EXPLORING A GRADIENT-BASED EXPLAINABLE AI TECHNIQUE FOR TIME-SERIES DATA: A CASE STUDY OF ASSESSING STROKE REHABILITATION EXERCISES

Min Hun Lee & Yi Jing Choy
School of Computing and Information Systems
Singapore Management University
Singapore, 178902, Singapore
mhlee@smu.edu.sg; yijing.choy.2019@scis.smu.edu.sg

ABSTRACT

Explainable artificial intelligence (AI) techniques are increasingly being explored to provide insights into why AI and machine learning (ML) models provide a certain outcome in various applications. However, there has been limited exploration of explainable AI techniques on time-series data, especially in the healthcare context. In this paper, we describe a threshold-based method that utilizes a weakly supervised model and a gradient-based explainable AI technique (i.e. saliency map) and explore its feasibility to identify salient frames of time-series data. Using the dataset from 15 post-stroke survivors performing three upper-limb exercises and labels on whether a compensatory motion is observed or not, we implemented a feed-forward neural network model and utilized gradients of each input on model outcomes to identify salient frames that involve compensatory motions. According to the evaluation using frame-level annotations, our approach achieved a recall of 0.96 and an F2-score of 0.91. Our results demonstrated the potential of a gradient-based explainable AI technique (e.g. saliency map) for time-series data, such as highlighting the frames of a video that therapists should focus on reviewing and reducing the efforts on frame-level labeling for model training.

1 INTRODUCTION

Artificial intelligence (AI) models are increasingly being explored to support high-stake decision-making, such as industrial predictive maintenance (Mobley, 2002) and healthcare (e.g. heart-beat anomaly detection (Chuah & Fu, 2007), cancer diagnosis McKinney et al. (2020), rehabilitation assessment Lee et al. (2021)). A misprediction by AI models in these high-stake domains could have serious and lasting impacts on people and communities. As the end-users are unwilling to adopt these technologies without explaining AI model outcomes regardless of its performance (Wang et al., 2020), there is a growing emphasis to explain an AI model to end-users for their informed decision making Cai et al. (2019); Lee et al. (2021); Rojat et al. (2021).

There has been an expanding focus and research in explainable AI (XAI) methods for image and tabular data (Rojat et al., 2021). For instance, post-hoc explanations generate approximations of an AI model outcome by producing understandable information or explore saliency maps of an AI model outcome for image data Simonyan et al. (2013). Specifically, these saliency maps calculate the gradient of the loss function with respect to the input pixels for the class of an AI model outcome. Thus, these maps are useful to provide insights into which part of an image is considered important for the model outputs.

While saliency maps are traditionally for image data, previous studies have explored adapting saliency map methods for time series data (Rojat et al., 2021). Goodfellow et al. (2018) used Class Activation Maps (CAM), a variant of saliency maps, (Zhou et al., 2016) to explain classifications made by a convolutional neural network (CNN) on univariate electrocardiogram (ECG) data. Assaf & Schumann (2019) used a CNN to forecast the average power consumption of a photo-voltaic power plant. This study utilized a more versatile form of CAM, Grad-CAM (Selvaraju et al., 2017)

to explain the predictions from the CNN model. However, these previous works focus on exploring saliency maps as an explanation tool for a model's output through qualitative evaluations. For instance, after generating sequences of saliency maps of a video, these work overlaid these maps on the input data and checked qualitatively whether the highlighted areas of time-series data are relevant to a model output or not Samek et al. (2017). It remains unclear how well saliency maps can be used to identify salient frames in time-series data.

In this paper, we present a threshold score method that utilizes a weakly supervised machine learning (ML) model and a gradient-based explainable AI technique, saliency map to explain time-series data by identifying salient frames of the data. Using the dataset of three stroke rehabilitation exercises from 15 post-stroke survivors and labels on whether an exercise involves compensatory motions or not, we trained a feed-forward neural network model. Given the feed-forward neural network model, we computed the gradients of the loss function with respect to input data over all frames and normalized these gradient scores to determine whether a compensatory motion has occurred in a specific frame or not. After implementing a threshold score method, we utilized frame-level annotations on the presence of a compensatory motion to evaluate the feasibility of our method. Our results showed that our method can detect salient frames of a frame-level compensatory motion with a recall of 0.96 and an F2-score of 0.91. This work contributed to an empirical study on how well a weakly supervised ML model with a gradient-based explainable AI can be utilized as explainable AI techniques for time-series data (e.g. identifying salient frames of data).

