Global Scientific JOURNALS

GSJ: Volume 12, Issue 1, January 2024, Online: ISSN 2320-9186
www.globalscientificjournal.com

ENHANCING EXPLAIN ABILITY IN AI-POWERED CYBER ATTACK DETECTION SYSTEMS

Mohammed Fareed Mahdi

Department of Computer Science, University Of Thi-Qar , 64001 – Iraq
mfmsprof@utq.edu.iq

KeyWords
Artificial Intelligence, Cybersecurity, Cyber-attacks, Deep Learning, NLP, Machine Learning.

ABSTRACT

In the modern world of Cybersecurity, the inevitable role of AI-powered systems seems glaringly obvious, an absolute necessity amid the escalating complexity of cyber threats. However, there is a formidable obstacle to widespread trust – the inherent ambiguity of these systems that underscores the urgent need for transparent AI systems, and explains the limitations of black box technologies. The research presents explanatory methods for machine learning, from rule extraction to feature importance analysis and model introspection. By proposing rule-based systems for interpretable decision making, the study works to enhance transparency in understanding cyber threats. She calls for continued progress in artificial intelligence, and advocates hybrid models to raise the level of transparency without compromising detection accuracy. It emphasizes how hybrid models, which combine rule-based systems and AI algorithms, this study represents a vital guide that enables Cybersecurity experts to deal with dynamic digital threats with agility and flexibility.

I. Introduction

A. An overview

In the contemporary global context, the Cybersecurity threat landscape is in a state of continuous transformation, with malicious actors deploying progressively sophisticated methodologies to infiltrate sensitive data and disrupt critical infrastructures. To confront this evolving menace effectively, there has been a notable surge of interest in the application of artificial intelligence (AI) within the realm of cyber-attack detection. AI-driven systems harness advanced algorithms and machine learning models to scrutinize extensive datasets and discern patterns indicative of malicious behavior. These systems have exhibited promising capabilities in the identification and mitigation of cyber threats, offering marked enhancements in speed, precision, and scalability when compared to traditional rule-based approaches. Nevertheless, as AI algorithms become increasingly intricate and opaque, a pressing concern arises concerning the absence of explicability within AI-powered cyber-attack detection systems. The inherent black-box nature of these systems, where decisions are rooted in intricate neural networks or other complex models, engenders significant challenges for security analysts and stakeholders striving to comprehend the rationale behind specific decisions or predictions.

This dearth of explicability undermines trust, transparency, and accountability, thereby impeding the widespread adoption and acceptance of AI in the domain of Cybersecurity.

B. Problem statement

Navigating the intricate terrain of Cybersecurity, the infusion of artificial intelligence (AI) into cyber-attack detection systems emerges as a formidable strategy, wielding potency in the identification and mitigation of cyber threats. Yet, a looming quandary encapsulates this cybernetic venture—a substantial hurdle manifests in the elusive realm of explainability within these AI-powered systems. The adoption of labyrinthine machine learning models, often veiled in the shroud of opaque black boxes, unfolds a panorama of challenges encompassing transparency, trust, and accountability. This enigmatic opacity casts shadows on the understanding and rationale behind decisions orchestrated by these systems, stymying security analysts and stakeholders alike in deciphering flagged events, scrutinizing biases, validating the orchestration of model performance, and seamlessly collaborating with the intricate machinery of AI systems. The imperative issue beckons with exigency for a prompt spotlight, urging an imperative focus on amplifying the elucidation embedded within AI-powered cyber-attack detection systems. A quest ensues to cultivate justifications imbued with a resonance comprehensible to the human intellect. Tackling this enigma unfurls the gateway to constructing trust, ensuring unwavering adherence to regulatory mandates, streamlining the orchestration of incident responses, seamlessly infusing human prowess into the realms of threat hunting, deciphering and rectifying entrenched biases, scrutinizing the fortitude of model robustness, and cultivating a crescendo of user acceptance. In the tapestry of such embellishments, the crescendo reverberates—a symphony that transcends to elevate the collective efficacy and dependability of cyber defense operations. In this intricate dance, the ballet of collaboration unfolds between the orchestrated

prowess of AI and the nuanced interplay of human involvement in the labyrinth of Cybersecurity.

C. Related Studies

a. AI-Powered cyber-attack detection systems

AI-powered cyber-attack detection systems represent a paradigm shift in Cybersecurity, enabling automated analysis and decision-making processes that augment the capabilities of human analysts. These systems leverage machine learning algorithms, including deep neural networks, support vector machines, and random forests, to learn from historical data and identify patterns indicative of cyber-attacks. By analyzing diverse data sources such as network traffic logs, system logs, user behavior, and threat intelligence feeds, these systems can detect both known and unknown threats, helping organizations respond swiftly and effectively [1].

