Explainable AI: An Empirical Analysis of Attribution Methods for Time Series Classification

Yesh Lohchab (Roll No: 230102115)
yesh.3119@iitg.ac.in

Amanganti Chethan Reddy (Roll No: 230102117)
r.amanaganti@iitg.ac.in

Garvil Sethi (Roll No: 230123082)
s.garvil@iitg.ac.in

Mihir Hitesh Birani (Roll No: 230123035)
b.mihir@iitg.ac.in

***Abstract*—Explainable Artificial Intelligence (XAI) provides insight into complex machine learning models, offering transparency and interpretability in domains where decision accountability is essential. This work investigates the application of several XAI techniques on a deep learning model trained for time-series classification. Drawing inspiration from Bayesian robustness-centric approaches, we explore the effectiveness of saliency-based attribution methods including Saliency Maps, Integrated Gradients, and DeepLIFT, applied to a convolutional neural network (CNN) trained on the FordA dataset. We also present the importance of model selection through Randomized Search for hyperparameter tuning to ensure high performance and interpretability.**

I. Introduction

In recent years, the integration of deep learning models into real-world applications has raised critical concerns about their transparency and interpretability. While these models exhibit exceptional accuracy, especially in high-dimensional and complex tasks such as image recognition and time-series classification, they often function as black boxes with little insight into their decision-making processes.

Explainable Artificial Intelligence (XAI) has emerged as a vital subfield aimed at developing tools and methods that make the predictions of machine learning models understandable to humans. XAI is particularly important in safety-critical and regulated domains such as healthcare, finance, and autonomous systems, where understanding why a model made a specific decision can be as important as the decision itself.

Our project builds upon the ideas proposed in the paper titled *"Bayesian XAI Methods: Towards a Robustness-Centric Approach to Deep Learning"*, which emphasizes evaluating the reliability of explanations under uncertainty. While their application domain was medical imaging, we adapt similar methodology for time-series classification tasks, aiming to derive meaningful insights from model predictions and to evaluate how different XAI methods explain the learned representations of time-series data.

II. Model Selection and Hyperparameter Tuning

A critical component in building interpretable machine learning systems is choosing the right model architecture. A poorly tuned model not only underperforms in accuracy but also generates noisy and misleading explanations. Thus, before delving into XAI techniques, we conducted an extensive hyperparameter tuning procedure to select the best-performing model for subsequent analysis.

We adopted `RandomizedSearchCV` from the `scikit-learn` library to explore a wide range of hyperparameters. The hyperparameters tuned included the number of convolutional filters, kernel sizes, number of layers, dropout probabilities, and batch sizes. This approach allowed for efficient exploration of the hyperparameter space while avoiding the combinatorial explosion associated with grid search.

During each iteration, the training dataset was split into training and validation subsets. The model was trained and evaluated based on classification accuracy and validation loss. The best performing model was selected using the model with the lowest

validation loss and highest accuracy. This model achieved a robust trade-off between predictive performance and generalization, providing a reliable base for interpretability experiments.

III. Implementation

The FordA dataset, taken from the UCR Time Series Archive, consists of one-dimensional signals representing automotive sensor readings. The dataset is labeled into two classes and is known for its non-trivial decision boundaries, making it an ideal candidate for interpretability research.

We began by standardizing the input features using `StandardScaler` to ensure zero mean and unit variance across the dataset. Label encoding was also performed to convert the original labels {-1, 1} into a binary format {0, 1} for compatibility with PyTorch models.

Our model architecture comprises multiple 1D convolutional layers interleaved with ReLU activations and max pooling operations. These layers were followed by a series of fully connected layers. The model was trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 64. Early stopping was employed based on validation loss to prevent overfitting.

To probe the internal workings of the trained model, we applied several XAI methods using the Captum library:

- **Saliency Maps**: Compute the gradient of the output class with respect to input features, identifying which time steps most influence the prediction.
- **Integrated Gradients**: Average gradients along a path from a baseline input to the actual input, offering more stable and smoother attributions.
- **DeepLIFT**: Assigns contribution scores by comparing the activation of neurons to a reference input, highlighting both positive and negative influences.

These methods were applied on test samples to evaluate the consistency and interpretability of explanations across correct and incorrect predictions.

