RIT

Explainable AI

Research Project Contribution

By
Srujan Vaddiparthi

Agenda

1. Introduction and Motivation
2. Research Questions
3. Literature Review
4. Paper 1: "Notions of Explainability and Evaluation Approaches for Explainable Artificial Intelligence"
5. Paper 2: "Explanations in Terms of Hierarchically Organised Middle Level Features"
6. Paper 3: "Feature Importance Explanations for Temporal Black-Box Models"
7. Paper 4: "A Hierarchical Explanation Generation Method Based on Feature Interaction Detection"
8. Personal Contributions
9. What I Learned
10. Future Work
11. Conclusion
12. References

Introduction and Motivation

I'll cover the research questions I've developed, the papers I've reviewed, my personal contributions, and what I've learned throughout this project.

Enhancing explainability in cybersecurity

- Complexity of AI models in cybersecurity
    - AI models used in Network Intrusion Detection Systems (NIDS) are increasingly complex.
- Need for explainability
    - Sysanalysts need to interpret model decisions for effective responses.
- Hierarchical Explanations, what are they?
    - They break down model decisions into different levels.
    - Provide both high-level overviews and detailed specifics.

Research Questions

- Hierarchical Explanations for complex data
  - Question: How can hierarchical explainability methods improve interpretability and decision-making in Network Intrusion Detection Systems (NIDS) by enabling sysanalysts to access both high-level insights and low-level feature-specific explanations?
  - Goal: Design explanations that are both broad and specific, helping analysts understand general patterns and drill down into detailed feature interactions when needed.
- Limitations of existing methods
  - Question: What are the limitations of existing explainability methods in handling temporal dependencies and feature correlations within cybersecurity datasets, and how can these be addressed to provide more accurate and actionable insights?
  - Goal: Improve the real-time usefulness of XAI models in detecting evolving threats by addressing challenges around temporal data and feature correlation.
- Actionable insights for sysAnalysts
  - Question: How can explainable AI models be adapted or developed to generate insights that are actionable for sysanalysts, rather than merely visualizing model behavior, particularly for anomaly detection and intrusion prevention?
  - Goal: Create explanations that translate directly into actions, helping analysts take immediate steps rather than just interpreting model outputs.
- Usability challenges
  - Question: What usability challenges arise when SOC analysts use hierarchical explanations in real-world cybersecurity contexts, and how can we measure and optimize the cognitive load, response time, and confidence levels in their decision-making?
  - Goal: Explore how hierarchical explanations impact usability and confidence for SOC analysts, especially when handling complex cybersecurity events.
- Robustness and Reliability
  - Question: How robust are hierarchical explanation models in replicating consistent explanations across similar network security events, and what evaluation metrics can be established to measure robustness and reliability for cybersecurity applications?
  - Goal: Establish metrics that measure consistency and reliability in explanations, crucial for consistent accuracy in cybersecurity models.

Literature review overview and my purpose..

Some papers I had reviewed which felt were important.

- "Notions of Explainability and Evaluation Approaches for Explainable Artificial Intelligence"
- "Explanations in Terms of Hierarchically Organised Middle Level Features"
- "A Hierarchical Explanation Generation Method Based on Feature Interaction Detection"
- “Feature Importance Explanations for Temporal Black-Box Models”

My Purpose:

- Explore hierarchical explanations in AI.
- Understand how they can be applied to cybersecurity.
- Identify gaps and opportunities for further research.

Paper 1: Notions of Explainability and Evaluation Approaches for Explainable Artificial Intelligence

What does it contribute?:

- Categorizes evaluation methods into objective evaluations and human-centered evaluations.
- Analyzes various metrics and methodologies used to assess the quality of explanations, including quantitative and qualitative measures.
- Defines and clusters various notions and attributes related to explainability, such as interpretability, transparency, understandability, causality, and more, which is helpful in framing research questions.
- Discusses the attributes of explainability and theoretical approaches for structuring explanations.
- Highlights limitations of existing methods and emphasizes on human centered design. Focuses on developing hierarchical explanations that improve interpretability and decision-making for SOC analysts.

