Explainable AI frameworks for Time-series in healthcare - A Comparative Evaluation

Aagaaz Ali Sayed
A59026423
asayed@ucsd.edu

Shrihari Jhawar
A59026575
sjhawar@ucsd.edu

Asher Jacob
A59026368
asjacob@ucsd.edu

Ahmed Mostafa
A17401383
ahmostafa@ucsd.edu

Abdullah Ashfaq
A59026296
aashfaq@ucsd.edu

1 PROBLEM STATEMENT

Our project aims to evaluate and identify the most effective explainable AI methods for time series data in healthcare by experimenting with causal-based, backpropagation-based, perturbation-based, and surrogate-based techniques. Through this comprehensive analysis, we aim to identify the best possible approach for this domain.

2 INTRODUCTION

Time-series interpretability enables understanding of how deep learning models derive predictions from sequential data. Unlike static data, time-series data involves temporal dependencies, requiring interpretability approaches that can capture how patterns evolve over time. This capacity is crucial in domains where insight into the decision-making process is as important as the predictions themselves.

In healthcare, where time-series data such as ECG and EEG are fundamental, explainable AI (XAI) is essential for safe and effective decision-making. Lack of model transparency can hinder trust, especially in critical health decisions [9]. By making AI-driven insights more interpretable, healthcare professionals can better trust and integrate these tools, ensuring that predictions align with clinical expectations and patient safety.

Interpretability methods for time-series can be broadly categorized into post-hoc and ante-hoc approaches. Post-hoc methods, such as backpropagation and perturbation based techniques, explain model behavior after training, while ante-hoc approaches integrate interpretability during model development. Traditional methods often highlight associations but fall short of identifying causal factors. In healthcare, causal-based techniques offer more meaningful insights, helping isolate factors that may influence patient outcomes more directly than correlation-focused methods.

3 METHODOLOGY / DESIGN

Our project focuses on evaluating four key explainable AI (XAI) methodologies—backpropagation-based, perturbation-based, surrogate-based, and causal-based techniques—to determine the most effective approach for interpreting time-series models in healthcare. By leveraging real-world and synthetic datasets, we aim to identify the method that best clarifies model behavior for critical applications in this domain.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3.1 Dataset

We will utilize the ECG5000 and Epileptic Seizure Recognition datasets, chosen for their relevance in healthcare and their complexity in time-series analysis. The ECG5000 dataset contains 5,000 heartbeats from the MIT-BIH Long-Term ECG Database, with each sequence represented by 140 time steps. The Epileptic Seizure Recognition dataset consists of EEG recordings from 500 subjects, each sampled over 23.6 seconds at 173.61 Hz, yielding 4097 data points per recording. Additionally, we may incorporate a synthetic dataset to streamline evaluation, as the data-generating mechanism is known and controllable, allowing for clear benchmarking of explainability methods.

3.2 Approaches

(1) **Backpropagation-Based Techniques**: Backpropagation-based methods use gradient information to identify which time-series segments most significantly influence model predictions. These techniques, such as saliency maps [8] and integrated gradients [7], trace the flow of model outputs back to inputs, highlighting important features. We will apply these methods to both ECG and EEG datasets to reveal which heartbeat or brainwave segments are most influential in the classification. This interpretability approach enables us to understand which features play a critical role in predictions, making model behavior more transparent and aligned with human judgment in healthcare.
(2) **Perturbation-Based Techniques**: Perturbation-based methods improve interpretability by modifying or occluding certain input features and observing the change in model output. For instance, by masking segments of the time series (such as specific ECG or EEG time steps) and analyzing the resulting drop in classification accuracy, we can identify which parts of the data are essential for accurate predictions. Techniques like feature erasure [4], occlusion sensitivity [9], and counterfactual generation [10][2] will be utilized to create alternative scenarios, providing insights into how subtle changes in the data can impact predictions. This approach is particularly useful for understanding model robustness and identifying decision boundaries.
(3) **Causal-Based Techniques (CausalConceptTS)**: Causal-based techniques, such as CausalConceptTS [1], go beyond traditional association-based interpretability by focusing on identifying causal factors within the data. By distinguishing causative factors from mere correlations, causal-based

methods can identify which features have a direct influence on classification outcomes. CausalConceptTS employs advanced techniques, including diffusion models, to estimate counterfactuals—enabling us to simulate alternative scenarios and observe how certain factors may impact predictions. This approach is especially valuable in healthcare, where understanding causality can inform targeted treatments and interventions, providing actionable insights beyond associations.

