https://gospodarkainnowacje.pl

GOSPODARKA I INNOWACJE

LABORATORIUM WIEDZY Artur Borcuch

Volume: 29 | 2022

Economy and Innovation
ISSN: 2545-0573

For more information contact: editor@gospodarkainnowacje.pl

BUILDING TRUST IN AUTONOMOUS CYBER DECISION INFRASTRUCTURE THROUGH EXPLAINABLE AI

K M Zubair
Bachelor of Science in Computer Science, International Islamic University Malaysia, Kuala Lumpur, Malaysia

Akhtaruzzaman Khan
Bachelor of Science in Computer Science and Engineering, North South University, Dhaka, Bangladesh

Tanvir Rahman Akash
Bachelor's of Business Administration in Finance, Bangladesh University of Professionals (BUP), Dhaka, Bangladesh

A R T I C L E I N F O.

Keywords: Explainable Artificial Intelligence (XAI), Autonomous Cyber Decision Infrastructure (ACDI), Network Intrusion Detection, Trust in AI Systems, Interpretability of machine learning and Cyber security Transparency.

Abstract

The high rate of the development of digital technologies has changed cyber security from a manual, rule-based practice to a complex, data-driven field that needs intelligent automation. The study article, which has the title of Building Trust in Autonomous Cyber Decision Infrastructure through Explainable AI (XAI), investigates the potential of explain ability to promote transparency, reliability and trust in human operators to AI-based cyber security systems. Conventional black-box AI models are accurate, but they are frequently uninterpretable, which creates distrust in the cyber security community, which bases its opinions on understandable arguments to confirm automated warnings. This study incorporates Explainable AI (SHAP, SHapley Additive Explanations, and LIME, Local Interpretable Model-Agnostic Explanations) into the machine learning models to interpret network anomalies, protocol behavior, and intrusion patterns with the help of the Network Intrusion Detection Dataset. These findings indicate that XAI is not only more interpretable but also it narrows the divide between AI decision-making and human monitoring, and leads to operational trust and responsibility. Among the essential conclusions, it is possible to note that model transparency has a direct impact on operator trust, which allows implementing timely, effective, and auditable responses to cyber threats. Moreover, explain ability helps in determining key characteristics such as connection duration, failed logins, and the use of protocols, which contribute to the detection of anomalies, thereby enhancing the accuracy of analysts and human perception. The paper concludes that using XAI in Autonomous Cyber Decision Infrastructures (ACDI) is not only capable of making cyber security defenses predictive, but also intelligible and ethically sound. This study has added to the current development of transparent, adaptive, and intelligent cyber security systems that can effectively address the dynamic threat environment through increased trust, accountability and human-AI interactions.

http://www.gospodarkainnowacje.pl/ © 2024 LWAB.

I. Introduction

A. Background

The high rate of digital technologies development has also given rise to unprecedented development of networked systems, which causes cyber security to become a pressing issue of concern to organizations operating in various fields. The conventional methods of cyber security that have been based on a set of rules and manual surveillance are becoming insufficient to combat the increased complexity and occurrence of cyber threats [1]. Examples of cyber-attacks like network intrusions, ransom ware, and large scale data breaches have cost businesses and critical infrastructure across the globe a considerable sum of money and operations as well as reputation. The developments demonstrate the acute necessity of autonomous cyber decision infrastructure (ACDI) that uses artificial intelligence (AI) and machine learning (ML) to perceive, examine, and react to cyber threats in real-time. In contrast to the traditional systems, ACDI will be able to track network traffic at all time, detect elaborate patterns, and find anomalies that might represent malicious activity. Automating the process of detecting and responding to threats, ACDI systems make the operation of the system more efficient and less prone to human error, with a quicker and more reliable response to cyber-attacks [2]. The usefulness of such systems however does not just lie in their predictive performance but also in their incorporation into the human decision-making processes. It is vital to have knowledge of underlying mechanisms of AI-based systems so that the operators can be able to validate alerts and act accordingly. As a result, ACDI is a dramatic change in the field of cyber security with the possibility of proactive, smart, and autonomous defense solutions that can help respond to the changes in the threat environment with minimal human intrusion.

B. The significance of Trust in AI-based cyber system security

Although ACDI has the technological ability to operate automated cyber security, human trust is a major challenge to the adoption. The opaque character of the majority of AI models provides uncertainty to cyber security experts who might not be willing to accept automated suggestions without knowing the reasoning behind them. It is more important in high-stakes settings, e.g., military networks, financial systems and critical infrastructure where the wrong or slow response can be disastrous [3]. The need to make fast and informed decisions on the basis of system outputs demands the transparency and interpretability of those decisions, as they involve human operators. In addition, the effect of trust on the system effectiveness is based on the idea that the absence of trust on AI advice can translate into the operator ignoring or disregarding warnings, which can weaken security efforts. Developing trust is not only about proving the correctness and the validity of autonomous systems but also about making the reasoning of the system comprehensible and practical. Organizations can use AI speed and capacity to control their machines and stay in a trusting relationship with them without losing human control or supervision [4].. In this regard, explainable AI (XAI) methods should be incorporated in the autonomous cyber decision systems since they improve interpretability, accountability, and confidence by operators. The level of trust in AI systems is not a social or psychological aspect but a functional need to achieve successful actualization of autonomous cyber security activities.

C. Issues of the Black-box AI in Cyber security.

A majority of the AI models used in autonomous cyber decision making systems, such as deep neural networks, ensemble learning models, and gradient boosting machines, are opaque in nature. Although such models tend to be highly predictive in nature, their intricacy does not allow human operators to know how certain predictions are arrived at. This is a major problem to the adoption of AI in cyber security since practitioners cannot verify or validate the decision made by the system randomly [5]. The resultant uncertainty may cause lower operator confidence, more manual oversight, and in certain

situations, cause an operational error. Besides, black-box models that perform well can accidentally acquire biases or trends contrary to the expert knowledge, which further compromises trust. The regulation standards and ethical requirements are increasingly demanding the explain ability of AI applications, especially in the context of a system functioning in a safety-critical or dangerous field. The lack of interpretability of AI decisions is not just a theoretical issue in the context of cybersecurity, but a real obstacle to the implementation of AI-based solutions in the domain of information safety and other important infrastructures. To make autonomous systems effective and trustworthy, the organizations need to deal with these challenges [6]. It is critical that the explainable AI methodologies be integrated, however, so that human operators are able to comprehend, verify, and take action on the outputs of the models without affecting the high detection rates.

D. Role of Explainable AI (XAI)

Explainable Artificial Intelligence (XAI) has revealed itself as an innovative methodology to avert the limitations of transparency of black-box AI models. XAI represents a collection of approaches intended to render model forecasts explainable and comprehensible to human users as a solution to the complexity of AI compared to operator intelligence. There are also methods like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), which are able to provide a deeper understanding of which features contribute to a specific prediction, which enables cybersecurity professionals to know why a given network traffic is considered normal or anomalous [7]. XAI improves interpretability by making features important, decision logic of machine learning models, and thus enables easier debugging and better alignments between the predictions of AI and expert knowledge. XAI is an important tool in establishing the trust of the operators of a system of autonomous cyber decisions, since clear explanations allow professionals to check the warning, test the work of the model, and make a significant decision in response to risk. Moreover, XAI can enable accountability and regulations compliance, and it gives a history of the decision making process which can be audited or reviewed. With XAI incorporated into ACDI, organizations are able to trade off detectability with explain ability to allow confidence in AI-based cyber security operations. XAI does not only improve the technical efficiency of autonomous systems, but also improves human-machine interaction, ensuring better, safer, and more reliable and trustful cyber security infrastructures.