- How is it relevant?:
  - Highlights the importance of tailoring explanations to SOC analysts' needs.
  - Inspired consideration of user-centric design in hierarchical explanations.

Paper 2: Explanations in Terms of Hierarchically Organised Middle Level Features

- Limitations of Existing Explanations:
  - Traditional XAI methods like Layer-wise Relevance Propagation (LRP) provide explanations at the low-level feature (e.g., pixel) level, placing a heavy interpretive burden on users to identify meaningful patterns.
  - Concept Activation Vectors (CAVs) introduce high-level concepts from external labeled datasets to explain model decisions, but they may not directly relate to specific input instances and can lead to interpretive challenges or misleading explanations if the model is unreliable.
- Identified Gap:
  - There's a need for explanations that are directly tied to the input data and more interpretable than low-level features.
  - Hierarchical organization of explanations can help users understand model decisions at different levels of granularity.
- Key Contributions:
  - Middle-Level Features (MLFs):
    - Propose using MLFs derived directly from the input, representing perceptually salient parts (e.g., bird heads, car wheels).
    - Avoid reliance on external concepts, enhancing relevance to specific inputs.
  - Hierarchical Organization:
    - Organize MLFs hierarchically to provide explanations at multiple levels of detail.
    - Reduces cognitive load by aligning explanations with human perception.
- Methodology:
  - Hierarchical Segmentation:
    - Break down images into hierarchically organized segments.
  - Autoencoder Framework:
    - Use an autoencoder where decoder layers correspond to hierarchical levels of MLFs.
  - Layer-wise Relevance Propagation (LRP):
    - Apply LRP to identify which MLFs at each level are most relevant to the prediction.

Paper 2: Explanations in Terms of Hierarchically Organised Middle Level Features

- What were the key takeaways:
  - Improved interpretability:
    - Explanations align better with human perception by focusing on MLFs derived from the input itself.
    - Hierarchical organization helps users understand model decisions at both global and local levels.
  - Addresses limitations of previous methods:
    - Unlike CAVs, their method doesn't rely on external datasets or predefined high-level concepts.
    - Avoids potential biases introduced by external concept definitions.
    - Directly relates explanations to the specific input instance, enhancing relevance and accuracy.
  - Reduces cognitive load:
    - By presenting explanations at different granularities, users can focus on the level of detail that suits their needs.
    - Makes it easier for users to comprehend and act upon the explanations.
- Relevance to My Research:
  - Hierarchical Explanations:
    - Validates the importance of hierarchical explanations for better interpretability in cybersecurity.
    - Supports my goal of helping SOC analysts understand threats at multiple levels.
  - User-Centric Design:
    - Emphasizes reducing cognitive load by presenting explanations directly tied to input data.
    - Enhances trust and usability through perceptually salient explanations.
  - Applicability to Cybersecurity:
    - Provides a framework for hierarchical explanations in cybersecurity contexts, such as network patterns and multi-stage attacks.
    - Assists SOC analysts in understanding both high-level attack strategies and specific indicators of compromise.
- This paper offers a method that combines hierarchical organization with Middle-Level Features directly derived from input data, hence provides valuable insights for developing effective, user-centric explanations in cybersecurity; aligning closely with the research objectives.

Paper 3: A Hierarchical Explanation Generation Method Based on Feature Interaction Detection

