Applying Explainable Artificial Intelligence in the Field of Cybersecurity

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Jacqueline Lainhart

Spring 2024

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Richard D. Jacques
Department of Engineering and Society

Introduction

As more technologies arise and more people are integrating tech into their day-to-day lives, there are more possibilities and targets for cybercriminals. Defensive measures must be taken place to detect cyber-attacks and vulnerabilities as quickly as they occur. Unfortunately, even the most advanced individuals in cyber defense can overlook weaknesses in code. The more vulnerabilities that go unnoticed, the more susceptible they are cyber attackers. Cyber attackers can then obtain sensitive information about the system and its user and use that for harm (financially and personally) towards the users. However, using something like Artificial intelligence (AI) can help identify these overlooked weaknesses. AI has been a hot topic for the past couple of years in which numerous AI tools have been developed for many reasons. Some AI tools examples include AI for coding (CoPilot) and AI for presentations (Tome). AI has made it easier to develop the applications and systems that the AI is applied to, even if it is simply for generating ideas or suggestions. Combing AI and cybersecurity is something that has also been developing as it is believed the benefits of AI can cross into the cybersecurity realm in terms of identifying potential viruses, attacks, and vulnerabilities. There are agencies, such as the CISA, that work meticulously in integrating the use of AI with cyber defense systems (CISA, 2024). However, with something as sensitive as cybersecurity, blindly trusting the output of an AI can lead to dangerous consequences if the AI has a false negative. To evolve the use of AI in cybersecurity, implementing more explainable artificial intelligence (XAI) tools is the next step forward. XAI is a kind of AI that allows its users to understand how the AI itself works. When applying this to cybersecurity, it could allow for a better understanding of why the machine learning model is making suggestions for addressing cyber vulnerabilities and attacks.

Technical Discussion

One of the key things to a good machine learning model is one that gives a good prediction meaning it has a high prediction score or confidence. Several types of machine learning models have been developed to apply to different scenarios to yield the highest confidence. Confidence is rated from 0 to 1: 1 meaning the model will always predict correctly (Google, 2024). Depending on the system, a prediction score extremely close to one is more desired whereas other systems can get away with a lower number. When using AI for cybersecurity, it is important to have models with a prediction with high confidence, due to the excessive cost of missing an attack or vulnerability in the system. This can result in both monetary loss and the leaking of personal and sensitive information. Currently, the integration of AI with cybersecurity exists with tools such as Tessian, which claim to use AI to help prevent email-based threats (phishing) (Jones & Towse, 2019). There are also tools like Charlotte AI by CrowdStrike that monitor potential threats by using data of prior attacks and cyber criminals (Petronaci & Driggs, 2023). AI is a tool used proactively. Commonly, it is seen in network security and antivirus detection, “The product (AI tool) eliminates the threat exactly at the point of execution without the need for any human intervention” (Srivastava, et al., 2022). As useful as these tools are, the problem arises when cyberattacks on them go unnoticed. Using XAI can help find attacks that are occurring, which otherwise can be nondetectable with current cybersecurity AI tools. Recently, there have been attacks that specifically target these AI systems. Cybercriminals can create malware that goes undetected by an AI system. They performed poisoning attacks, adversarial attacks, and backdoor attacks (Mi, Jiang, Luo, & Gao, 2024). In the case of poison attacks, they are attacks where data gets injected into an AI model to compromise its performance. Therefore, predictions can become more biased or inaccurate.

However, a user might not realize the inaccuracies or why they are occurring. If XAI is used, then the understanding of an AI system and how it reaches its predictions can help cybersecurity analysts improve their system to prevent these anomalies.

Deep Neural Networks (DNN) are a type of machine learning that is meant to mimic the brain (IBM Data and AI Team, 2023). The smallest component of a DNN is a neuron that contains different layers referred to as the node layers: the input layer, the hidden layer, and the output layer. Figure 1 below shows the hierarchical layers containing neuron nodes starting from the input layer to the hidden layer and ending with the output layer (La Rosa, et al., 2023). DNNs are black box in nature which means that you can only insert inputs and get outputs. The inner workings of a model and its algorithms at work remain a mystery. Meanwhile, if it was a white box, then the AI algorithm's behavior can be observed. XAI aims to turn the black box AI into essentially white box systems. The interpretability of a model allows developers to understand it more through the internal workings of the AI model. This would be for people who already understand AI concepts and the AI model. Meanwhile, the explainability of a model allows regular users to understand why the AI makes the decisions it makes. Both aim to build trust in the AI tool, but XAI caters to a broader audience which is ideal (Ali, et al., 2023).