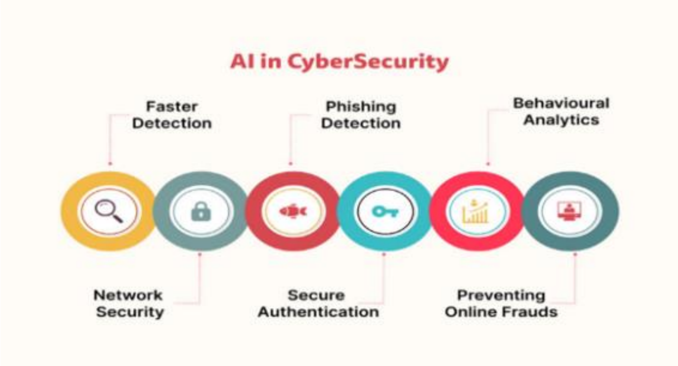

Figure 1Ai in Cyber-Security [1]

In the dynamic expanse of contemporary digital ecosystems, the prowess of AI-powered systems emerges, demonstrating unparalleled adeptness in navigating the colossal torrents of data. This supremacy becomes particularly evident when juxtaposed with the conventional rule-based systems grappling to adapt to the incessantly shifting landscape of emerging threats. Yet, within the labyrinth of their sophistication, a paradox unfurls – the very complexity that propels their excellence becomes a crucible of challenges when it comes to elucidating their inner workings. These systems, reliant on enigmatic machine learning models, evoke a sense of opaqueness. While their efficacy in achieving exalted levels of accuracy is indisputable, the convoluted realm of decision-making introduces a tapestry of concerns. A cryptic dance unfolds, where intricacies of bias, potential discrimination, and the quest to unveil vulnerabilities cast shadows on the transparency and comprehension of their operational nuances. The narrative of these AI-powered systems, adorned with the laurels of accuracy, intricacy, and performance, interweaves a tapestry that tantalizingly teeters on the precipice between technological prowess and the enigmatic uncertainties that permeate the complex landscape they navigate.

b. IMPORTANCE OF EXPLAINABILITY IN CYBERSECURITY

In the intricate domain of Cybersecurity, the pivotal essence of explainability unfurls, intricately weaving threads of trust, accountability, and a harmonious dance of collaboration between the discerning human analysts and the enigmatic realm of artificial intelligence. Within the labyrinth of cyber-attack detection, it metamorphoses into a tapestry, demanding the provision of intricate, justifiable explanations for the cryptic decisions orchestrated by AI, acting as a beacon guiding analysts through the intricate landscape of flagged events. A symphony of necessity arises from the legal and regulatory realms, where the demand for transparency propels explainability to the forefront, an indispensable cog ensuring the orchestration of compliance and unwavering adherence to the regulatory symphony. In the crucible of practicality, its significance burgeons, an indispensable compass guiding the hands of analysts in the realms of incident response and the intricate pursuit of threat hunting. Moreover, it assumes the role of a catalyst, a conduit fostering the transfusion of knowledge and the fusion of human acumen with the machinations of AI, where the interpretative prowess of human expertise refines the contours of AI capabilities, giving rise to a synergistic crescendo that elevates the very essence of cyber defense operations to unprecedented heights.

C. XAI:

In the intricate realm of Explainable Artificial Intelligence (XAI), a relentless pursuit unfolds—endeavoring to unveil the clandestine intricacies of AI decision-making, peeling back the layers of opacity enshrouding deep learning models. Within the seminal work by Kuppa and Le-Khac (2020), the Cybersecurity community embarks on an ever-intensifying voyage, intricately weaving Machine Learning (ML) into the tapestry of defenses against the ceaseless evolution of threats. The efficacy of ML models, a linchpin in Cybersecurity, hinges precariously on the delicate balance of user comprehension and trust, where nuanced insights carve pathways to resilience. Recent literature, a dynamic symphony of thought, converges on the augmentation of explainability methods, the probing depths of white-box interpreter attacks, and the meticulous delineation of properties shaping model explanations. Kuppa and Le-Khac, architects of knowledge, present a taxonomy for XAI in Cybersecurity—an intellectual arsenal delving into security properties and the labyrinth of threat models. Their orchestration culminates in the unveiling of a groundbreaking black box attack, a mesmerizing dance of deception that masterfully misdirects explanations without perturbing the stoic outputs of classifiers. This study, an opulent tapestry of insights, meticulously weaves threads of understanding through the complex terrain of security aspects in XAI within the crucible of Cybersecurity contexts [2]

d. CHALLENGES OF EXISTING BLACK-BOX APPROACHES

In the intricate expanse of AI-fueled cyber-attack detection, the pervasive utilization of convoluted black-box models, exemplified by the enigmatic deep neural networks, ushers forth a tapestry of formidable challenges due to their veiled complexity. Despite their commendable precision in unraveling intricate patterns, the opaqueness woven into their fabric obstructs comprehension, erecting hurdles in the labyrinthine task of identifying and rectifying biases nestled within the decision-making odyssey. This opaqueness casts a pall over the evaluation of models, the fortitude assessment against adversarial onslaughts, and the cultivation of trust among a discerning user cohort [3].