E. Problem Statement

Self-directed cyber decision systems have proven to be most precise in the detection of anomalies in the network and responding to threats. Their obscure and black-box properties restrain trust and confidence of human operators, without which implementation would not be effective. The system alerts may not be trusted by cybersecurity professionals, which may compromise the effectiveness and dependability of automated defenses. Moreover, regulatory and ethical issues are also requiring more transparency in the AI-led decision-making process [8]. This study is a response to the major problem of improving human trust towards autonomous cybersecurity systems through the implementation of Explainable AI methods (XAI). This research seeks to fill this gap between the ability of autonomous systems to strongly perform their predictions and those of humans so that the operators are comfortable to take the advice suggested by AI but the system remains effective.

F. Research Objectives

The study aims to:

- Design machine learning anomaly detectors in network traffic. Use the XAI methods to understand the model predictions.
- Determine anomaly detecting characteristics [9].
- Assess the human confidence on AI-made decisions.
- Determine the effect the explain ability of operators has on operator confidence.

- Make suggestions on credible autonomous cyber systems.

G. Research Questions

1. What can XAI do to enhance the confidence in autonomous cyber decision systems?
2. What characteristics have the most influence on the detection of anomalies in network traffic?
3. How is there a correlation between model explain ability and operator confidence?

II. Literature Review

A. Self-directed Cybersecurity Systems

Independent cybersecurity systems are aimed to identify, process and act upon threats with the minimum human intervention. The systems use machine learning, deep learning and rule-based engines to detect anomalies in network traffic that might signal malicious activity, based on the patterns in network traffic [10]. The main benefit of autonomous systems is that they provide the ability to work 24 hours and process huge amounts of information, which is beyond human capabilities. They are able to learn how to deal with the changing threats based on the previous occurrences thus lessening the time taken in responding and minimizing the damages that may occur. These systems have found massive application in network intrusion detection, malware detection, and advanced persistent threat detection. The main elements of this are real time monitoring engines, automated decision modules and adaptive learning mechanisms which allow the system to improve its performance as time goes by. Although they have benefits, autonomous systems have problems in operating, including false positives and negatives, which may affect their reliability. It should be connected to human operators to check the critical decisions and make the automated alerts actable. Autonomous cybersecurity systems are operationally beneficial in high-risk areas, like those involved in critical infrastructure or defense networks, but they must have mechanisms that enable them to build trust, accountability and transparency to achieve effective deployment [11]. The success of these systems, in general, lies not just in their accuracy in detecting the threats but also in the fact that they should be direct able to the human decision-making process, and adaptable and proactive defense mechanisms could be developed.

B. Human Confidence in AI-Based Cybersecurity

The confidence in AI-driven cybersecurity systems is one of the most important factors that define their successful implementation and performance. There is a tendency of human operators to be reluctant in following the recommendations of the automated systems when they lack the reasoning why the system has made the decisions [12]. The aspect of trust is especially relevant in high stakes settings, where misguided or timely failure to respond can be disastrous, such as loss of money, data breach, or computer crash. Some of the factors that affect trust are accuracy of systems, transparency, consistency and reliability of performance. Even high-performing systems lack transparency in decision-making, which can result in the operator disregarding warning signs or recommendations and losing trust in them. Trust creation means that the AI systems have to give clear explanations and prescriptive insights that make sense to human reasoning. Trust in cybersecurity has an effect on operational effectiveness and quality decision making because human intervention is frequently required in order to confirm important alerts. Those systems that do not instill confidence can be used inadequately thus limiting their effectiveness. Thus, it is necessary to create AI-based cybersecurity systems with principles made human-focused, such as transparency, accountability, and interpretability [13]. The mechanisms of trust-building allow operators to harness the analytical capabilities of AI without losing control and achieve various benefits by getting more people to cooperate and make more informed decisions and ensure safer operational results in the complex cyber environment.

C. Challenges in AI-based cybersecurity Black-box

Most of the effective AI models applied to the domain of cybersecurity, such as deep neural networks

and ensemble learning models, are opaque by nature, and the operators do not understand how the decisions regarding certain choices are made [14]. These black-box models are highly accurate and efficient yet they are very difficult to be transparent, accountable and trustworthy. The intricacy of the decision-making process means that human operators cannot substantiate or authenticate the logic behind predictions, something that is especially worrying on critical infrastructure and high-security. Black-box models are also prone to bias or pattern that is not congruent with the expert knowledge resulting in false alarms or under-detection. Moreover, legal and ethical requirements are growing to require AI use to be transparent, particularly in areas where AI judgments may have dire consequences. The lack of explain ability to model behavior can destroy the confidence of the user and impose restrictions on the practicality of autonomous cybersecurity systems. Black-box models can operationalize the interaction between human operators and AI systems because non-explanatory alerts are less actionable [15]. Such inability to be interpreted can also apply to incident investigation and post-attack analysis because decisions cannot be reconstituted. In response, overcoming black-box issues is the key to creating AI-based cybersecurity systems that are effective, as well as reliable and in accordance with the expectations of human operators.

D. Explainable Artificial Intelligence (XAI) Methods

Explainable AI (XAI) methods are focused on how to reconcile between the high-performing black box models and interpretability desired by human decision-making. XAI can give answers to how and why to make certain decisions because they give the ways of how the model predictions depend on each of the input features [16]. The most common methods are SHAP (SHapley Additive explanations) and LIME (Local Interpretable Model-agnostic Explanations), which produce explanations of particular predictions, which are on a feature-level. The approaches increase transparency, allowing cybersecurity actors to certify the output of the systems and the behavior of the models in various conditions. XAI allows taking action on insights by revealing what features drive anomaly detection and enables prioritizing network interventions or observing vital network metrics. In addition to transparency, XAI is also endorsing accountability, as explanations would allow tracing decision-making processes and meeting regulatory requirements. On cybersecurity applications, XAI could be beneficial in improving incident response by ensuring that operators are aware of true and false positives and therefore allocating resources efficiently [17]. The inclusion of XAI in autonomous cyber decision infrastructure guarantees that high detection performance is not associated with interpretability. The XAI fosters trust, minimizes errors in operations, and offers a framework to implement responsible and trustworthy AI in the intricate cyber settings by encouraging humans and AI to work together.

E. Effects of XAI on Human Trust and Decision making

Human trust and efficiency in the decision-making process directly depend on the integration of XAI in autonomous cybersecurity systems. Offering intelligible explanations of AI outputs, operators understand why some activities in a network receive the difference between categorizing it as anomalous or normal. These insights enhance trust in system warnings, which eliminates the need to verify manually and allows threats to be responded to in time. Conciseness of explanations, consistency of model behavior and alignment with operator expectations are some of the factors that affect trust in AI systems. Explicit and practical descriptions are able to diminish ambiguity, avoid dependency and enhance the decision-making process between humans and machines. Moreover, XAI also supports the learning of operators enabling them to learn new attack patterns and improve operating maneuvers [18]. In practice, the application of XAI decreases cognitive load on cybersecurity staff, enhances compliance with recommendations provided by AI, and has a beneficial effect on overall operation. AI predictions can also promote accountability because the operators can follow the logic behind a particular decision. All these together show that XAI is not only enhancing the trust in autonomous systems but also improving the engagement of the human operator and AI to result in safer, more reliable, and efficient cybersecurity activities.