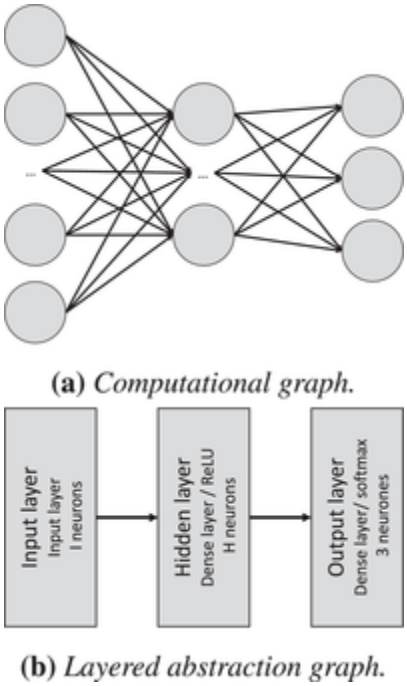

Figure 1. Graphical depiction of the layers of a DNN.

The importance of using XAI is not just about understanding AI systems but about being able to trust them. This is especially for those who are using an AI tool but may not understand much about machine learning in general. If someone were explained to why a conclusion is made, they would give more consideration to trusting the information as opposed to blindly trusting the information; therefore, the user would be making more informed decisions. This can be related to presenting facts about a subject without a source. Since XAI is centered around explaining how an AI system makes its decisions, there needs to be a way to show this to the user. These explanations can be displayed to a user visually, textually, or orally. In addition to having effective explanations of outputs, XAI effectively presents these explanations to the user. It is the marriage between AI and the user interface. According to Grigoryan, for XAI to be recognized as explainable, then it must be able to deliver content, provide clarification, have justification, and establish fidelity. The difference between justification and establishing fidelity

is that justification is about providing evidence to support predictions or decisions while establishing fidelity is to ensure that a machine learning model presents the information in a way that aligns with the model itself and its purpose (Grigoryan, 2022). In terms of cybersecurity, as an example, an AI system could detect network traffic, conclude it is a cyberattack, and justify this claim by saying there were spikes in traffic volume or suspicious communication patterns. It would then further describe where vulnerabilities lie and ways to address those vulnerabilities aligning with standard industry practices. The system must have faithful depictions of the model's behavior while ensuring understandable and transparent insights for the user, "If an explainable AI method's output satisfies 'faithfulness' and 'understandability', it can be further extended for functional purposes such as model understanding, model debugging, and detecting dataset biases" (Sattarzadeh, Sudhakar, & Plataniotis, 2021). XAI was proposed due to AI's unethical usage, lack of transparency, and unintended biases (Ali, et al., 2023).

Currently, XAI or explainable deep learning (XDL) can be classified using various methods such as post-hac methods, model agnostic methods, model specific methods, self-explainable DNNs, and other means of creating an XAI. Another classification for XAI methods is local and global: local refers to explaining the decision of a model using a specific sample of input and global often utilizes an entire dataset (Mi, Jiang, Luo, & Gao, 2024). There are also even more classifications that involve different ways of searching, perturbation-based vs. gradient-based. Post hoc methods are commonly used due to their ability to be applied to current systems without having to change the algorithm. This is because post hoc "receives a trained and/or tested AI model as input, then generates useful approximations of the model's inner working and decision logic by producing understandable representations in the form of feature importance scores, rule sets, heatmaps, or natural language" (Moradi & Samwald, 2021). Ali, et

al. mentions how using post hoc provides clarity for numerous features using various sorts of explanations. The post hoc methods, Latin for “after this,” replicate the workings of a black-box system by uncovering connections between feature values and predictions (Moradi & Samwald, 2021). This type of method can be classified further by whether it is model-specific or model-agnostic. If the post hoc method is model-specific, naturally, it can only be used on that specific model. Meanwhile, model agnostic can be generally applied to any machine learning model (Capuano, Fenza, Loia, & Claudio, 2022). This XAI method is ideal for cybersecurity purposes as many organizations would rather build off existing structures instead of creating new ones as it would take time and resources. There are a variety of post hoc methods used to explain AI models outputs. Some popular methods are LIME, SHAP, GRAD-CAM, and CEM. According to previous research, the areas of cybersecurity that would benefit from utilizing XAI are Intrusion Detection Systems, malware detection, phishing and spam detection, and BotNet detection (Capuano, Fenza, Loia, & Claudio, 2022). When looking at information such as network traffic, it would be important to have the specific post hoc method tailored to the complexities of the data and the model architecture being utilized. The benefit of using Local Interpretable Model agnostic Explanations (LIME) is that LIME can provide explanations for any black-box classifier (Srivastava, et al., 2022), while also providing individual insights making it advantageous due to a higher specificity. Analysts prefer something that produces a more detailed report on the decision-making process and can help in finding the root of a vulnerability quickly (Mi, Jiang, Luo, & Gao, 2024). LIME also works faster than SHAP, which is ideal when dealing with substantial amounts of data. When analyzing a Convolutional neural network (CNN) in an Android app that was used for malware detection, LIME was used to explain their predictions (Kinkead, Millar, McLaughlin, & O'Kane, 2022). Kinkhead, Millar, McLaughlin & O’Kane goes

