IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 11, November 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.524

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311191|

Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research

Shobha Agasibagil, Arpitha L

Assistant Professor, Department of Computer Science and Engineering, CIT, Gubbi, Tumkur, Karnataka, India

U.G. Student, Department of Computer Science and Engineering, CIT, Gubbi, Tumkur, Karnataka, India

ABSTRACT: The paper *"Explainable Artificial Intelligence (XAI) Applications in Cyber Security: State-of-the-Art in Research"* delves into how XAI enhances trust and usability in AI-driven cybersecurity systems by providing insights into model decision-making. It emphasizes the critical need for transparency in cybersecurity, where decisions made by AI systems—such as identifying malware or detecting phishing—must be understandable to human operators. The lack of interpretability in traditional machine learning models poses challenges, especially as cyber threats evolve and become more sophisticated. This survey addresses the gap by summarizing popular datasets, tools, and techniques for applying XAI in defending against different types of cyberattacks, including malware, botnets, and phishing. Moreover, the study highlights the risks posed by adversarial attacks targeting AI models themselves and stresses the importance of building robust, explainable systems that can defend against such threats. By providing human-readable explanations, XAI not only improves the trustworthiness of cybersecurity systems but also enhances their overall effectiveness. The paper outlines the balance between maintaining high accuracy and providing explanations, a key challenge in the field. It concludes by identifying future research directions, such as integrating XAI with emerging technologies and developing standardized evaluation frameworks for explainability in cybersecurity

KEYWORDS: Explainable Artificial Intelligence (XAI), Cybersecurity, Machin Learning (ML), Deep Learning (DL), Intrusion Detection Systems (IDS), Malware Detection, Phishing Detection,Adversarial Attacks, Transparency.

I. INTRODUCTION

The introduction of the paper *"Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research"* highlights the critical need for integrating Explainable AI (XAI) in cyber security systems. With the increasing reliance on AI, particularly machine learning (ML) and deep learning (DL), for tasks like intrusion detection and malware analysis, there is a significant challenge due to the "black-box" nature of these models. This lack of interpretability makes it difficult for security professionals to understand how decisions are made, reducing trust and confidence in AI systems. The authors emphasize the importance of XAI in overcoming these limitations by creating models that are not only accurate but also interpretable. They note that while there are extensive reviews of AI in cybersecurity and XAI in other fields like healthcare and finance, a comprehensive survey of XAI specifically in cybersecurity is missing. This paper aims to fill that gap by providing a roadmap for the application of XAI techniques in cybersecurity, promoting both transparency and trust in automated defense mechanisms

II. RELATED WORK

Surveys and Frameworks: Comprehensive reviews have been conducted on the use of XAI techniques like SHAP and LIME to explain AI models used in cybersecurity, emphasizing model interpretability and trust.

Intrusion Detection Systems (IDS): XAI approaches help in making AI-driven IDS more transparent, enabling analysts to understand why specific network activities are flagged as threats.

Malware Detection: Studies highlight the role of XAI in identifying critical features that influence AI models' decisions to classify files as malicious or benign, aiding in malware analysis and detection.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311191|

Adversarial Attacks: Research on adversarial attacks explores how XAI can be used to detect and understand vulnerabilities in AI models, improving defenses against adversarial manipulations.

Balancing Accuracy and Interpretability: Several works discuss the trade-off between model performance and explainability, proposing hybrid models that maintain a balance between these aspects.

Human-Centered AI: Emphasis on human-centered approaches in cybersecurity, where XAI improves collaboration between AI systems and security analysts by providing actionable explanations.

Future Directions: Suggestions for future research include integrating XAI with real-time systems, visual analytics, and exploring its use in emerging technologies like blockchain and edge computing

III. METHODOLOGY

1. **Black-Box Model Analysis:**

   Challenge: AI models, especially deep learning models, often function as black boxes, meaning their internal decision-making processes are opaque.

- **Solution:** Techniques like SHAP **(SHapley Additive Explanations)** and Layer-Wise Relevance Propagation (LRP) are used to break down and visualize how models make decisions, pinpointing influential factors in predictions

2. **Interpretable Machine Learning (IML):**

   Approach: IML focuses on using inherently interpretable models (like decision trees) or post-hoc explanation methods to clarify complex models.

- **Example Techniques:**
  - **Saliency Maps:** Highlight which input features (like parts of an image or text) are pivotal for a prediction.
  - **Semantic Graphs:** Capture the relationships and hierarchy within neural networks to reveal how different layers contribute to decision-making

3. Types of Interpretability:

Global Interpretability: Understanding the overall behavior of a model across all predictions.

- **Local Interpretability:** Explaining individual predictions to understand why specific decisions were made in certain contexts.