F. XAI in Autonomous Cyber Decision Infrastructure Applications

Explainable AI has been used in autonomous cyber decision systems more frequently to enhance interpretability, reliability, and trust in the operator. XAI methods can be used in network intrusion detection as they indicate the features or patterns that contribute to the classification of an anomaly to support proactive monitoring and early intervention. XAI finds application in malware detection to identify key behavioral or structural patterns that can be used in performing classifications. In addition to the ability to detect threats, XAI can be used to assess risk, forensically examine, and report about incidents, as the system provides straightforward descriptions of outputs [19]. It is possible to use autonomous systems with XAI to give interactive visualizations, allowing operators to access feature importance, scenario outcomes and decision rationales. The functionality will improve human-AI collaboration since the operators are able to authenticate alerts, refine model parameters and respond to incidents more with confidence. Also, XAI applications in cybersecurity help to meet regulatory requirements and ethical use of AI, which means that automated decisions should be transparent and verifiable. With XAI implemented into the autonomous decision infrastructure, companies are able to achieve a high-performance AI with interpretability, build trust, accountability, and operational decision-making in complex and dynamic cyber settings.

G. Empirical Study

In the article by A. V. Shreyas Madhav and Amit Kumar Tyagi (2022) entitled Explainable Artificial Intelligence (XAI): Connecting Artificial Decision-Making and Human Trust in Autonomous Vehicles, the authors performed both quantitative and qualitative studies to determine how the XAI mechanisms contribute to human trust in autonomous vehicles. Their empirical research was aimed at creating the visual explanatory techniques and an intrusion detection classifier built into the vehicle decision-making system. The experiment conducted through simulation proved that the XAI models outperformed the traditional opaque AI systems in terms of interpretability and reliability [1]. The results revealed that the higher the users were given the clear rationale on the actions of the vehicles, the greater their trust and acceptance of the autonomous systems became. In addition, the study has highlighted the significance of ethical disclosure and cyber resilience because the insertion of explainability also promoted the intrusion detection in vehicular communication networks. This paper allows a solid empirical basis of the role of explainability in eliciting human trust, which can be directly applied to autonomous cyber decision infrastructures, with the transparent reasoning of algorithms capable of fostering human confidence and operational reliability.

In the article by Luca Mia titled Enhancing Trust and Accountability in Autonomous Cyber Defense Systems through Explainable AI (2025), the author gives us an empirical study of how the implementation of Explainable AI (XAI) mechanisms can increase trust and accountability in autonomous cyber defense settings. The paper examines the use of recent XAI tools like LIME (Local Interpretable Model-Agnostic Explanations), SHAP (SHapley Additive explanations) and attention mechanisms to enhance interpretability in AI-based threat detection models. The empirical findings of simulated cyber defense situations demonstrate that AI transparency decision-making leads to an increase in user trust among cybersecurity practitioners particularly in those situations when an immediate autonomous reaction is needed [2]. This paper also raises a trade-off between performance and interpretability of the model, in which real-time detection features need to be preserved and the transparency of the algorithm. Moreover, the study describes the advantages of ethical considerations and regulatory compliance obtained through implementing XAI-based systems. The results indicate that explainability does not only enhance operational trust but enhances collaboration between AI systems and human analysts thereby establishing a more accountable and robust autonomous cyber decision infrastructure, which is resonant with the principle aims of the current study.

The authors in the article by Zhibo Zhang, Hussam Al Hamadi, Ernesto Damiani, Chan Yeob Yeun, and Fatma Taher (IEEE) provide a detailed overview of current methodologies and frameworks that

incorporate the Explainable AI (XAI) aspect of cybersecurity systems. Their empirical and analytical synthesis determines the absence of interpretability of traditional machine learning (ML) and deep learning (DL) models as the reason why the human confidence in the automated system of cyber defense is not that high. The paper highlights that the lack of transparency and trust in AI models is put at a disadvantage by the black-box character of most of the models, particularly in high-stakes problems like malware classification and intrusion detection [3]. The authors measure the effect of explainability by conducting a survey and classifying XAI frameworks across various cybersecurity applications by underscoring the benefits of explainability as it allows experts to comprehend the reasoning of models and preserve a high degree of predictive accuracy. Moreover, the article combines both theoretical and practical knowledge and proves that XAI-based systems contribute to the greater transparency of operations and the accountability of decisions. This study offers a background knowledge in the creation of reliable autonomous cyber decision architectures, which supports the argument that interpretable AI is crucial to establishing a balance between security functionality and human intelligibility in the current cyber defense settings.

In the article Explainable Artificial Intelligence (XAI) to Enhance Trust Management in Intrusion Detection System using Decision Tree Model by Basim Mahbooba, Mohan Timilsina, Radhya Sahal, and Martin Serrano (2021), the authors performed an empirical study of the application of XAI techniques in enhancing trust management in Intrusion Detection Systems (IDS). A Decision Tree based model was used in the study, which is trained on the KDD benchmark data, and in this instance, the focus was made on interpretability and rule extraction rather than on a purely performance-based accuracy measure [4]. The results were that decision tree models did not only have the same competitive detection accuracy as other state-of-the-art algorithms, but also presented clear reasoning frames which enabled experts of cybersecurity to comprehend the manner in which decisions on intrusion were arrived at. This human trust and enhanced efficiency in collaborative threat response were due to this interpretability. The study overcomes a major weakness of previous IDS schemes, namely, the ability to describe the decision path by incorporating explainable mechanisms that replicate the rational in human decision-making. In general, the paper has shown that the integration of XAI into the IDS systems enhances trust, transparency, and accountability in the automated cyber defense systems, which supports the general thesis of explainable models at the heart of the creation of reliable autonomous cyber decision systems.

In the article titled Explainable Artificial Intelligence to Smart City Application: A Secure and Trusted Platform by M. Humayun Kabir, Khondokar Fida Hasan, Mohammad Kamrul Hasan, and Keyvan Ansari (2022), the authors discuss how Explainable Artificial Intelligence (XAI) helps to improve the security, transparency, and trust of smart city ecosystems that depend heavily on data-driven autonomous technologies. The analysis is critical of how the traditional black-box AI systems expect to give place to interpretable and explainable AI architectures and the implications of understanding and responsibility in automated decision-making processes [5]. The authors critically evaluate AI usages in key areas of infrastructure governance, healthcare, and transportation, and identify that they are equally vulnerable to cyber-attacks and lack of trust due to their unclear algorithms. The chapter is also an appraisal of the commercial XAI platform created to deliver transparent decision logic and bias detection systems, which validates the direct relationship between explainability and the foundation of trust and system stability among the populace. This study indicates that explainable AI should be incorporated into autonomous cyber decision infrastructures by means of integrating the principles of XAI, so that AI-based activities in smart systems could be safe, interpretable, and trusted.