on to explain that LIME involves sampling data points from both nearby and distant regions relative to the input data point that requires explanation, “Sampling modifies this input data point, and LIME creates an explanation using these sample points by first determining their classifications according to the model to be explained, and then building a weighted linear model of how the input features influence the classification decision” (Kinkead, Millar, McLaughlin, & O'Kane, 2022). They selected LIME due to its previously mentioned fact of it being able to be applied to any black box model. They found that their model had a 0.98 score of accuracy and precision. They observed that LIME's highest activations coincided with those of CNN in the same areas. For neural networks, activations are the output values of individual neurons within the network. In this context, the activations represent the suspected locations within the opcode sequences of the Android apps for malware detection. Figure 2 below shows a graph for the CgFinder malware. CNN and LIME are visibly shown to have peaks in similar locations, indicating the approximated model’s behavior had similar outputs (Kinkead, Millar, McLaughlin, & O'Kane, 2022).

.

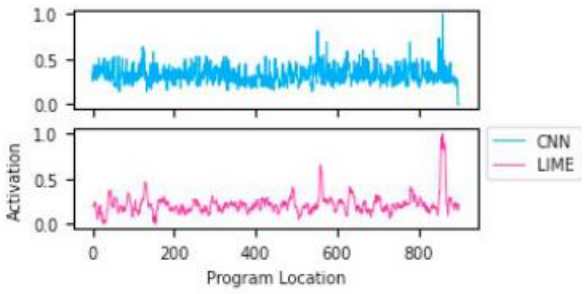

Figure 2. Activation for the CgFinder malware.

Therefore, by using LIME, they were able to conclude that their model is learning correctly for malware detection. Of course, all these methods have their limitations by themselves. Even with the positive results achieved by Kinkead, Millar, McLaughlin, & O'Kane, LIME has a limitation in which it would not be able to explain models that do not have linear decision boundaries. This makes LIME not be ideal for the more high complex, deep learning algorithms. Furthermore, it has problems explaining surrounding oberservations. In the scope of cybersecurity, if the method is unable to provide context for detected anomalies or threats, it may limit its effectiveness in identifying sophisticated attacks or understanding the full scope of a security incident (Ali, et al., 2023). However, hybrid frameworks are being designed by combining different methodologies. Some of these XAI tool kits and libraries include iNNvestigate, InterprateML, Captum, etc. (Hariharan, Velicheti, Anagha, Thomas, & Balakrishnan, 2021). By using tool kits and libraries and creating new ones, applying XAI will be easier and more widespread. With more research and development, there will be further advancements in XAI methods. This can allow it to overcome current time complexity constraints, make its explanations of an AI model's decisions more user-friendly, and enhance its applicability in cybersecurity.

STS Discussion

Although implementing XAI can be effective in analysis, it is not the solution to everything. Just because the AI can now present how it reached an output, does not mean analysts should put their full trust in the system. The data that the AI is trained on has the potential to still be biased. With many use cases, it will need a large data set to train on. When managing such large amounts of data, the process of training the model increasingly gets slower with larger amounts of data which is why XAI algorithms need to have good optimization

(Srivastava, et al., 2022). A huge concern in the cybersecurity field is gathering the data used to train for XAI. It raises some privacy concerns as it often involves data about user's behavior and network traffic logs. There exists a tradeoff between accuracy and interpretability in which the more complex a model is, the harder it is to interpret (Ali, et al., 2023). It would require more research and dedication to be able to develop XAI for a complex cybersecurity system. Overall, it should be used as a tool for suggestion and not as a final say. Humans still play a crucial role in cybersecurity, and XAI cannot detect every vulnerability. One popular cyber vulnerability is a data breach, but they do not only occur through cyber-attacks. In fact, according to Jones & Towse around 88% of data breaches in the UK occur due to human error of the employees working at the company. There needs to be a stronger emphasis on cybersecurity awareness and better training for personnel. Once that is in place, organizations can leverage XAI as a complementary tool to improve their cybersecurity and mitigate risks associated with cyber-attacks and human error.