4. Use in Cybersecurity: XAI methods are applied to detect malware, identify phishing attacks, and monitor network traffic anomalies. They allow security analysts to trust AI outputs by making the decision process transparent and auditable.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311191|

IV. EXPERIMENTAL RESULTS

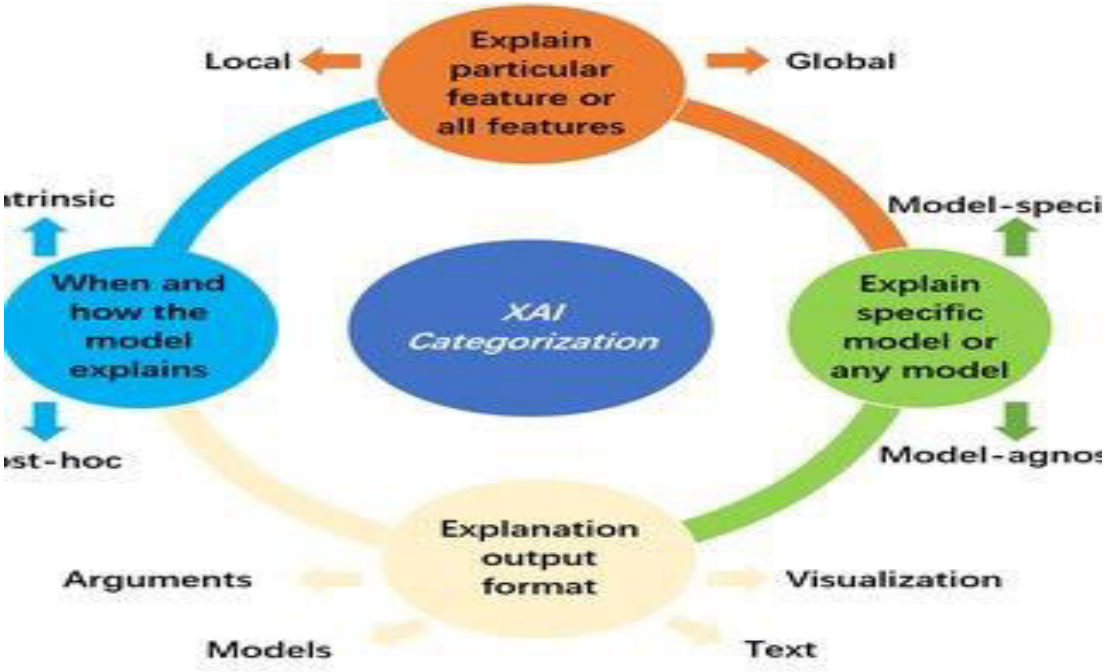

FIGURE 1: EXPLAINABLE ARTIFICIAL INTELLIGENCE APPLICATIONS IN CYBER SECURITY

V. CONCLUSION

The **conclusion** of the paper *"Explainable Artificial Intelligence (XAI) Applications in Cyber Security: State-of-the-Art in Research"* emphasizes the importance of integrating explainability into AI models used in cybersecurity. The authors underline that while AI has significantly improved threat detection and risk management, the lack of transparency in "black-box" models poses challenges for trust, accountability, and decision-making. The paper advocates for adopting XAI techniques to enhance model interpretability, helping cybersecurity professionals understand and trust AI-generated insights. It also highlights the need for further research to develop hybrid models that balance accuracy and explainability, as well as standardized frameworks for evaluating the effectiveness of XAI in cybersecurity.

REFERENCES

1. G. Samek, T. Wiegand, and K. Müller, "Explainable AI: Understanding, visualizing and interpreting deep learning models," *Frontiers in Robotics and AI*, vol. 7, pp. 1-21, 2020.
2. D. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier," in *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144, 2016.
3. S. Bhattacharyya, S. Jha, and J. West, "Data mining for credit card fraud detection – A survey," in *Proc. Int. Conf. Communication and Signal Processing*, pp. 345-352, 2011.
4. L. A. Bazzan, "Artificial intelligence in cybersecurity," *Journal of the Brazilian Computer Society*, vol. 18, no. 2, pp. 45-62, 2012.
5. T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1-38, 2019.
6. W. Zhang, L. Xie, and H. Li, "Deep learning in cybersecurity: A survey," *Future Generation Computer Systems*, vol. 80, pp. 11-27, 2018.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311191|

7. S. M. LaPlante, "Explainable AI and its role in cybersecurity decision support," *Journal of Cybersecurity*, vol. 14, no. 3, pp. 105-117, 2020.
8. L. S. R. S. Ali, R. K. Gupta, and M. Gupta, "A survey of explainable AI techniques for cybersecurity," *IEEE Access*, vol. 9, pp. 17187-17206, 2021.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details