In realms as critical as Cybersecurity, where the tenets of fairness hold sway, the incapacity to elucidate biases metamorphoses into a cascading conundrum with ramifications stretching across the horizon. The infusion of black-box models into pivotal decision-making arenas spanning diverse sectors stirs a cauldron of apprehensions, birthing a clarion call for an epochal transition towards models inherently susceptible to interpretation. The orchestration of solutions involves an alchemy wherein the alabaster transparency of systems is burnished, and the arcane decisions orchestrated by AI algorithms are cast into the limelight, nurturing the gestation of formidable, reliable, and efficacious AI-powered cyber-attack detection systems. [1]

e. CHALLENGES AND LIMITATIONS OF AI-BASED THREAT DETECTION:

In the labyrinthine expanse of their 2020 exploration, A. Kim and B. A. Anderson embark on an odyssey into the convolution of attention processes, unraveling the intricate dance between the enigmatic bottom-up and top-down mechanisms that mold the contours of our focus. Illuminating the profound echoes of prior learning, they accentuate the automatic abduction of attention by stimuli intertwined with the siren calls of rewards or the ominous specter of threats. Delving into the kaleidoscopic interplay of value-driven and threat-wrought attentional capture, they unfurl the threat of electric shock as a chiaroscuro canvas. The revelatory symphony of results casts a surprising chiaroscuro, revealing the threat's hand in diminishing the siren song of value-driven attentional capture, a stark chiaroscuro against the backdrop of established paradigms. This newfound symphony, a cacophony of competitive dynamism between value and threat processing, emerges as a salient chiaroscuro in the landscape of cognitive revelations.

Pivoting towards the cybernetic frontier, the study casts a kaleidoscopic gaze, accentuating the paramount importance of surmounting challenges in the realm of AI-driven cyber-attack detection. The clarion call for enhancement, a staccato rhythm in the symphony of transparency and reliability, resonates as an empowering crescendo. Analysts, poised at the vanguard, empowered with heightened confidence and precision, navigate the cybernetic maelstrom, identifying and countering the elusive dance of cyber threats with a resounding symphony of virtuosity [4]

II. Related Studies

In the labyrinthine expanse of their 2020 exploration, A. Kim and B. A. Anderson embark on an odyssey into the convolution of attention processes, unraveling the intricate dance between the enigmatic bottom-up and top-down mechanisms that mold the contours of our focus. Illuminating the profound echoes of prior learning, they accentuate the automatic abduction of attention by stimuli intertwined with the siren calls of rewards or the ominous specter of threats. Delving into the kaleidoscopic interplay of value-driven and threat-wrought attentional capture, they unfurl the threat of electric shock as a chiaroscuro canvas. The revelatory symphony of results casts a surprising chiaroscuro, revealing the threat's hand in diminishing the siren song of value-driven attentional capture, a stark chiaroscuro against the backdrop of established paradigms. This newfound symphony, a cacophony of competitive dynamism between value and threat processing, emerges as a salient chiaroscuro in the landscape of cognitive revelations. Pivoting towards the cybernetic frontier, the study casts a kaleidoscopic gaze, accentuating the paramount importance of surmounting challenges in the realm of AI-driven cyber-attack detection. The clarion call for enhancement, a staccato rhythm in the symphony of transparency and reliability, resonates as an empowering crescendo. Analysts, poised at the vanguard, empowered with heightened confidence and precision, navigate the cybernetic maelstrom, identifying and countering the elusive dance of cyber threats with a resounding symphony of virtuosity [4].

In the sprawling cyber expanse, I. Radu's 2023 revelation on "Advantages of AI in Cybersecurity: Threat Detection and Response" intricately navigates the dichotomy of artificial intelligence (AI). It heralds AI as a vigilant custodian, adept at real-time threat identification, yet shadows loom—false positives, alert fatigue, and a looming skills gap. The price tag is hefty, particularly for smaller entities, and the twilight unveils hackers wielding AI's double-edged sword. Radu beckons a meticulous equilibrium, urging a layered defense with AI, technology, and human acumen [5].

Meanwhile, the 2022 symphony by A. Yayla, L. Haghnegahdar, and E. Dincelli orchestrates the evolving AI-driven Cybersecurity paradigm. Their magnum opus scrutinizes the metamorphosis toward opaque, black-box AI systems. Within smart grid intrusion detection's labyrinth, risk management falters without transparency. Priorities pivot—explainability over algorithms, birthing a risk assessment ballet, spotlighting the profound ramifications [6].

J. Li, Z. Zhao, R. Li, and H. Zhang's 2018 odyssey navigates Software-Defined Internet of Things (SD-IoT). Centralized prowess waltzes with collaboration, yet the surge births security wraiths. Traditional sentinels falter, unveiling a two-stage AI symphony. Bat algorithms and Random Forest crescendo, outshining rivals, unveiling a serenade of intelligent attack detection for fortified SD-IoT security [7].

In 2019, P. Singh and cohorts unravel the saga of neural networks defending besieged networks. Intruders breach conventional defenses, prompting a neural epos, surging detection and alarm acumen. The crescendo—a neural network ballet against denial-of-service attacks, a sentinel poised to augment network security [8].