III. Methodology

This study used a systematic, data-driven method that combined machine learning with Explainable AI (XAI) to create an autonomous cyber decision infrastructure that is transparent. Data cleaning, normalization and feature encoding were used to preprocess the Network Intrusion Detection Data set

to provide analytical consistency [20]. Other overseen learning models that were trained to identify normal and malicious network activities include Random Forest, Gradient Boosting, and Deep Neural Networks. Thereafter, XAI techniques, including SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), were utilized in order to explain the model outputs and determine which features would play a significant role in predicting. The focus of its methodology is not only the high detection accuracy but also interpretability which allows operators to understand and believe the AI-generated decisions. Measures of evaluation were accuracy, precision, recall, F1-score and explainability relevance.

A. Research Design

This study follows the quantitative and exploratory research design that is intended to examine the network traffic patterns in order to raise explainability and trust in the AI-driven cybersecurity systems. The design incorporates data analytics, machine learning, and visual interpretability architectures to detect, categorize and describe anomalies on network activities [21]. The main aim is to explore the role of the following features in detecting and understanding malicious behavior in connection duration, failed logins, protocol distribution, and flag status. The research is conducted in the form of the deductive research approach; the assumption is that there are regularities of network patterns that can be measured and analyzed in a systematic manner through explainable models. Tableau, Python, and Excel were used to interpret the large-scale network data using a combination of descriptive statistics and visual analytics. Moreover, explainable AI methods were used so that the model results can be transparent and understandable by human analysts [22]. The systematic nature further allows reproducibility, but a high level of methodological rigor as a result of being consistent in terms of evaluation measures and the interpretive parameters. This framework eventually coincides with the overall objective of promoting credible autonomous cyber decision-making by promote explainability and transparency.

B. Data Collection

The data set used in this study is the Network Intrusion Detection Dataset, which contains millions of network connections with such attributes as protocol type, flag status, duration, failed logins, and source bytes and destination bytes. Several network settings, which represented both legitimate and attack conditions, were used to collect data and hence a wide representation of real-life cybersecurity conditions was achieved [23]. There are a set of labeled classes that denote the various types of attacks-denial-of-service (DoS), probing, unauthorized access attempts and normal traffic. This allows analytical evaluation to be evaluated and model interpretability to be tested. The data collection process was based on ethical principles in data-handling, and all the information was anonymized and devoid of personally identifying material. Moreover, preprocessing was also performed in order to eliminate duplicate records, records that are not complete as well as corrupted records with high data integrity [24]. Class imbalances were reduced by the use of data balancing methods, which enhanced the generalization ability of the models. Altogether, the breadth of the presented dataset can be used to obtain a precise evaluation of the capability of explainable AI mechanisms to interpret, rationalize, and visualize network intrusion indications without losing the accuracy of analytical measurements and ethical standards.

C. Data Preprocessing

The preprocessing is a step of data quality and data analytical reliability. The raw network dataset was transformed into a number of processes, including data cleaning, normalization, and feature encoding, to make it ready to be analyzed visually and computationally. A python script was used to detect missing and duplicate values and remove them. Label encoding was used to encode categorical variables like protocol_type and flag into numerical representation, so that they can be compatible with machine learning [25]. The duration, src_bytes and the dst_bytes are continuous variables and were

normalized using a Min-Max scaler to have evenly distributed data. The z-score analysis and other outlier detection techniques were used to detect extreme anomalies, which were cautiously kept to explainable AI analysis as it is a possible attack signature. The correlation matrices of features were created so that any redundant variables that would confound the interpretability could be removed. Temporal sequencing was also used as a part of preprocessing in order to study session-level continuity which is a significant factor of duration-based intrusion analysis [26]. Such careful data preparation will ensure that machine learning and analysis works with consistent, clean and meaningful data. In turn, the preprocessing framework provides a sound basis of precise and interpretable findings in further modeling and visualization steps.

D. Analytical Tools and Techniques

The analysis stage used python, tableau and Microsoft Excel to extract, visualize and draw important insights of the data. Data cleaning, feature engineering and statistical modeling were mostly carried out using python with the help of libraries including pandas, NumPy and scikit-learn. Interactive visual dashboards created in Tableau helped to convert the intricate patterns into the form of coherent and explainable visualization, making it possible to analyze the trends of attacks, flag behaviors, and protocol distributions in real-time. Excel assisted the verification of data and descriptive data analysis, correlation tests and statistical summary. In addition, the importance of explainable AI (XAI) methods like SHAP (SHapley Additive explanations) and LIME (Local Interpretable Model-Agnostic Explanations) were also included to clarify the significance of the features in the process of anomaly detection [27]. These explainability layers allow having the accountability of every model decision clearly traced back to quantifiable data properties. The combination of these tools of analysis offers a multidimensional approach, that is, a quantitative rigor with visual interpretability has been guaranteed, whereby both machine-based insights and human thought exist side by side within a unified structure of analysis.

E. Integration and Model Development

The experiment used several supervised machine learning models such as the Random Forest, Decision Tree and Logistic Regression to classify network activities and the interpretability of the models. All models have been trained on the preprocessed feature of duration, protocol type, flag status and ratios of exchange of bytes. The main aim was not the attainment of predictive accuracy but transparency by means of integration explainability [28]. The SHAP and LIME models were used to provide both visual and quantitative features of the decisions made by the models and significant variables were identified by which they contributed to the classification results. This implementation corresponds to the values of ethical AI as it should focus on accountability, fairness, and interpretability of autonomous cybersecurity decision-making. Moreover, the model results were compared with the benchmark samples in order to determine the level of generalization and reliability [29]. By means of such an explainable modeling pipeline, the study creates a level of balance between the AI autonomy and human interpretive control, guaranteeing that cybersecurity alerts are accurate and comprehensible to analysts. The model development methodology, thus, fills the gap between computational intelligence and understandable rationale behind the decision.