(4) **Surrogate-Based Techniques (TimeX)**: Surrogate-based methods involve training an interpretable model to approximate the behavior of a more complex model, preserving latent relationships in the time series. TimeX [5], a surrogate model designed for time-series data, replicates the predictions of the original model while offering simpler, more interpretable representations. This surrogate method generates discrete attribution maps and provides a latent space for explanations, allowing us to visualize temporal patterns and assess consistency with the underlying model. By approximating the model's decision-making process in an interpretable way, TimeX enables us to understand the factors influencing predictions without compromising model fidelity.

Together, these approaches will provide a multifaceted evaluation of interpretability methods for time-series healthcare data, allowing us to systematically compare their effectiveness and determine the approach that offers the most reliable and actionable insights

3.3 Evaluation

Evaluating the effectiveness of explainability methods in time-series analysis is an active research area. For our project, we will utilize established quantitative evaluation metrics to assess how well each explainable AI approach provides meaningful insights into model behavior. Our evaluation will include three primary metrics:

(1) **Perturbation Analysis**: By occluding the most relevant features identified by the model and observing the resulting drop in accuracy, we can measure how crucial these features are for prediction, helping validate the explainability method's focus on essential data segments [6].
(2) **Precision and Recall**: We will assess the model's ability to identify salient features, using precision and recall metrics to ensure that all identified features are informative and all informative features are flagged as important. This approach offers insight into the explainability method's completeness and accuracy [3].
(3) **AUC and F1 Modifications**: Adjusted AUC and F1 scores will allow us to evaluate how well models rank time steps by importance, ensuring that the explainability method aligns with the temporal significance of data features in time-series contexts [11].

These quantitative metrics will enable us to systematically compare backpropagation, perturbation, causal-based [1], and surrogate methods [5], identifying the approach that provides the most accurate, actionable insights in healthcare applications.

REFERENCES

[1] Juan Miguel Lopez Alcaraz and Nils Strodthoff. 2024. CausalConceptTS: Causal Attributions for Time Series Classification using High Fidelity Diffusion Models. (2024). arXiv:cs.LG/2405.15871 https://arxiv.org/abs/2405.15871

[2] Eoin Delaney, Derek Greene, and Mark T. Keane. 2021. Instance-based Counterfactual Explanations for Time Series Classification. (2021). arXiv:cs.LG/2009.13211 https://arxiv.org/abs/2009.13211

[3] Aya Abdelsalam Ismail, Mohamed Gunady, Héctor Corrada Bravo, and Soheil Feizi. 2020. Benchmarking Deep Learning Interpretability in Time Series Predictions. (2020). arXiv:cs.LG/2010.13924 https://arxiv.org/abs/2010.13924

[4] Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Understanding Neural Networks through Representation Erasure. (2017). arXiv:cs.CL/1612.08220 https://arxiv.org/abs/1612.08220

[5] Owen Queen, Tom Hartvigsen, Teddy Koker, Huan He, Theodoros Tsiligkaridis, and Marinka Zitnik. 2023. Encoding Time-Series Explanations through Self-Supervised Model Behavior Consistency. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 32129–32159. https://proceedings.neurips.cc/paper_files/paper/2023/file/65ea878cb90b440e8b4cd34fe0959914-Paper-Conference.pdf

[6] Udo Schlegel and Daniel A. Keim. 2023. A Deep Dive into Perturbations as Evaluation Technique for Time Series XAI. (2023). arXiv:cs.LG/2307.05104 https://arxiv.org/abs/2307.05104

[7] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2017. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. (2017). arXiv:cs.LG/1605.01713 https://arxiv.org/abs/1605.01713

[8] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. (2014). arXiv:cs.CV/1312.6034 https://arxiv.org/abs/1312.6034

[9] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. 2017. Clinical Intervention Prediction and Understanding using Deep Networks. (2017). arXiv:cs.LG/1705.08498 https://arxiv.org/abs/1705.08498

[10] Sana Tonekaboni, Shalmali Joshi, David Duvenaud, and Anna Goldenberg. 2020. Explaining Time Series by Counterfactuals. (2020). https://openreview.net/forum?id=HygDF1rYDB

[11] Hugues Turbé, Mina Bjelogrlic, Christian Lovis, and Gianmarco Mengaldo. 2023. Evaluation of post-hoc interpretability methods in time-series classification. *Nature Machine Intelligence* 5, 3 (01 Mar 2023), 250–260. https://doi.org/10.1038/s42256-023-00620-w