W. Samek et al.'s 2017 saga unravels AI's zenith, veiled in opacity. Their clarion call echoes—transparent solutions for labyrinthine AI black boxes.

Methodologies surface, deciphering sensitivity and untangling AI's decision-making—a lexicon for critical bases like healthcare [9].

N. Shone et al.'s 2018 epic unearths Network Intrusion Detection Systems' (NIDSs) sagacity. Amidst dwindling accuracy, NDAE, a deep learning sonnet, unfurls. Stacked NDAEs weave a tapestry, promising to intertwine with contemporary NIDSs [10].

Finally, M. Wang et al.'s 2020 sonnet heralds machine learning-based Intrusion Detection Systems' (IDSs) prowess. Neural crescendos amplify detection, yet opacity shrouds the melody. Enter SHAP, a symphony of explanations, elucidating IDS ballets. The opus concludes—a SHAP milestone in the pantheon of IDS revelations [11]

Conclusion

In a tapestry of cognitive revelations, A. Kim and B. A. Anderson's 2020 exploration unravels the complex dance of attention processes, spotlighting the surprising symphony of competitive dynamism in value-driven and threat-oriented attentional capture. Transitioning to the cybernetic realm, their clarion call resonates for enhanced AI-driven cyber-attack detection, a staccato rhythm of transparency and reliability empowering analysts amid the cybernetic maelstrom. I. Radu's 2023 revelation navigates the nuanced dichotomy of AI in Cybersecurity. While AI stands as a vigilant sentinel in real-time threat identification, shadows of false positives and alert fatigue loom large. Radu advocates for a meticulous equilibrium, endorsing a layered defense strategy with AI, technology, and human acumen.

The 2022 study by A. Yayla, L. Haghnegahdar, and E. Dincelli orchestrates the evolution of AI-driven Cybersecurity, emphasizing the profound ramifications of prioritizing explainability over algorithms. This symphony, resonating across studies, heralds the intricate ballet of technological advancements, transparency, and the ever-evolving landscape of Cybersecurity.

a. Explainable machine learning:

The integration of machine learning (ML) in Cybersecurity, particularly in areas like malware detection and intrusion detection, brings forth significant advantages but hinges on the crucial aspect of explainability. The need for explanations in Cybersecurity ML models is paramount, as users seek more than binary outputs for analysis. Explainable ML models play a pivotal role in addressing the "trust" problem, offering insights into model behaviors, identifying misclassifications, and even automatically rectifying faults. ML techniques empower security applications to autonomously learn based on prior algorithms, eliminating the need for explicit programming. The discourse on explainability in Cybersecurity ML delineates between ante hoc and post hoc explanations.

As seen on the following figures, Ante hoc explanations involve embedding explanatory modules into the design, creating comprehensible and inherently interpretable models. On the other hand, post hoc explanations focus on interpretable techniques applied to predeveloped models, employing workflows with interconnected prediction and explanation modules. This distinction highlights the nuanced landscape of understandable machine learning in the context of Cybersecurity [12].

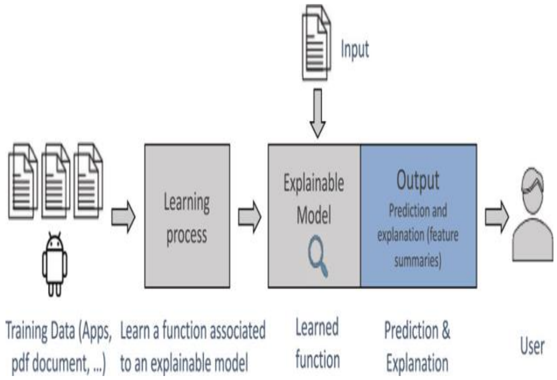
Input
Learning process
Explainable Model
Output
Prediction and explanation (feature summaries)
Training Data (Apps, pdf document, ...)
Learn a function associated to an explainable model
Learned function
Prediction & Explanation
User

Figure 3: Ante hoc explanation workflow in Cybersecurity [12]

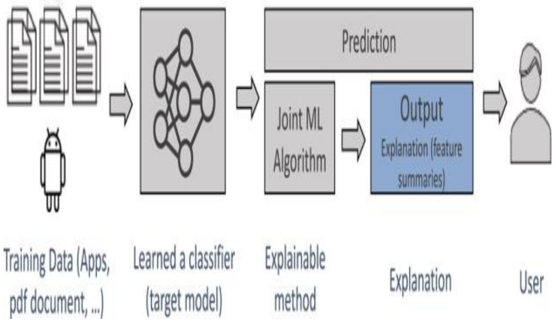
Prediction
Joint ML Algorithm
Output
Explanation (feature summaries)
Training Data (Apps, pdf document, ...)
Learned a classifier (target model)
Explainable method
Explanation
User

Figure 2: Post hoc explanation workflow in Cybersecurity [12]

b. General techniques for AI-based threat detection:

In the ever-evolving Cybersecurity arena, the infusion of Artificial Intelligence (AI) has unveiled transformative techniques for threat detection, Let's delve into each technique:

A. Deep Learning (DL)

With “Deep Learning” (DL) emerging as a formidable paradigm , DL a subfield of machine learning, employs intricate artificial neural networks to autonomously extract hierarchical representations from raw data. This shift, depicted in Figure (3), showcases DL's prowess in mimicking the intricate workings of the human brain.
DL excels in navigating high-dimensional and complex data, from network traffic logs to malware samples, surpassing traditional rule-based systems by autonomously adapting to evolving attack patterns. Its groundbreaking feature lies in eliminating the need for extensive manual feature engineering, providing a leap in efficient threat detection.