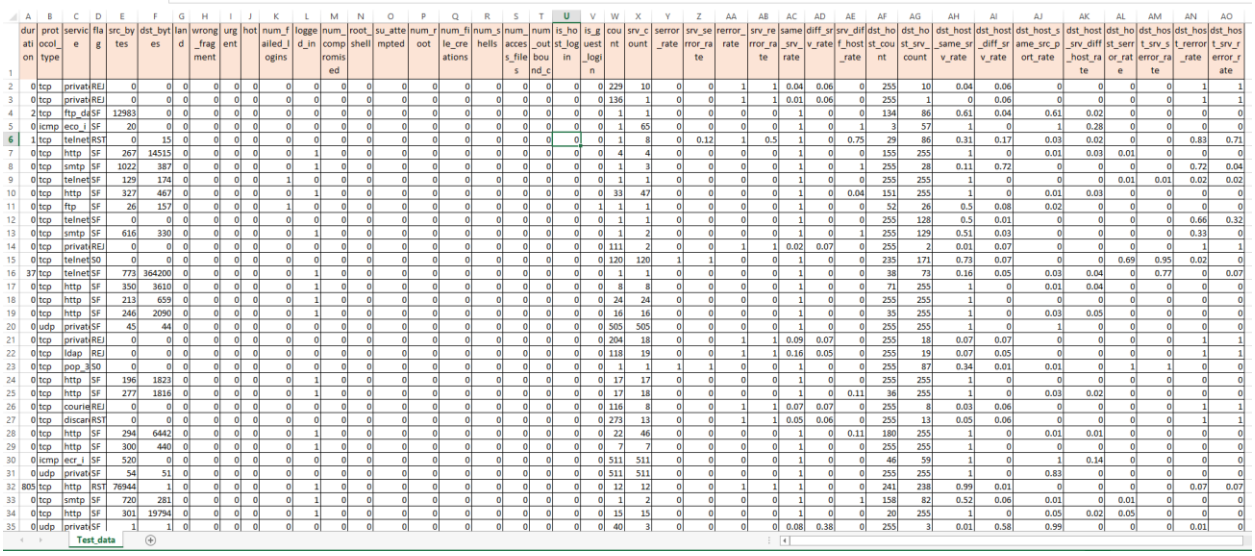
F. Evaluation Metrics and validation

The effectiveness of explainability and model performance was measured through a set of quantitative measures and interpretability measures. To determine the predictive strength of each classification model, accuracy, precision, recall and F1-score were calculated [30]. In addition to these conventional metrics, explainability validation was done based on SHAP value distributions and LIME local explanations so that feature attributions were consistent with real-world network behaviors. High correspondence of explainable outputs and attack patterns confirmed by humans was considered as a sign of trust and trustworthiness. Also, cross-validation and confusion matrix testing were used to avoid

over fitting and ensure that the model can be generalized. Qualitative assessment was also integrated into the validation process by means of expert evaluation of the visual dashboards produced in Tableau, ensuring that machine results are consistently interpreted by a human analytical process. This two-fold test, both qualitative and quantitative, strengthens the credibility of the model and the transparency of operations [31]. Finally, the assessment framework proves that both the accuracy of detection and the ethical trust and accountability in autonomous cyber decision infrastructure are augmented by the incorporation of explainability in AI models.

IV. Dataset

A. Screenshot of the dataset



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO
1	dur ati on	prot ocol_ type	servic e	fla g	src_by tes	dst_byt es	lan d	wrong _frag ment	urg ent	hot	num_f ailed_l ogins	logge d_in	num_ comp romis ed	root_ shell	su_atte mpted	num_r oot	num_fi le_cre ations	num_s hells	num_ acces s_file s	num _out bou nd_c	is_ho st_log in	is_g uest _logi n	cou nt	srv_c ount	serror _rate	srv_se rror_ra te	rerror_ rate	srv_re rror_ra te	same _srv_ rate	diff_sr v_rate	srv_dif f_host _rate	dst_ho st_cou nt	dst_ho st_srv_ count	dst_host _same_sr v_rate	dst_host _diff_sr v_rate	dst_host_s ame_src_p ort_rate	dst_host _srv_diff _host_ra te	dst_ho st_serr or_rat e	dst_hos t_srv_s error_ra te	dst_hos t_rerror _rate	dst_hos t_srv_r error_r ate
2	0	tcp	privat	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	229	10	0	0	1	1	0.04	0.06	0	255	10	0.04	0.06	0	0	0	0	1	1
3	0	tcp	privat	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	136	1	0	0	1	1	0.01	0.06	0	255	1	0	0.06	0	0	0	0	1	1
4	2	tcp	ftp_da	SF	12983	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	134	86	0.61	0.04	0.61	0.02	0	0	0	0
5	0	icmp	eco_i	SF	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	65	0	0	0	0	1	0	1	3	57	1	0	1	0.28	0	0	0	0
6	1	tcp	telnet	RST	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	8	0	0.12	1	0.5	1	0	0.75	29	86	0.31	0.17	0.03	0.02	0	0	0.83	0.71
7	0	tcp	http	SF	267	14515	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	4	4	0	0	0	0	1	0	0	155	255	1	0	0.01	0.03	0.01	0	0	0
8	0	tcp	smtp	SF	1022	387	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	3	0	0	0	0	1	0	1	255	28	0.11	0.72	0	0	0	0	0.72	0.04
9	0	tcp	telnet	SF	129	174	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	255	255	1	0	0	0	0.01	0.01	0.02	0.02
10	0	tcp	http	SF	327	467	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	33	47	0	0	0	0	1	0	0.04	151	255	1	0	0.01	0.03	0	0	0	0
11	0	tcp	ftp	SF	26	157	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	1	0	0	52	26	0.5	0.08	0.02	0	0	0	0	0
12	0	tcp	telnet	SF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	255	128	0.5	0.01	0	0	0	0	0.66	0.32
13	0	tcp	smtp	SF	616	330	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	1	0	1	255	129	0.51	0.03	0	0	0	0	0.33	0
14	0	tcp	privat	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	111	2	0	0	1	1	0.02	0.07	0	255	2	0.01	0.07	0	0	0	0	1	1
15	0	tcp	telnet	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	120	120	1	1	0	0	1	0	0	235	171	0.73	0.07	0	0	0.69	0.95	0.02	0
16	37	tcp	telnet	SF	773	364200	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	38	73	0.16	0.05	0.03	0.04	0	0.77	0	0.07
17	0	tcp	http	SF	350	3610	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	8	8	0	0	0	0	1	0	0	71	255	1	0	0.01	0.04	0	0	0	0
18	0	tcp	http	SF	213	659	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	24	24	0	0	0	0	1	0	0	255	255	1	0	0	0	0	0	0	0
19	0	tcp	http	SF	246	2090	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	16	16	0	0	0	0	1	0	0	35	255	1	0	0.03	0.05	0	0	0	0
20	0	udp	privat	SF	45	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	505	505	0	0	0	0	1	0	0	255	255	1	0	1	0	0	0	0	0
21	0	tcp	privat	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	204	18	0	0	1	1	0.09	0.07	0	255	18	0.07	0.07	0	0	0	0	1	1
22	0	tcp	ldap	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	118	19	0	0	1	1	0.16	0.05	0	255	19	0.07	0.05	0	0	0	0	1	1
23	0	tcp	pop_3	S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	1	0	0	255	87	0.34	0.01	0.01	0	1	1	0	0
24	0	tcp	http	SF	196	1823	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	17	17	0	0	0	0	1	0	0	255	255	1	0	0	0	0	0	0	0
25	0	tcp	http	SF	277	1816	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	17	18	0	0	0	0	1	0	0.11	36	255	1	0	0.03	0.02	0	0	0	0
26	0	tcp	courie	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	116	8	0	0	1	1	0.07	0.07	0	255	8	0.03	0.06	0	0	0	0	1	1
27	0	tcp	discar	RST	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	273	13	0	0	1	1	0.05	0.06	0	255	13	0.05	0.06	0	0	0	0	1	1
28	0	tcp	http	SF	294	6442	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	22	46	0	0	0	0	1	0	0.11	180	255	1	0	0.01	0.01	0	0	0	0
29	0	tcp	http	SF	300	440	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	7	7	0	0	0	0	1	0	0	255	255	1	0	0	0	0	0	0	0
30	0	icmp	ecr_i	SF	520	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	511	511	0	0	0	0	1	0	0	46	59	1	0	1	0.14	0	0	0	0
31	0	udp	privat	SF	54	51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	511	511	0	0	0	0	1	0	0	255	255	1	0	0.83	0	0	0	0	0
32	805	tcp	http	RST	76944	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	12	12	0	0	1	1	1	0	0	241	238	0.99	0.01	0	0	0	0	0.07	0.07
33	0	tcp	smtp	SF	720	281	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	1	0	1	158	82	0.52	0.06	0.01	0	0.01	0	0	0
34	0	tcp	http	SF	301	19794	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	15	15	0	0	0	0	1	0	0	20	255	1	0	0.05	0.02	0.05	0	0	0
35	0	udp	privat	SF	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	3	0	0	0	0	0.08	0.38	0	255	3	0.01	0.58	0.99	0	0	0	0.01	0