- Limitations of Existing Hierarchical Explanations:
  - Traditional attribution methods provide explanations at predefined text granularities (e.g., words, phrases), which may miss compositional semantics.
  - Recent hierarchical attribution methods build explanations using continuous text spans, following the connecting rule, where only adjacent words are grouped together.
- Gap Identified:
  - There's a need for hierarchical explanations that can capture long-distance feature interactions without being constrained to continuous text spans.
  - Existing methods do not adequately reflect the decision-making process of models that consider non-local interactions.
- Key Contributions:
  - Novel Hierarchical Explanation Method Without the Connecting Rule:
    - Propose a method that allows grouping of non-adjacent words to capture meaningful feature interactions.
    - Overcomes limitations of previous methods by reflecting the true interaction patterns in transformer-based models.
  - Feature Interaction Detection Strategy:
    - Introduce a strategy to quantify and detect interactions between features (words or phrases) based on their influence on attribution scores.
    - Measures how the presence or absence of one feature affects the attribution score of another.
  - Conversion of Non-Hierarchical Explanations into Hierarchical Versions:
    - Unlike previous methods that rely on specific algorithms, their approach can transform any non-hierarchical attribution method (e.g., LIME, LOO) into a hierarchical explanation.
    - Provides flexibility and broad applicability across different attribution techniques.
- Methodology:
  - Detecting Feature Interactions:
    - Compute interaction scores between pairs of text groups by measuring how the attribution score of one group changes when another group is removed.
  - Building Hierarchical Explanations:
    - Start with each word as an individual text group.
    - Iteratively merge pairs with the highest interaction scores, regardless of their positions, forming a hierarchy.
    - Continue until all words are merged into a single group.
  - Visualization:
    - Display the newly formed, possibly non-continuous groups and their attribution scores at each hierarchical level.
    - Traditional tree structures aren't feasible due to non-continuity.

Paper 3: A Hierarchical Explanation Generation Method Based on Feature Interaction Detection

- Relevance to My Research:
  - Hierarchical Explanations:
    - Validates the importance of capturing complex feature interactions for interpretability in cybersecurity.
  - Overcoming Limitations:
    - Provides a strategy to enhance methods like SHAP and LIME to handle non-local interactions.
  - Applicability to Cybersecurity:
    - Cybersecurity Data Involves Complex Interactions:
      - In cybersecurity, we often analyze data where important signals or indicators of threats are spread out over different times or across different parts of a network.
      - For example, an attacker might perform one action on Monday and a related action on Wednesday, or they might interact with one server and then with another.
    - Traditional Methods might not capture important relationships between features that are far apart in time or space.
    - Their Method Captures Non-Local Interactions:
      - The paper’s method allows us to detect and explain interactions between features that aren't next to each other.
      - It identifies how features from different times or parts of the network influence each other and the model's predictions.
    - Improving Threat Detection and Response:
      - By understanding these complex, spread-out interactions, security analysts can get a clearer picture of potential threats.
      - It leads to better and faster responses because analysts can see how separate events are connected.
    - In Simple Terms:
      - Think of cybersecurity threats as puzzle pieces scattered all over.
      - Traditional explanations might only look at pieces that are side by side.
      - But this new method helps us see how pieces from different parts fit together to complete the puzzle.
      - It makes it easier to spot and understand threats that aren't immediately obvious.

Paper 4: Feature Importance Explanations for Temporal Black-Box Models | https://github.com/Craven-Biostat-Lab/anamod

- Given a learned model and a hierarchy over features, (i) it tests feature groups, in addition to base features, and tries to determine the level of resolution at which important features can be determined, (ii) uses hypothesis testing to rigorously assess the effect of each feature on the model's loss, (iii) employs a hierarchical approach to control the false discovery rate when testing feature groups and individual base features for importance, and (iv) uses hypothesis testing to identify important interactions among features and feature groups.
- Need for Temporal Model Explanation:
  - Temporal models capture complex patterns over time, making their decisions hard to interpret.
  - There's a lack of model-agnostic methods to explain temporal models, especially at a global level (understanding overall model behavior).
- Key Contributions:
  - TIME (Temporal Importance Model Explanation):
    - Proposes a model-agnostic, global explanation method tailored for temporal models.
  - Identifies globally important features and their temporal properties, including:
    - The most important time windows for each feature.
    - Whether the ordering of values within these windows is important.
    - Uses permutation-based hypothesis testing for statistical rigor and controls the false discovery rate.
- Methodology:
  - Permutation-Based Feature Importance:
    - Extends permutation tests to temporal data by permuting entire time windows of features.
    - Measures the change in model loss when a feature or window is permuted to assess importance.
  - Identifying Important Windows:
    - Uses a binary search to find the most important contiguous time window for each feature.
    - The window is the smallest segment where permuting it significantly affects the model's loss.
  - Determining Importance of Feature Ordering:
    - Assesses whether the order of values within the important window matters to the model.
    - Permutes the ordering within the window and observes the impact on model loss.
  - Hypothesis Testing and False Discovery Rate Control:
    - Applies statistical hypothesis testing to determine the significance of feature importance, window importance, and ordering importance.
    - Uses a hierarchical testing framework with false discovery rate control for statistical rigor.