2 Physical Stroke Rehabilitation & Data Preprocessing

Post-stroke survivors require conducting physical stroke rehabilitation exercises to regain their functional abilities. While conducting exercises, they often compensate (e.g. lean their trunk to the side - Figure 1b) for their limited functional abilities, which may be detrimental to long-term recovery (Alaverdashvili et al., 2008; Pain et al., 2015). Physical therapists monitor and correct compensatory movement in rehabilitation exercises, which are often recorded as patients perform them at home. However, examining through a video can be a long and tedious process. To address this challenge, this work explores a threshold-based method that utilizes a weakly supervised ML model and saliency maps to detect salient frames of time-series data (e.g. compensatory periods during an exercise).

This work utilizes the dataset that consists of 300 exercise trials performed by 15 patients (Lee et al., 2019). Each patient performed 10 trials each with his/her affected and unaffected side. The affected side refers to the side of the body affected by the stroke, and vice versa for the unaffected side (Figure 2). Given patients' exercise motions, this dataset includes two types of labels: (1) labels on each trial and (2) labels on each frame that indicates whether a compensatory motion occurs (0) or not (1 - normal).

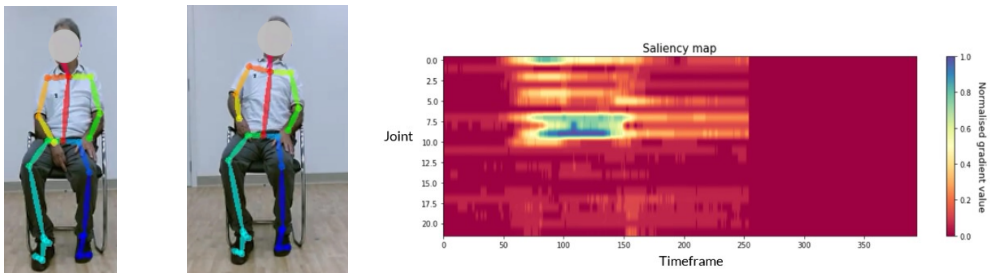
For processing the videos of patients performing an exercise, we utilized the Openpose(Cao et al., 2019) to estimate the body key points of patients from the videos and extracted features that indicate the distance from an initial to a current position at every frame of patients' exercise videos. We identified the maximum frame number of the dataset (i.e. 394) and padded an initial position to samples that have a smaller number of frames than the maximum frame number for downstream model training (Appendix. Figure 3 and Figure 4).

3 Gradient-based Explainable Technique for Time-Series Data

We describe the implementations of a weakly supervised machine learning model, a gradient-based explainable AI technique (i.e. saliency map), and a threshold-based method to explain time-series data (i.e. identifying salient frames of data) in the following sections.

3.1 A feed-forward neural network Model

We applied an 80-20 split on the dataset to obtain the training (i.e. 240 exercises) and test (i.e. 60 exercises) data sets to train a feed-forward neural network model to detect whether a compensatory motion occurs in an exercise or not. Using Skorch (Tietz et al., 2017) and three-fold cross-validation, we grid-searched different architectures and parameters and implemented a feed-forward neural



(a) (b) (c)

Figure 1: (a) The initial position and (b) compensatory motion of a post-stroke survivor. (c) the saliency map of a corresponding post-stroke survivor's exercise

network model with a test accuracy of 0.98. We summarize the architectures, hyper-parameters, and performances of the models in the Appendix. Figure 5 and describe the final, best-performing architecture in the Appendix. Figure 6.

3.2 Computing a Frame Score from Saliency Map

After implementing a feed-forward neural network, we computed gradients of the loss function with respect to input data over all frames. We then aggregated gradient scores of features at each frame level and normalized this frame-level score into [0, 1] (Appendix. Figure 7).

We hypothesize that there is a threshold value between the score range [0,1] where scores less than or equal to the threshold score indicates normal movement, and scores more than the threshold score indicates compensation. Frames involving compensation will be deemed more important, and thus result in a higher gradient score for the time frame.

4 Experiments

We conducted both qualitative and quantitative evaluations to explore the feasibility of a gradient-based explainable technique for time-series data. For the qualitative evaluation, we utilized a sample exercise and checked whether the most important frames and features of the exercise using a saliency map can provide relevant insights into a compensatory motion (Figure 1b).