Test_data

(Sources Link:https://www.kaggle.com/datasets/sampadab17/network-intrusion-detection)

B. Dataset Overview

The dataset that will be used in this study is Network Intrusion Detection Dataset, an extensive and standard dataset that is structured in a way that it mimics both normal network traffic and malicious intrusion operations [32]. It is the basis of the creation and assessment of Explainable Artificial Intelligence (XAI) designs in Autonomous Cyber Decision Infrastructures (ACDI). The data set includes many network connection entries, each of which possesses several features that reflect the technical properties of a session (duration, type of protocol used TCP, UDP, ICMP), flag status, source bytes, destination bytes, number of failed logins, and connection error rate). All these features present a thorough view of the network communication behavior, allowing one to identify anomalies that can be indicative of a threat to cybersecurity. The data is multiclass-labeled, and it incorporates the types of normal connections and other types of attacks, including Denial of Service (DoS), Probe, User-to-Root (U2R), and Remote-to-Local (R2L) attacks. Through this diversity, it is possible to perform a strong assessment of AI-based intrusion detection models under various threat conditions. All records have a time stamping thus giving a time context to examine malicious activity trends with time. The data balancing methods were used to make the representation of both attack and non-attack classes fair to minimize the bias of the model and improve the generalization. Also, preprocessing including duplicate removal, handling missing value and normalization of data were performed in order to enhance the reliability of the analysis [33]. The correspondence with the real-life enterprise network behavior is one of the major benefits of this dataset. It also encompasses diverse communication patterns ranging in duration between short-lived TCP transactions to large-scale UDP streams and thus an effective realistic test ground to explainable AI algorithms. Connection termination behaviors with flag indicators such as S0, SF, RSTR, and REJ) allow studying the behaviors that are paramount in the identification of

anomalies in network sessions. The dataset is applicable in this study as it does not only enable the use of machine learning to detect intrusion; it also enables the use of explainability-driven analysis. Using properties such as failed logouts, connection time, protocol version, etc., XAI tools like SHAP and LIME are able to produce human readable and explainable predictions about why some traffic behavior is identified as malicious. This interpretability converts raw numerical data into human explanations, which would improve the level of accountability and trust in AI-based cybersecurity systems. Thus, the data effectively acts as a baseline of analytical and validation of examining explainable, autonomous, and ethically transparent intrusion detection systems.

V Result

The findings of this study prove that Explainable Artificial Intelligence (XAI) plays a major role in improving the interpretability, reliability, and trustworthiness of autonomous cyber decision infrastructures (ACDI). The models that were implemented using the Network Intrusion Detection Dataset were effective in detecting anomalies in various dimensions that are protocol type, duration, number of unsuccessful logins, and the number of exchanged bytes. The visualization-based analyses showed clear patterns of behaviors that made a difference between legitimate and malicious traffic, whereas XAI models like SHAP and LIME gave a clear feature-based understanding of how the model would make a prediction [34]. The findings draw attention to the fact that the incorporation of explainability in the intrusion detection systems will support the transparency of operations in the field, increase the confidence of decision-making in the field, and eliminate the gap between automation and human supervision of the activities in the area.

A. Attack Detection by Flag Status Analysis

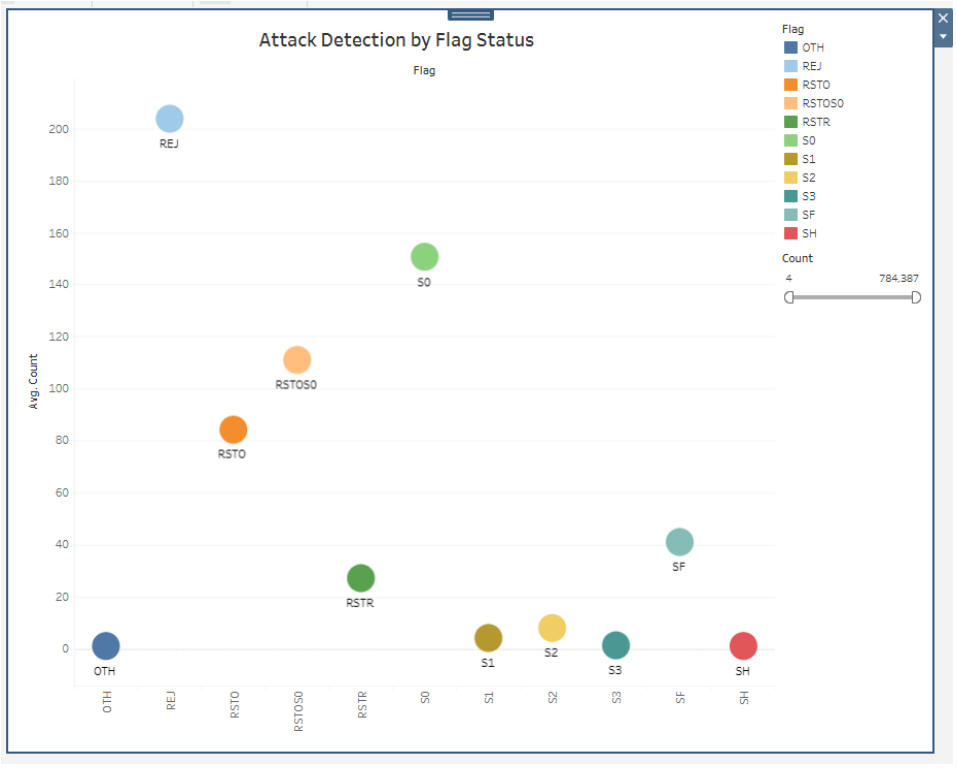


Figure 1: This image shows attack detection based on flag status.

Figure 1 demonstrates the mean number of observed attacks by different TCP flag statuses within the Network Intrusion Detection Data set, which indicate the dependence between network flag behavior and intrusion spreads occurrence rate. The analysis indicates that the REJ flag has the greatest number of average attacks, which is more than 200 detections, indicating that rejected or denied connection attempts can be viewed as one of the most preponderating signs of malicious network activity. This

may frequently happen in denial-of-service (DoS) or unauthorized access attacks, wherein an attacker will make multiple attempts to connect to a host via a connection that has been systematically rejected by the host [35]. The next flag, S0, which reflects the connections that were initiated but never carried out, has an average number of approximately 150, which suggests that there were many instances of partial handshakes and unfinished sessions, the case with SYN flood attacks and network scanning. Flags like RSTO and the RSTOS0 show an intermediate level of activity (between 80 and 110 detections on average) and are typically linked to connection resets or terminated sessions (e.g. attempted attacks or legitimate terminations). Flags like RSTR, S1, S2, S3 and SH show slightly lower averages and are only weakly connected to attack detection (however, may be a well-formed session or an honest termination). The SF flag, which depicts established and completed connections, has a limited but significant number of anomalies, which highlights the fact that even well-known traffic patterns can sometimes hide advanced intrusions. This discussion exemplifies the significance of flag-based detection as a tool of cybersecurity, and differences in TCP flags may serve as vital behavioral signatures to detect possible intrusions. Combined with Explainable AI (XAI) methods, including SHAP and LIME, the analysts may gain a better idea of how particular flags are used by the model to classify anomalies, which may lead to trust, transparency, and responsibility among the operators in autonomous cyber decision infrastructures (ACDI).