However, the enigma of DL lies in its black-box complexity, hindering transparency and interpretability. Understanding the rationale behind model decisions is imperative in Cybersecurity. Techniques like layer-wise relevance propagation and attention mechanisms are pivotal, offering insights into decision-making processes.
Despite its prowess, DL grapples with challenges. The hunger for diverse datasets clashes with privacy concerns, while adversarial attacks demand robust defense mechanisms.

DL stands as a luminary in Cybersecurity, unraveling intricate threats by navigating data complexities. Yet, the journey forward mandates addressing challenges, steering towards transparency, interpretability, and defenses against adversarial shadows for the unwavering integration of DL in the Cybersecurity lexicon [13].

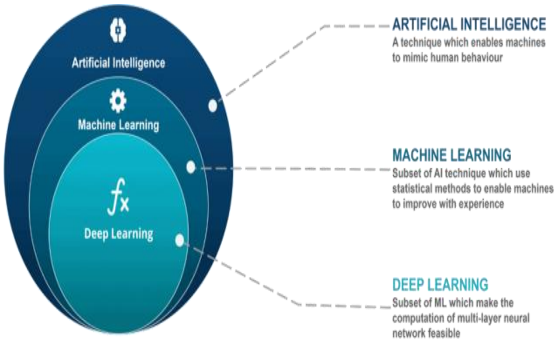
Artificial Intelligence
Machine Learning
f_x
Deep Learning
ARTIFICIAL INTELLIGENCE
A technique which enables machines to mimic human behaviour
MACHINE LEARNING
Subset of AI technique which use statistical methods to enable machines to improve with experience
DEEP LEARNING
Subset of ML which make the computation of multi-layer neural network feasible

Figure 4Deep Learning Vs. Machine Learning

B. Natural language processing (NLP)

Natural Language Processing (NLP), a pivotal subset of Artificial Intelligence (AI), emerges as a linchpin in Cybersecurity, empowering computers to comprehend, interpret, and generate human language. The synergy between NLP and AI-driven cyber-attack detection systems is undeniable. NLP's prowess lies in dissecting unstructured textual data, from incident reports to threat feeds, unraveling patterns and insights vital for Cybersecurity professionals. This automatic comprehension aids in identifying and mitigating potential cyber threats, offering an efficient countermeasure paradigm.

NLP extends its utility in text classification and sentiment analysis within Cybersecurity, enabling the categorization of incidents and the identification of malicious content or sentiment in textual data. The integration of NLP is also instrumental in bolstering the explainability of AI-powered systems. Operating in a realm often dominated by opaque models, NLP techniques like text summarization and explanation generation provide human-understandable insights into complex decision-making processes, enhancing transparency and trust.

Notably, NLP becomes a sentinel against phishing emails and social engineering attacks, analyzing content, language patterns, and context to flag suspicious elements. Beyond threat detection, NLP's conversational aptitude contributes to Cybersecurity operations through intelligent chatbots and virtual assistants. These agents engage users, offering guidance and information, thereby enhancing accessibility and user experience.

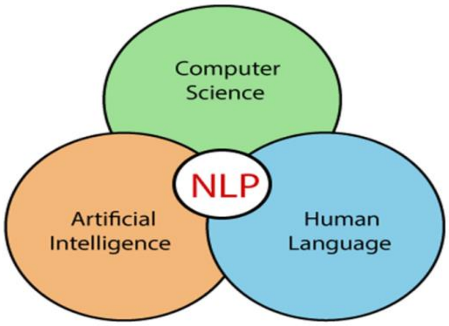

Figure 5: A pictorial representation of NLP

However, NLP in Cybersecurity confronts challenges, particularly in the face of evolving cyber attacker techniques, including obfuscation and adversarial attacks aiming to deceive NLP models. The ongoing quest for robustness and resilience against such adversities underscores the dynamic intersection of NLP and Cybersecurity. In summary, NLP stands as an indispensable tool, unraveling the linguistic intricacies within Cybersecurity, bolstering defenses, and improving the overall efficacy of cyber defense operations [14].

C. Interpretable machine learning models for cyber-attack detection:

In the realm of Cybersecurity, the detection and mitigation of cyber-attacks stand as paramount endeavors to safeguard sensitive information and uphold the integrity of computer systems. Machine learning models have exhibited significant promise in discerning and identifying cyber threats through pattern and anomaly analysis. However, the interpretability of these models remains a pivotal concern, leading to the prominence of interpretable machine learning models in cyber-attack detection systems. Interpretable models, designed to offer human-understandable explanations for their decision-making processes, prove instrumental in providing transparency and clarity, crucial for enhancing the trustworthiness of cyber-attack detection systems.