In addition, we conducted quantitative evaluations of how well our threshold-based method can detect compensatory motions at the frame level. We explored the threshold value ranging from 0 to 1 and empirically selected an optimal threshold value. For the evaluation of our threshold-based approach, we utilized frame-level annotations and computed recall and f2-score with the $\beta = 2$. As our dataset is heavily skewed (e.g. containing a larger number of normal frames than abnormal frames with compensatory motions), we explored the effect of (i) removing zero-padded normal frames and (ii) utilizing only samples that include compensatory movements. Also, we explored the effect of aggregating a score over multiple windows $(5, 10, 15, 20)$. For post-analysis, the scores have been sorted into 2 groups: Group 0 contains scores where their corresponding frame label = 0, while Group 1 contains scores where their frame label = 1. The data distributions of Group 0 and 1 are listed in Appendix. Table 2.

5 Results

5.1 Qualitative Evaluation

Figure 1c shows a saliency map for a post-stroke survivor's exercise trial (Figure 1a and 1b). Blue-green regions denote the frames and features with the highest importance, while orange-red regions indicate those with the least importance. The flat red region at the end of the map indicates zero-padded frames. For this exercise trial, our method identified 50 - 60 frames and joint features (0:

HeadX, 2: NeckX, 5: ShoulderRightY, 6: ElbowRightX, 7: ElbowRightY, 8: WristRightX, 9: WristRightY, and 10:ShoulderLeftX) are most important with respect to the model classification. After qualitatively reviewing the corresponding compensatory frames (Figure 1b), we found the potential of our method to provide an insight into a compensatory motion: When a post-stroke survivor strives to raise his right wrist to complete an exercise, he involved compensatory motion by moving his head and left shoulder to the side and right shoulder upward.

5.2 Quantitative Evaluations

Table 1 describes the quantitative evaluations of how well our approach can identify salient frames with compensatory motions. We found that using only compensatory trials without padded frame scores gave the best results of recall = 0.96, F2-score = 0.91. As we utilized the dataset of more balanced distributions from all trials (90% of normal frames and 9% of compensatory frames in Appendix. Table 2) to compensatory trials without padded frames (34% of normal frames and 65 of compensatory frames in Appendix. Table 2), our approach significantly improved its performance to identify frame-level compensatory motions. When it comes to the computation of a gradient with different window sizes (i.e. 5, 10, 15, 20), we did not find any significantly different results. We achieved the best performance to detect frame-level compensatory motions by using only scores from samples involving compensatory motions and removing padded frame scores.

Experiment	Best Threshold Score	Recall	F2-Score
(1) All trials	0.4	0.44	0.44
(2) All trials without padded frames	0.4	0.44	0.44
(3) **Compensatory trials without padded frames**	**0.36**	**0.96**	**0.91**

Table 1: Experiment results using datasets with different data distributions of normal and compensatory frames and window size of 5. The best result is boldfaced

5.3 Discussion

In this work, we contributed to an empirical study that explores the feasibility of using a weakly supervised ML model and a gradient-based explainable AI (XAI) technique, saliency map for explaining time-series data (e.g. post-stroke rehabilitation exercises). Our qualitative and quantitative results are aligned with the hypothesis discussed in Section 3.2. Specifically, our results demonstrate the potential of our approach to detect frame-level compensatory motions of post-stroke survivors with a recall of 0.96 and an F2-score of 0.91. Our approach enables a therapist to pinpoint an important time period of a video that the therapist should prioritize reviewing and support human-AI collaborative annotations instead of a time-consuming manual labeling process Lee et al. (2022). However, there are several limitations to be addressed.

First, our quantitative evaluations using datasets with different distributions between normal and abnormal frames suggest the importance of utilizing a balanced dataset for applying our threshold-based method using a gradient-based XAI technique. When we utilized the unbalanced dataset with both unaffected and affected motions with padded frames (9% of compensatory frames and 90% of normal frames - Appendix. Table 2), we found that some scores were normalized to higher values and led to higher misdetections of frame-level compensatory motions.

In addition, it is also important to consider a more systematic approach to dynamically determine an *"ideal"* threshold value and address overlapping regions of scores between normal and compensatory frames (Appendix. Figure 8). Also, as this work only explores a feed-forward neural network model and a saliency map to compute gradient scores using video data from 15 post-stroke survivors, it is necessary to conduct additional studies to demonstrate the generalizability of our approach to other ML algorithms, gradient-based explainable AI methods, and data modalities (e.g. text).

Acknowledgments

This work has been supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant.

REFERENCES

Mariam Alaverdashvili, Afra Foroud, Diana H Lim, and Ian Q Whishaw. "learned baduse" limits recovery of skilled reaching for food after forelimb motor cortex stroke in rats: a new analysis of the effect of gestures on success. *Behavioural brain research*, 188(2):281–290, 2008.