Paper 4: Feature Importance Explanations for Temporal Black-Box Models

- Relevance to My Research:
  - Handling Temporal Dependencies:
    - Addresses the challenge of explaining models that process temporal data with time-varying features.
  - Feature Correlations and Hierarchical Structures:
    - By identifying important time windows and feature ordering, TIME captures feature interactions over time.
    - Inspires methods to handle feature correlations and temporal hierarchies in cybersecurity data.
  - Applicability to Cybersecurity:
    - Temporal Nature of Cyber Threats and Relevance to KAIROS:
      - In cybersecurity, attacks often unfold over time, involving sequences of events or actions.
      - Understanding how threats evolve is crucial for effective detection and response.
      - KAIROS aims to understand complex events by analyzing sequences and patterns over time.
      - This paper's method aligns with KAIROS by explaining temporal models that process such sequences.
    - Applying TIME to Cybersecurity:
      - Identifying Important Features Over Time:
        - TIME helps pinpoint which features (specific network activities or user behaviors) are most significant during certain periods.
        - For instance, it can reveal that unusual login attempts are most critical during off-peak hours.
      - Understanding Feature Ordering:
        - Determines if the sequence of events matters to the model's predictions.
        - In cyber attacks, the order of actions is often important.
    - Benefits for SOC Analysts:
      - Enhanced Interpretability: By explaining when and how certain features influence the model, analysts can better understand complex attack patterns.
      - Improved Threat Detection: Helps in identifying critical time windows where security measures should be intensified.
      - Supports KAIROS Objectives:Aids in constructing narratives of complex cyber events by highlighting key actions over time.
- TIME provides a way to explain temporal models in cybersecurity by showing which features are important at specific times and whether the order of events matters. This aligns with KAIROS's goal of understanding complex sequences of events, helping analysts detect and respond to threats more effectively.

Personal Contributions

- Formulated Key Research Questions:
  - Identified critical challenges in applying hierarchical explainability methods to Network Intrusion Detection Systems (NIDS).
  - Developed focused research questions addressing interpretability, temporal dependencies, feature correlations, and actionable insights for SOC analysts.
- In-Depth Literature Review:
  - Conducted comprehensive analysis of existing hierarchical explanation methods.
  - Evaluated the applicability of these methods to cybersecurity contexts.
  - Identified limitations in handling temporal data and complex feature interactions.
- Identified Gaps and Opportunities:
  - Highlighted the inadequacy of current explainability methods in capturing temporal dependencies and non-local feature interactions in cybersecurity data.
  - Recognized the need for user-centric design to reduce cognitive load and enhance decision-making for SOC analysts.
- Proposed Conceptual Framework:
  - Developed initial ideas for a hierarchical explanation model tailored to cybersecurity applications.
  - Emphasized the integration of temporal models and feature interaction detection to improve interpretability.
- Set the Foundation for Future Research:
  - Established groundwork for developing new methods that address identified gaps.
  - Prepared to conduct usability studies to measure the effectiveness and usability of hierarchical explanations in real-world settings.

What I Learned

- Significance of Hierarchical Explanations:
  - Hierarchical explanations can enhance interpretability by providing insights at multiple levels, from high-level overviews to detailed specifics.
  - They align closely with human cognitive processes, making complex model decisions more understandable.
- Limitations of Existing Methods:
  - Many current explainability methods like SHAP and LIME are not well-suited for temporal data or capturing non-local feature interactions.
  - These limitations hinder their effectiveness in domains like cybersecurity, where data is complex and time-dependent.
- Importance of User-Centric Design:
  - Explanations must be tailored to the needs of the end-users (SOC analysts) to be truly effective.
  - Reducing cognitive load and providing actionable insights are crucial for practical usability.
- Challenges with Temporal and Sequential Data:
  - Temporal dependencies and the ordering of events are critical in cybersecurity but often overlooked in explainability methods.
  - Capturing these aspects is essential for accurate and meaningful explanations.
- Need for Statistical Rigor and Robustness:
  - Incorporating statistical methods like hypothesis testing and controlling the false discovery rate improves the reliability of explanations.
  - Ensuring consistency and robustness is vital for trust in AI models.
- Alignment with KAIROS Objectives:
  - Understanding complex sequences of events is central to both KAIROS and the development of effective hierarchical explanations.
  - There is significant potential for synergy between hierarchical explanation methods and KAIROS's goals.