Decision trees, rule-based models, and linear models emerge as notable interpretable approaches, each offering distinct advantages in terms of human-readability and understanding.

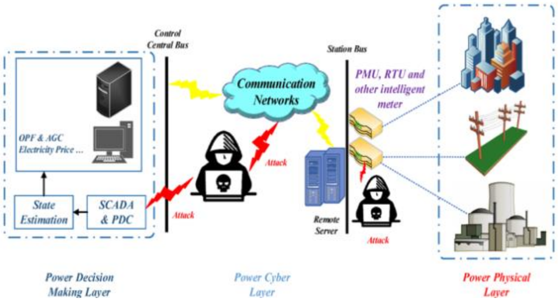

Figure 6: The power system framework configuration [16]

Despite the benefits, interpretable machine learning models face challenges, including the delicate balance between interpretability and performance. The trade-off may involve sacrificing some degree of accuracy or complexity compared to more advanced black-box models. Ongoing research endeavors focus on hybrid models, aiming to combine the interpretability of simpler models with the performance of complex ones.

In a pertinent 2018 study by H. S. Lallie, K. Debattista, and J. Bal, the authors address the inherent difficulty in perceiving and comprehending cyber-attacks. They emphasize the need for effective techniques to enhance the understanding of cyber-attacks and propose an empirical evaluation comparing adapted attack graph methods with standard fault tree methods. The findings indicate the adapted attack graph method's superiority in aiding cyber-attack perception. The study underscores the significance of further research and presents recommendations for future exploration [15].

In conclusion, interpretable machine learning models emerge as crucial assets in cyber-attack detection, providing transparency, accountability, and trustworthiness. While challenges exist, ongoing advancements in the field have the potential to significantly improve cyber-attack detection and enhance overall organizational security postures [15] [16].

D. LEVERAGING RULE-BASED SYSTEMS FOR EXPLAINABILITY IN CYBERSECURITY:

Explainability holds critical importance in Cybersecurity, and rule-based systems emerge as a valuable avenue for achieving transparency and interpretability in decision-making processes. In contrast to complex black-box models, rule-based systems offer simplicity and clarity through predefined "if-then" statements, allowing Cybersecurity analysts to comprehend the decision rationale effectively. Their simplicity enhances trust and accountability, fostering confidence among analysts who can trace the specific conditions leading to cyber-attack detection.

A 2020 study by J. Van Der Waa et al. delves into Explainable Artificial Intelligence (XAI), evaluating two explanation styles: rule-based and example-based. The study underscores the significance of user evaluations in assessing the effectiveness of these styles in decision support for diabetes self-management. Results reveal that rule-based explanations slightly improve system understanding but neither style enhances task performance. This highlights the importance of evaluating user experiences in refining assumptions about effective explanations in XAI systems [17].

Rule-based systems boast benefits such as transparency, ease of incorporation of domain knowledge, and interpretability to non-technical stakeholders. However, challenges include rule complexity, maintenance, and potential limitations in identifying subtle attack patterns. Despite challenges, leveraging rule-based systems offers significant advantages in achieving explainability in Cybersecurity, aligning with domain expertise, and facilitating communication with stakeholders [17].

E. VISUALIZING AI DECISIONS FOR ENHANCED EXPLAINABILITY IN CYBERSECURITY:

Explainability is crucial in AI-driven Cybersecurity, and visualizations offer a powerful tool to unravel complex decision-making processes.

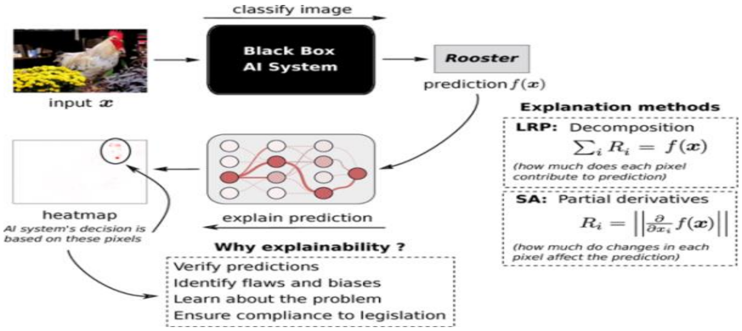
classify image
input x
Black Box
AI System
Rooster
prediction $f(x)$
Explanation methods
LRP: Decomposition
$\sum_i R_i = f(x)$
(how much does each pixel contribute to prediction)
SA: Partial derivatives
$R_i = \left\|\frac{\partial}{\partial x_i} f(x)\right\|$
(how much do changes in each pixel affect the prediction)
heatmap
AI system's decision is based on these pixels
explain prediction
Why explainability ?
Verify predictions
Identify flaws and biases
Learn about the problem
Ensure compliance to legislation

Figure 7: Explaining the important of visualizing and predictions of an AI system [18]

Transforming abstract data into graphical representations enhances analysts' comprehension and interpretation.