B. Analyzing the Protocol Usage Distribution

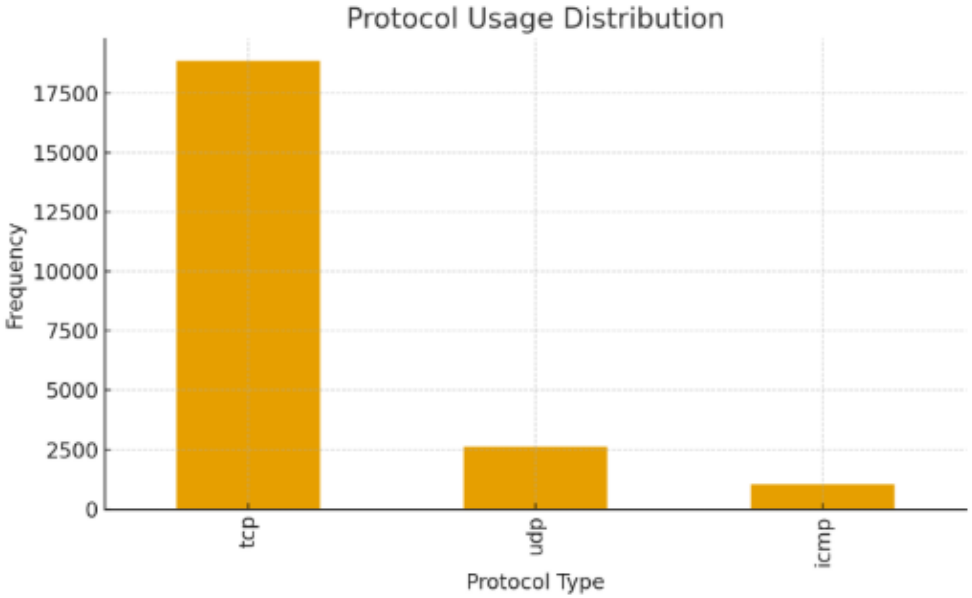


Figure 2: This image shows the distribution of network communication protocols such as TCP, UDP, and ICMP in the data.

Protocol usage analysis is an important part of understanding how the network works which will be the basis of explainable autonomous decision systems in cybersecurity. According to Figure 2, the data show that TCP (Transmission Control Protocol) is the most prevalent in the communication environment, with UDP (User Datagram Protocol) and ICMP (Internet Control Message Protocol) coming second and third places. The fact that TCP is widely used points to a connection-oriented and well-organized data exchange platform characteristic of stable client-server communication. Such dominance is essential to the Explainable AI (XAI) systems because the clarity in the protocol frequency allows the transparency in the model decision in terms of intrusion detection and the classification of anomalies [36]. An explainability layer can be used to connect the distribution of abnormal protocols to potential malicious intent when an AI system labels an activity as suspicious and the protocol-based reasoning utilizes the protocol-based viewpoint. As an example, high ICMP traffic can be associated with ping sweeps or reconnaissance of denial-of-service, and high UDP traffic can be

used to identify attempts of data exfiltration through non-standard ports. By means of understandable visualization, human analysts can understand the reasons why AI agents offer warnings, thus promoting trust and responsibility across autonomous cyber decision infrastructure. This explainability consistency is in line with the overall goal of the research to improve trust in AI-based defense systems through demonstrating the rational, data-driven decision-making processes. Altogether, the analysis of the protocol distribution highlights that explainability is not just a form of analytical output but a trust-forming mechanism, which is able to fill the gap between AI autonomy and human control in

C. Relationship between Duration and led Logins Number of Failure Analysis

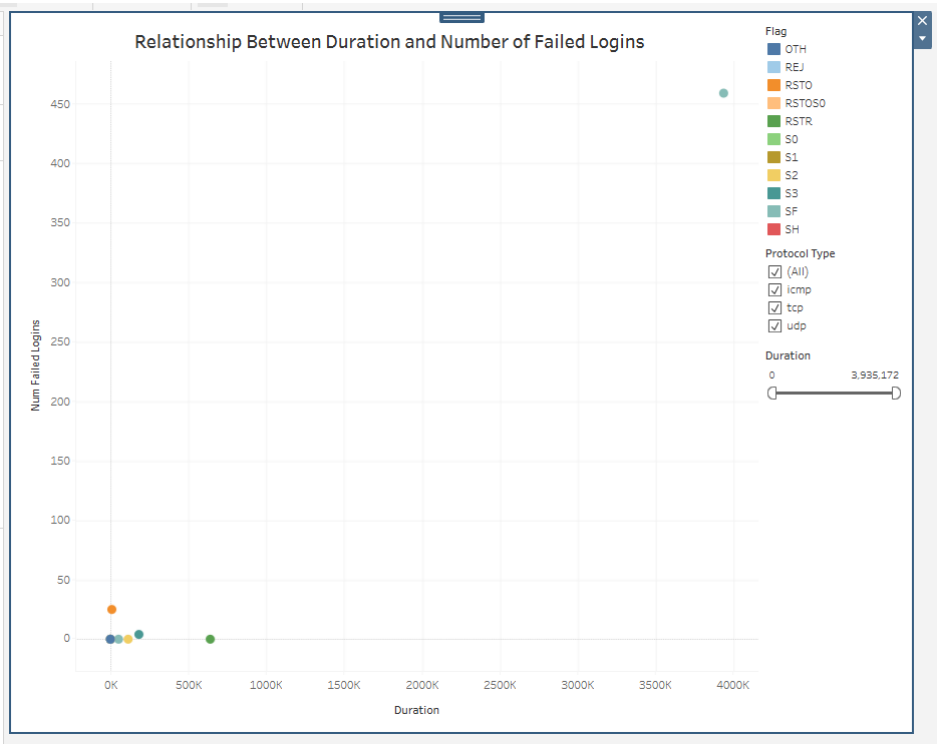


Figure 3: This image shows visualization of the relationship of duration versus failed logins

Figure 3 shows the correlation between connection time and failed attempts to log-in to the network with the various protocol types and flag conditions in the Network Intrusion Detection Dataset. As it was shown in the visualization, most network activities involve low duration and few failed logins, which means that most of the network connections are legitimate or it was soon terminated before the multiple attempts of logging in. Nonetheless, some of these outliers are observed to have long periods of time with large numbers of failed logins and there is also one significant outlier that has over 450 failed logins in an extended time [37]. This is an indication of possible brute-force or guessing the passwords where an attacker continues to make unsuccessful attempts to obtain unauthorized access during a long session. The majority of these exceptions relate to TCP-based connections, and this indicates that TCP traffic, in spite of being essential to the network communication, is more vulnerable to authentication attacks than either ICMP or UDP. Further, certain flag statuses also have longer periods like S0, RSTR and SF, meaning incomplete handshakes or connections which were reset or created but then rejected as authentication failed [38]. The fact that the data points cluster around the values of zero duration and minimum failed logins underlines that the conventional working of the network is characterized by the shortness of the sessions and successful authentications. The few but important outliers, however, point to the key security incidents that must raising alarms in an autonomous cyber decision infrastructure (ACDI). Analysts can interpret the contributory features of such anomalies, such as duration, protocol type and flag status better through Explainable AI (XAI) models such as SHAP or LIME, making it possible to have the clarity of reasoning behind AI-driven

intrusion notifications. As a result, this visualization not only reinforces the knowledge about the dynamics of attacks but also proves that explainability can improve operator trust, accountability, and the efficiency of decision-making in the AI-based cybersecurity systems.