IV. Results and Discussion

The attribution heatmaps generated by the different XAI methods offered several insights. Saliency Maps produced sharp and highly localized gradients. While they were computationally simple and fast to generate, their explanations were often noisy and difficult to interpret, especially when gradients were sparse or concentrated around high-frequency patterns.

Integrated Gradients offered smoother attribution profiles by integrating over multiple inputs. They consistently highlighted broader temporal regions associated with the predicted class. This made them more interpretable and aligned with expected patterns in the signal. DeepLIFT provided high-contrast attribution maps, particularly useful in identifying which features contributed positively or negatively to a class decision.

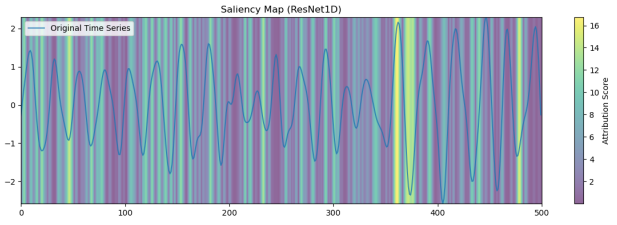

Fig. 1: Saliency Map Attribution on a Correctly Classified Sample

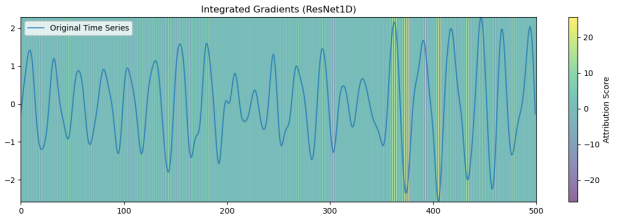

Fig. 2: Integrated Gradients Attribution on the Same Input

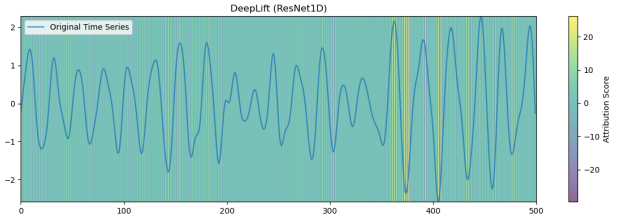

Fig. 3: DeepLIFT Attribution Emphasizing Contrasting Regions

For correctly classified examples, there was significant alignment across the attribution methods in identifying important regions. This indicates model consistency and supports the reliability of the learned features. On misclassified examples, however, the explanations were either too diffuse or focused on irrelevant segments, hinting at the model's uncertainty or misgeneralization. These cases point to the need for integrating uncertainty-aware methods like Bayesian neural networks for better robustness.

V. Conclusion

This study presented a comprehensive framework for implementing and evaluating attribution-based XAI methods on time-series classification problems. Through rigorous hyperparameter tuning, a performant and generalizable CNN model was selected as the base classifier. Using Saliency Maps, Integrated Gradients, and DeepLIFT, we generated visual explanations that revealed meaningful patterns behind the model's decisions.

Our findings suggest that smooth and reference-based methods like Integrated Gradients and DeepLIFT outperform basic gradient-based approaches in producing interpretable explanations. This aligns with existing literature and emphasizes the importance of explanation reliability in practice.

In future work, we aim to incorporate Bayesian uncertainty modeling to extend this study and align it more closely with the robustness-centric ideas proposed in our reference paper.

VI. Contributions

Yesh Lohchab: Code Implementation, Research Paper analysis, Git pushes, Mathematical realisations.

Amanaganti Chethan Reddy: Code Implementation, Git pushes, Hyperparameter tuning.

Garvil Sethi: Report, Research Paper Analysis, Mathematical realisations, Tuning.

Mihir Birani: Report, Research Paper Analysis.

References

[1] U. Budhwani, et al., "Bayesian XAI Methods: Towards a Robustness-Centric Approach to Deep Learning: An ABIDE I Study," in *arXiv preprint arXiv:2207.02688*, 2022.

[2] K. Kokhlikyan, et al., "Captum: A unified and generic model interpretability library for PyTorch," *arXiv preprint arXiv:2009.07896*, 2020.

[3] H. Dau, et al., "The UCR Time Series Classification Archive," 2019. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/

[4] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.