Future Work

- Develop Hierarchical Explanation Framework for NIDS:
  - Design and implement methods that provide hierarchical explanations specifically tailored for network intrusion detection.
  - Focus on integrating temporal dependencies and feature interactions into the explanation framework.
- Incorporate Temporal Dependencies and Feature Interactions:
  - Adapt existing methods or develop new ones to handle time-varying features and sequences of events common in cybersecurity data.
  - Utilize techniques like those proposed in TIME and feature interaction detection to capture complex patterns.
- User Studies with SOC Analysts:
  - Conduct usability studies to assess the effectiveness of hierarchical explanations in real-world cybersecurity scenarios.
  - Measure cognitive load, response time, and confidence levels to optimize explanation design.
- Establish Robustness and Reliability Metrics:
  - Develop evaluation metrics to measure the consistency and reliability of explanations in cybersecurity applications.
  - Ensure that explanations are replicable across similar network security events.
- Integration with KAIROS Project:
  - Explore collaboration opportunities with KAIROS to enhance understanding of complex event sequences.
  - Align hierarchical explanation methods with KAIROS's objectives to mutually benefit both projects.
- Implement Actionable Insight Generation:
  - Focus on creating explanations that translate directly into actionable steps for SOC analysts.
  - Ensure that explanations not only interpret model outputs but also guide effective responses.

Conclusion

- Summary of Findings:
  - Hierarchical explanations have significant potential to improve interpretability and decision-making in cybersecurity.
  - Existing methods have limitations in handling temporal dependencies and complex feature interactions.
- Addressing Research Questions:
  - Identified the need for hierarchical methods that provide both high-level insights and detailed explanations.
  - Recognized challenges in existing explainability methods and the importance of user-centric design.
- Path Forward:
  - Committed to developing new hierarchical explanation frameworks tailored to the needs of SOC analysts.
  - Aiming to incorporate temporal models, feature interactions, and statistical rigor into explanations.
- Final Thoughts:
  - Combining hierarchical explanations with user-centric design and robustness will significantly enhance the effectiveness of AI models in cybersecurity.
  - Collaboration and continuous evaluation are key to advancing explainable AI in this complex domain.
- Everything was uploaded to our git repository: https://github.com/dd9098/XAI-CyberSec.git

References

1. Apicella, Andrea et al. “Explanations in terms of Hierarchically organised Middle Level Features.” (2021).
2. Yiming Ju, Yuanzhe Zhang, Kang Liu, and Jun Zhao. 2023. A Hierarchical Explanation Generation Method Based on Feature Interaction Detection. In Findings of the Association for Computational Linguistics: ACL 2023, pages 12600–12611, Toronto, Canada. Association for Computational Linguistics.
    a. https://github.com/juyiming/HE_examples/tree/master
3. Sood, A., & Craven, M. (2021). Feature Importance Explanations for Temporal Black-Box Models. arXiv preprint arXiv:2102.11934.
    a. https://github.com/Craven-Biostat-Lab/anamod
4. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
5. Vilone, Giulia, and Luca Longo. "Notions of explainability and evaluation approaches for explainable artificial intelligence." Information Fusion 76 (2021): 89-106.
6. Kairos: Practical Intrusion Detection and Investigation using Whole-system Provenance. Zijun Cheng, Qiujian Lv, Jinyuan Liang, Yan Wang, Degang Sun, Thomas Pasquier, Xueyuan Han
    a. https://github.com/ProvenanceAnalytics/kairos.git
7. https://github.com/dd9098/XAI-CyberSec.git