Conclusion

The research into XAI aimed to learn more about how its application can benefit the field of cybersecurity. As technology continues to advance so does the threat landscape. There is a requirement for innovative approaches to defense mechanisms. By providing insights into the decision-making process of AI models, XAI enables cybersecurity analysts to make informed decisions to address potential biases or discrepancies. The examination of existing tests with XAI frameworks and toolkits has shown promising results for its further implementation in the cybersecurity field. It is important to explore what is currently existing in the field of XAI to see if can be utilized for such a crucial field. Based on the findings, cybersecurity businesses should be implementing XAI into their AI tools. Furthermore, researchers need to conduct more

research on the application of XAI with other AI cybersecurity tools to ensure it can have a wider range. XAI needs to be sufficient to be able to manage itself in risk-heavy situations. There needs to be an agreement on what is the best approach to present the explanations for users whether it be textually or visually. The next step for AI in cybersecurity lies in the integration of XAI frameworks along with research to broaden its effectiveness.

References

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., . . . Herrera, F. (2023, November). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*.

Capuano, N., Fenza, G., Loia, V., & Claudio, S. (2022, September 12). Explainable Artificial Intelligence in CyberSecurity: A Survey. *IEEE Access*.

CISA. (2024). *America's Cyber Defense Agency: Artificial Intelligence*. Retrieved from Cybersecurity & Infastructure Security Agency: https://www.cisa.gov/ai

Google. (2024). *Explorables*. Retrieved from Pair: https://pair.withgoogle.com/explorables/uncertainty-calibration/#:~:text=In%20general%2C%20machine%20learning%20classifiers,the%20input%20data%20belongs%20to

Grigoryan, G. (2022). Explainable Artificial Intelligence: Requirements for Explainability. *Special Interest Group on Simulation and Modeling (SIGSIM) Principles of Advanced Discrete Simulation (PADS)* (pp. 27-28). Atlana: ACM.

Hariharan, S., Velicheti, A., Anagha, A. S., Thomas, C., & Balakrishnan, N. (2021). Explainable Artificial Intelligence in Cybersecurity: A Brief Review. *4th International Conference on Security and Privacy (ISEA-ISAP).* Dhanbad: IEEE.

IBM Data and AI Team. (2023, July 6). *AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the difference?* Retrieved from IBM: https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/

Jones, H., & Towse, J. (2019). *Why Do People Make Mistakes? Changing the Narrative Around Human Errors Over Email.* University of Central Lancashire.

Kinkead, M., Millar, S., McLaughlin, N., & O'Kane, P. (2022). Towards Explainable CNNs for Android Malware Detection. *Procedia Computer Science*.

La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., . . . Angelini, M. (2023, February 1). State of the Art of Visual Analytics for eXplainable Deep Learning. *COMPUTER GRAPHICS forum*.

Mi, J.-X., Jiang, X., Luo, L., & Gao, Y. (2024, January 1). Toward explainable artificial intelligence: A survey and overview on their intrinsic properties. *Neurocomputing*.

Moradi, M., & Samwald, M. (2021, March 1). Post-hoc explanation of black-box classifiers using confident itemsets. *Expert Systems with Applications*.

Petronaci, M., & Driggs, T. (2023, December 20). *Five Questions Security Teams Need to Ask to Use Generative AI Responsibly*. Retrieved from CloudStrike: https://www.crowdstrike.com/blog/questions-to-ask-for-responsible-generative-ai-use/

Sattarzadeh, S., Sudhakar, M., & Plataniotis, K. N. (2021). SVEA: A Small-scale Benchmark for Validating the Usability of Post-hoc Explainable AI Solutions in Image and Signal Recognition. *International Conference on Computer Vision Workshops.* Montreal: IEEE/CVF.

Srivastava, G., Jhaveri, R. H., Bhattacharya, S., Pandya, S., Rajeswari, Reddy Maddikunta, P., . . . Reddy Gadekallu, T. (2022, June). XAI for Cybersecurity: State of the Art, Challenges, Open Issues and Future Directions. *ACM Computing Surveys*.