Roy Assaf and Anika Schumann. Explainable deep neural networks for multivariate time series predictions. In *IJCAI*, pp. 6488–6490, 2019.

Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–14, 2019.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

Mooi Choo Chuah and Fen Fu. Ecg anomaly detection via time series analysis. In *Frontiers of High Performance Computing and Networking ISPA 2007 Workshops: ISPA 2007 International Workshops SSDSN, UPWN, WISH, SGC, ParDMCom, HiPCoMB, and IST-AWSN Niagara Falls, Canada, August 28-September 1, 2007 Proceedings 5*, pp. 123–135. Springer, 2007.

Sebastian D Goodfellow, Andrew Goodwin, Robert Greer, Peter C Laussen, Mjaye Mazwi, and Danny Eytan. Towards understanding ecg rhythm classification using convolutional neural networks and attention mappings. In *Machine learning for healthcare conference*, pp. 83–101. PMLR, 2018.

Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. Learning to assess the quality of stroke rehabilitation exercises. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 218–228, 2019.

Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. A human-ai collaborative approach for clinical decision making on rehabilitation assessment. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–14, 2021.

Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. Towards efficient annotations for a human-ai collaborative, clinical decision support system: A case study on physical stroke rehabilitation assessment. In *27th International Conference on Intelligent User Interfaces*, pp. 4–14, 2022.

Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

R Keith Mobley. *An introduction to predictive maintenance*. Elsevier, 2002.

Liza M Pain, Ross Baker, Denyse Richardson, and Anne MR Agur. Effect of trunk-restraint training on function and compensatory trunk, shoulder and elbow patterns during post-stroke reach: a systematic review. *Disability and rehabilitation*, 37(7):553–562, 2015.

Thomas Rojat, Raphaël Puget, David Filliat, Javier Del Ser, Rodolphe Gelin, and Natalia Díaz-Rodríguez. Explainable artificial intelligence (xai) on timeseries data: A survey. *arXiv preprint arXiv:2104.00950*, 2021.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Marian Tietz, Thomas J. Fan, Daniel Nouri, Benjamin Bossan, and skorch Developers. *skorch: A scikit-learn compatible neural network library that wraps PyTorch*, July 2017. URL `https://skorch.readthedocs.io/en/stable/`.

Fei Wang, Rainu Kaushal, and Dhruv Khullar. Should health care demand interpretable artificial intelligence or accept "black box" medicine?, 2020.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

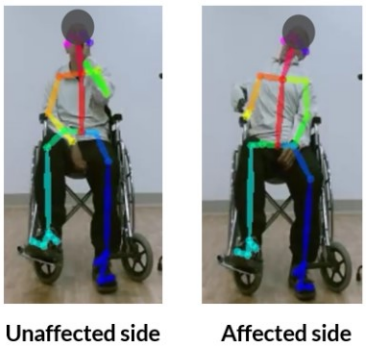


Figure 2: Movement of Unaffected and Affected Side

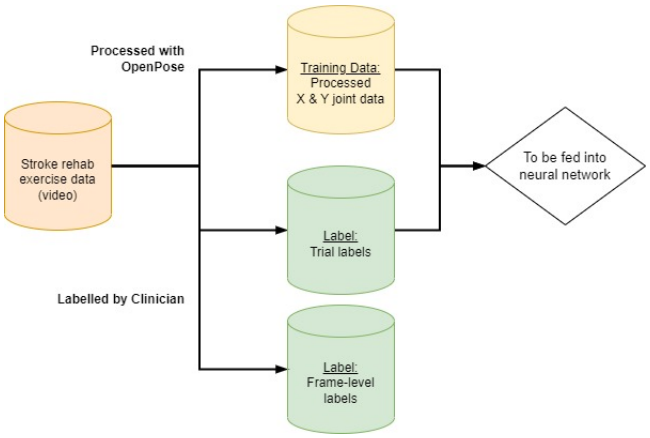


Figure 3: Data Preprocessing Diagram

A APPENDIX

Experiment	Group 0 Scores	Group 1 Scores	Total
(1) All trials	11,396 (9.64%)	106,804 (90.36%)	118,200
(2) All trials without padded frames	11,396 (9.64%)	37,048 (76.48%)	48,444
(3) Compensatory trials without padded frames	8,602 (65.75%)	4,485 (34.27%)	13,087
(4) Window size = 5 (best performance)	1,722 (65.25%)	917 (34.75%)	2,639