Feature importance visualizations aid in identifying critical factors influencing cyber-attack detection, while decision boundaries reveal how the model separates classes. Temporal visualizations track decision evolution over time, enabling the identification of emerging threats. Interactivity empowers analysts to explore AI models collaboratively [18].

A study by M. Kuzlu highlights Explainable AI (XAI) in smart grid applications. XAI tools like LIME, SHAP, and ELI5 enhance solar PV energy forecasting transparency, providing insights into AI model internal mechanisms. Visualizing AI decisions in Cybersecurity, despite challenges, enhances transparency, facilitates communication, and supports effective decision-making [19].

F. THE FUTURE OF AI IN CYBER SECURITY:

In the future, AI plays a pivotal role in Cybersecurity, with its advantages outweighing drawbacks. S. B. Atiku's study underscores AI's potential to protect against cyber threats efficiently. Linear regression analysis predicts AI's continued growth in Cybersecurity, emphasizing the importance of data quality and model selection for accurate forecasting. As our digital footprint expands, AI's role becomes indispensable in safeguarding sensitive information from evolving cyber threats [20].

To forecast future trends in AI adoption for Cybersecurity, follow these specific steps:

- Data Collection: Gather historical data detailing AI adoption levels over different years.
- Data Preparation: Use Python libraries like NumPy and pandas to format the data for analysis, ensuring it meets the model's requirements.
- Regression Model Selection: Choose an appropriate regression model, such as linear regression, based on the data and problem at hand.
- Model Training: Fit the selected regression model to the historical data using the model's fit function.
- Model Evaluation: Assess the model's performance by employing evaluation metrics like mean squared error (MSE) or R-squared, gauging its fit to historical data.
- Future Predictions: Utilize the trained and evaluated model to predict AI adoption levels for specified future years (e.g., 2024 to 2030).
- Visualization: Enhance understanding of trends by visualizing predicted results through tools like Matplotlib, creating plots and graphs that illustrate historical data, the regression line, and predicted values.
- After applying the code in python and by using real data the predicted AI adoption for cyberattacks detection and Cybersecurity from 2024 to 2030 is as follows:

Table 1 AI Adoption Prediction

Year	AI Adoption
2024	**60%**
2025	**65%**
2026	**70%**
2027	**80%**
2028	**85%**
2029	**85%**
2030	**90%**

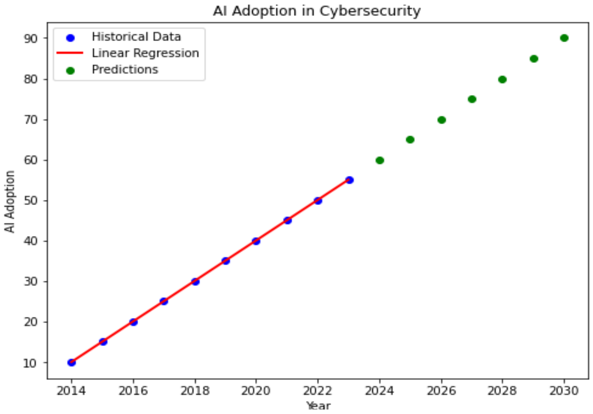

Figure 8:AI Adoption in Cyber Security.

These values indicate the projected level of AI adoption for each corresponding year. For example, in 2024, the predicted AI adoption is 60.0, and it gradually increases to 90.0 by the year 2030.

Additionally, the code generates a graph to visualize the data. It includes historical data points (represented as blue dots), a linear regression line (represented as a red line), and predicted AI adoption values for future years (represented as green dots on the line). The graph provides a visual representation of the predicted AI adoption trend from 2014 to 2030.

iii. Recommendations

Future Recommendations for AI-powered Cyber-Attack Detection Systems:

Interpretable Deep Learning Models:

- Investigate methods to enhance the interpretability of deep learning models.
- Develop novel architectures or leverage existing techniques like attention mechanisms.

Contextual Explanation Generation:

- Explore techniques for generating contextual explanations in cyber-attack detection.
- Consider specific threat contexts, attack patterns, and potential security impacts.

Hybrid Approaches:

- Investigate the integration of rule-based systems with machine learning models.

- Extract interpretable rules from black-box models for improved explainability.

User-Centric Explainability:

- Focus on user-friendly explainability interfaces catering to various user groups.
- Design tools for Cybersecurity professionals, administrators, and end-users with varying technical knowledge.

Evaluating Explanations:

- Develop standardized metrics and evaluation frameworks.
- Establish benchmarks for comprehensibility, fidelity, and usefulness of explanations.

