8] Ana Rita C6ias, Min Hun Lee, Alexandre Bernardino, Asim Smailagic, Mariana Mateus, David Fernandes, and Sofia Trapola. 2025. Learning Frame-Level Classifiers for Video-Based Real-Time Assessment of Stroke Rehabilitation Exercises from Weakly Annotated Datasets. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 33 (2025), 3334–3345. doi:10.1109/TNSRE.2025.3602548
- [9] Ramesh Doddaiiah, Prathyush S Parvatharaju, Thomas Hartvigsen, and Elke A Rundensteiner. 2022. Class-Specific Explainability for Deep Time Series Classifiers. *arXiv preprint arXiv:2210.05411* (2022).
- [10] Sebastian D Goodfellow, Andrew Goodwin, Robert Greer, Peter C Laussen, Mjaye Mazwi, and Danny Eytan. 2018. Towards understanding ECG rhythm classification using convolutional neural networks and attention mappings. In *Machine learning for healthcare conference*. PMLR, 83–101.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
- [12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780. doi:10.1162/neco.1997.9.8.1735
- [13] Aya Abdelsalam Ismail, Mohamed Gunady, Hector Corrada Bravo, and Soheil Feizi. 2020. Benchmarking deep learning interpretability in time series predictions. *Advances in neural information processing systems* 33 (2020), 6441–6452.
- [14] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324. doi:10.1109/5.726791
- [15] Min Hun Lee. 2024. Towards gradient-based time-series explanations through a spatiotemporal attention network. *arXiv preprint arXiv:2405.17444* (2024).
- [16] Min Hun Lee and Yi Jing Choy. 2023. Exploring a Gradient-based Explainable AI Technique for Time-Series Data: A Case Study of Assessing Stroke Rehabilitation Exercises. *arXiv preprint arXiv:2305.05525* (2023).
- [17] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2019. Learning to assess the quality of stroke rehabilitation exercises. In *Proceedings of the 24th international conference on intelligent user interfaces*. 218–228.
- [18] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2023. Design, development, and evaluation of an interactive personalized social robot to monitor and coach post-stroke rehabilitation exercises. *User Modeling and User-Adapted Interaction* 33, 2 (2023), 545–569.
- [19] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. 2022. Uniformer: Unified transformer for efficient spatiotemporal representation learning. *arXiv preprint arXiv:2201.04676* (2022).
- [20] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video Swin Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3202–3211.
- [21] Lili Meng, Bo Zhao, Bo Chang, Gao Huang, Wei Sun, Frederick Tung, and Leonid Sigal. 2019. Interpretable Spatio-Temporal Attention for Video Action Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 1513–1522. doi:10.1109/ICCVW.2019.00189
- [22] Prathyush S Parvatharaju, Ramesh Doddaiiah, Thomas Hartvigsen, and Elke A Rundensteiner. 2021. Learning saliency maps to explain deep time series classifiers. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1406–1415.
- [23] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. 2022. AI in health and medicine. *Nature medicine* 28, 1 (2022), 31–38.
- [24] Thomas Rojat, Raphaël Puget, David Filliat, Javier Del Ser, Rodolphe Gelin, and Natalia Diaz-Rodríguez. 2021. Explainable artificial intelligence (xai) on timeseries data: A survey. *arXiv preprint arXiv:2104.00950* (2021).
- [25] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* (2017).
- [26] Udo Schlegel, Hiba Arnout, Mennatallah El-Assady, Daniela Oelke, and Daniel A. Keim. 2019. Towards a Rigorous Evaluation of XAI Methods on Time Series. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 4197–4201. doi:10.1109/ICCVW.2019.00516
- [27] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.
- [28] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1010–1019.
- [29] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).
- [30] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* (2017).
- [31] Sana Tonekaboni, Shalmali Joshi, Kieran Campbell, David K Duvenaud, and Anna Goldenberg. 2020. What went wrong and when? Instance-wise feature importance for time-series black-box models. *Advances in Neural Information Processing Systems* 33 (2020), 799–809.
- [32] Akihiro Yamaguchi, Ken Ueno, Ryusei Shingaki, and Hisashi Kashima. 2024. Learning Counterfactual Explanations with Intervals for Time-series Classification. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 4158–4162. doi:10.1145/3627673.3679952
- [33] Zixin Zhu, Wei Tang, Le Wang, Nanning Zheng, and Gang Hua. 2021. Enriching Local and Global Contexts for Temporal Action Localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13516–13525.