Table 2: Data Distribution

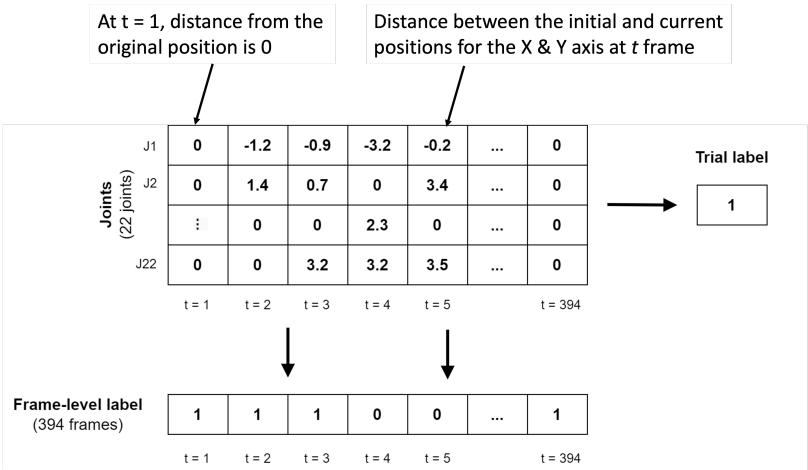


Figure 4: Shape of Data and Label after Preprocessing for an Exercise Trial

Hidden layers	Hidden layer sizes	lr	Epoch	Optimiser	Best lr	Best epoch	Best optimiser	Val set *Accuracy*	Test set *Accuracy*
2	512, 32	0.0001, 0.0005, 0.001, 0.005, 0.01, 0.1	10, 15, 20	SGD, Adam	0.001	15	Adam	0.98	0.97
	1024, 512				0.0001	15	Adam	0.98	0.98
	2048, 512				0.1	15	Adam	0.92	0.83
3	2048, 1024, 512				0.1	15	Adam	0.98	0.95
	4096, 2048, 1024				0.005	20	Adam	0.98	0.92

Figure 5: Summary of Model Parameters and Performance. The best performed architecture and parameters are higlighted in blue.

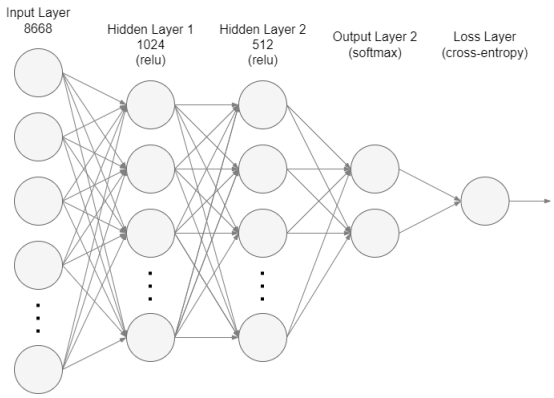


Figure 6: Final Model Architecture

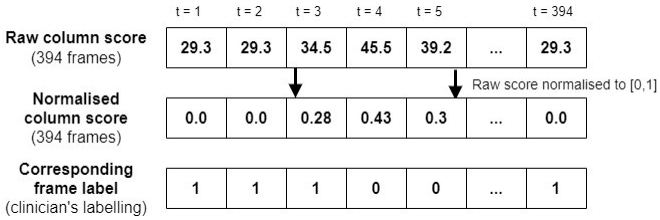


Figure 7: Score Computation Process

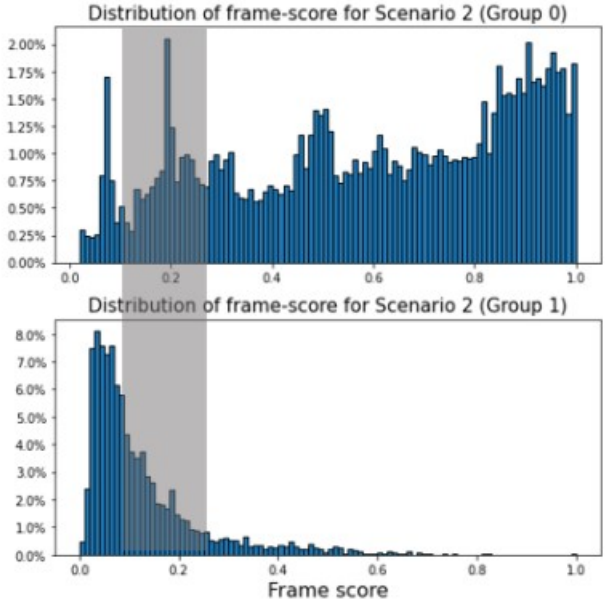


Figure 8: The histogram of normalized gradient scores using normal and abnormal, compensatory frames. The grey area indicates the overlapping areas between normal and abnormal frames