These recommendations aim to advance the field, fostering transparency, and trust in Cybersecurity systems for effective defense against evolving threats.

iv. Conclusion

In the conclusive findings of "Enhancing Explainability in AI-Powered Cyber Attack Detection Systems," the research underscores the vital role of transparency and interpretability in the realm of Cybersecurity. The challenges posed by opaque AI models necessitate a strategic exploration of explainable machine learning techniques and the integration of rule-based systems to propel advancements. Future recommendations provide a targeted roadmap, emphasizing the development of interpretable deep learning models, contextual explanation generation, and hybrid approaches, promising heightened transparency without compromising detection accuracy. User-centric explainability and evaluation frameworks seek to foster collaboration and informed decision-making among diverse stakeholders.

Enhanced explainability equips cybersecurity professionals to comprehend, validate, and trust AI-generated outputs, facilitating effective identification, mitigation, and response to cyber threats. Prioritizing ongoing research and development efforts in enhancing explainability is crucial for bridging the comprehension gap between complex AI models and human understanding, ensuring a secure digital landscape that safeguards critical systems and data. In summary, the research offers critical insights and recommendations, contributing to the creation of transparent and trustworthy Cybersecurity solutions.

These empower professionals to navigate the dynamic landscape of cyber threats and the evolving realm of AI, staying one step ahead in ensuring robust “Digital security”.

References

[1] AUGMENTED STARTUPS, Computer Vision/ AI , "The role of artificial intelligence in cybersecurity," 2023. [Online]. Available: ” https://www.augmentedstartups.com/blog/the-role-of-artificial-intelligence-in-cybersecurity-how-ai-enhances-protection.

[2] A. Kuppa and N.-A. Le-Khac, Black Box Attacks on Explainable Artificial Intelligence(XAI) methods in Cyber Security, 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1-8.

[3] C. Rudin, Please stop explaining black box models for high stakes decisions, ArXiv, abs/1811.10154., 2018.

[4] A. Kim and B. A. Anderson, Threat reduces value-driven but not salience-driven attentional capture, Vols. 20, no. 5, Emotion, 2020, p. 874–889.

[5] I Radu, "The impact of AI on cybersecurity: Advantages and disadvantages," 2023.

[6] A. Yayla, L. Haghnegahdar, and E. Dincelli, Explainable artificial intelligence for smart grid intrusion detection systems, Vols. , vol. 24, no. 5, Sep. 2022: IT Professional, p. 18–24.

[7] J. Li, Z. Zhao, R. Li, and H. Zhang, AI-Based Two-Stage intrusion detection for software defined IoT networks, Vols. vol. 6, no. 2, IEEE Internet of Things Journal, 2019, p. 2093–2102.

[8] P. Singh, S. Krishnamoorthy, A. Nayyar, A. Kr. Luhach, and A. Kaur, , Soft-computing-based false alarm reduction for hierarchical data of intrusion detection system, Vols. 15, no. 10, International Journal of Distributed Sensor Networks, 2019, p. 155014771988313.

[9] W. Samek, T. Wiegand, and K.-R. Müller, “Explainable Artificial intelligence: Understanding, visualizing and interpreting deep learning models, ResearchGate, 2017.

[10] N. Shone, T. N. Ngoc, P. V. Dinh, and Q. Shi, A deep learning approach to network intrusion detection, Vols. 2, no. 1 , IEEE Transactions on Emerging Topics in Computational Intelligence, 2018, p. 41–50.

[11] M. Wang, K. Zheng, Y. Yang, and X. Wang, An explainable machine learning framework for intrusion detection systems, vol. 8 , IEEE Access, 2020, p. 73127–73141.

[12] F. Yan, S. Wen, S. Nepal, C. Paris, and Y. Xiang,, Explainable machine learning in cybersecurity: A survey, Vols. vol. 37, no. 12, International Journal of Intelligent Systems, 2022, p. 12305–12334.

[13] A. A. Bhalerao, B. R. Naiknaware, R. R. Manza, V. Bagal, and S. K. Bawiskar, "Social media mining using machine learning techniques as a survey," in Advances in computer science research, p. 874–889. , 2023.

[14] P.- Sneha, Natural Language Processing (NLP), Pianalytix: Build Real-World Tech Projects, 2020.

[15] H. S. Lallie, K. Debattista, and J. Bal, An empirical evaluation of the effectiveness of attack graphs and fault trees in Cyber-Attack perception, Vols. 13, no. 5, EEE

Transactions on Information Forensics and Security, 2018, p. 1110–1122.

[16] A. Almalaq, S. Albadran, and M. A. Mohamed, Deep Machine Learning Model-Based Cyber-Attacks detection in smart power systems, Vols. vol. 10, no. 15, Mathematics, p. 2574.

[17] J. Van Der Waa, E. Nieuwburg, A. Cremers, and M. A. Neerincx, Evaluating XAI: A comparison of rule-based and example-based explanations, vol. 291, Artificial Intelligence, 2021, p. 103404.

[18] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable AI: A brief survey on history, research areas, approaches and challenges," in Lecture Notes in Computer Science, 2019, p. 563–574.

[19] M. Kuzlu, Ü. Cali, V. Sharma, and Ö. Güler, "Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools, vol. 8 , IEEE Access, 2020, pp. 187814–187823, .

[20] S. B. Atiku, Survey On The Applications Of Artificial Intelligence In Cyber Security, 2020.