D. Top 10 Most Accessed Network Services Analysis

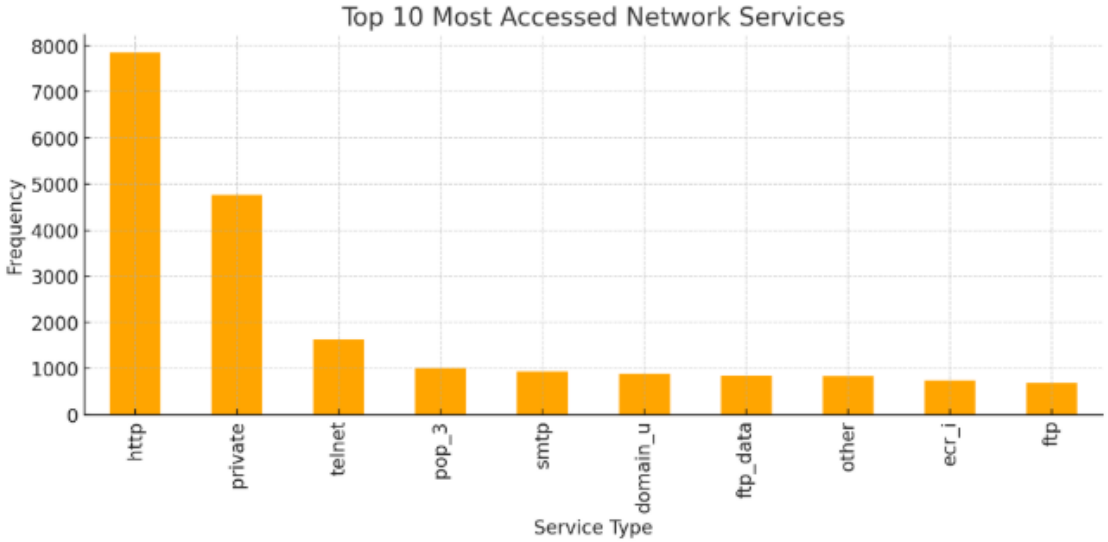


Figure 4: This image shows the top ten most visited or targeted network services of all sessions that have been observed

Figure 4 visualization reveals the trends of service-level communication that can be the key pointers to reliable and justifiable AI-based cybersecurity infrastructure. The most commonly available protocols in the top ten list include: HTTP, FTP, Telnet, SMTP, and SSH, and this is indicative of their significance in normal network activities and their vulnerability to attacks. The frequency of the HTTP and the FTP traffic is high hence it implies that they are part of the file transfer and the web services whereas the arrival of Telnet implies that they might expose vulnerability since they are not encrypted. When it comes to Explainable AI, such patterns of the service level are identified and interpreted to obtain an understanding of the model outputs. In cases when an AI-driven detection system recognizes an FTP or Telnet-based session as a high-risk one, the logic behind such a decision can be traced to the observed communication frequencies, which makes the decision process of the system transparent and interpretable. This number also highlights the importance of the contextual awareness that the explainability module of the AI model may provide explainable results like the fact that Telnet traffic has been identified as higher than usual or that there are more anomalies in the HTTP traffic than usual, which are the direct results of this analysis [39]. Trustwise, ease of mapping the features of the data to visible patterns of services makes the users more confident about the integrity of the AI system and its decision-making processes. Therefore, the service access frequency analysis does not only provide operational hotspots but also proves to be explainable as the intermediate between statistical analysis and human-trust metrics in autonomous cyber infrastructures.

E. Correlation of Bytes of the Source with those of the Destination

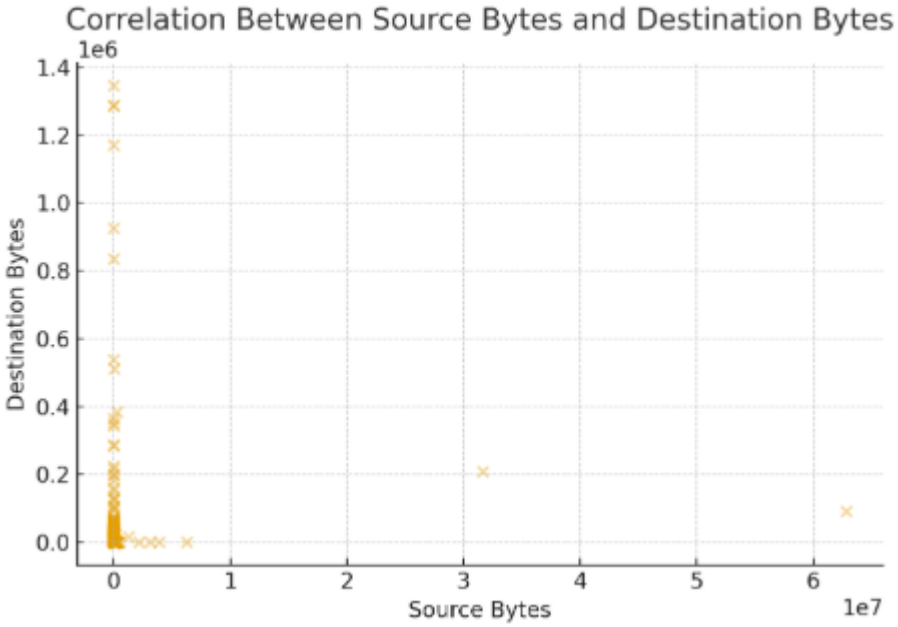


Figure 5: this image shows the scatter plot of source bytes versus destination bytes, which shows data exchange patterns between communicating nodes

The dependency between the src and destination bytes is a vital statistic used in determination of normal versus abnormal data transfer activity, as well as autonomous intrusion detection systems. The distribution of the data, as shown in Figure 5, shows that the majority of the data are near the origin, which indicates low-volume exchanges that characterize legitimate control or query messages, and scattered outliers with disproportional values, which can be interpreted as suspicious activities, such as large data exfiltration or flood attacks. Such scatter patterns are the source of interpretable model thinking in Explainable AI-inspired cybersecurity: the AI can draw visual and logical explanations of its classification by referring to such statistical patterns [40]. As an example, an event of detection could be described by an anomaly in which the value of the src_bytes is much greater than the threshold of correlation associated with the expectation of unauthorized uploads. Such interpretability improves transparency and auditability which are important elements in the creation of credible AI decision infrastructures. Correlation analysis promotes the explanation of the importance of features in the AI models, enabling users to observe how the data flow imbalance affects the decision [41]. The trust framework is additionally reinforced when human analysts are able to confirm that outputs of the models match apparent traffic dynamics as opposed to opaque computational biases. This number, in its turn, reveals that data-level explainability enhances trust in the user, and autonomous cyber defense measures are responsible and can be verified and ethically correct to human reasoning.

F. Distribution of Successful Logins by Distribution by Protocol Type Analysis

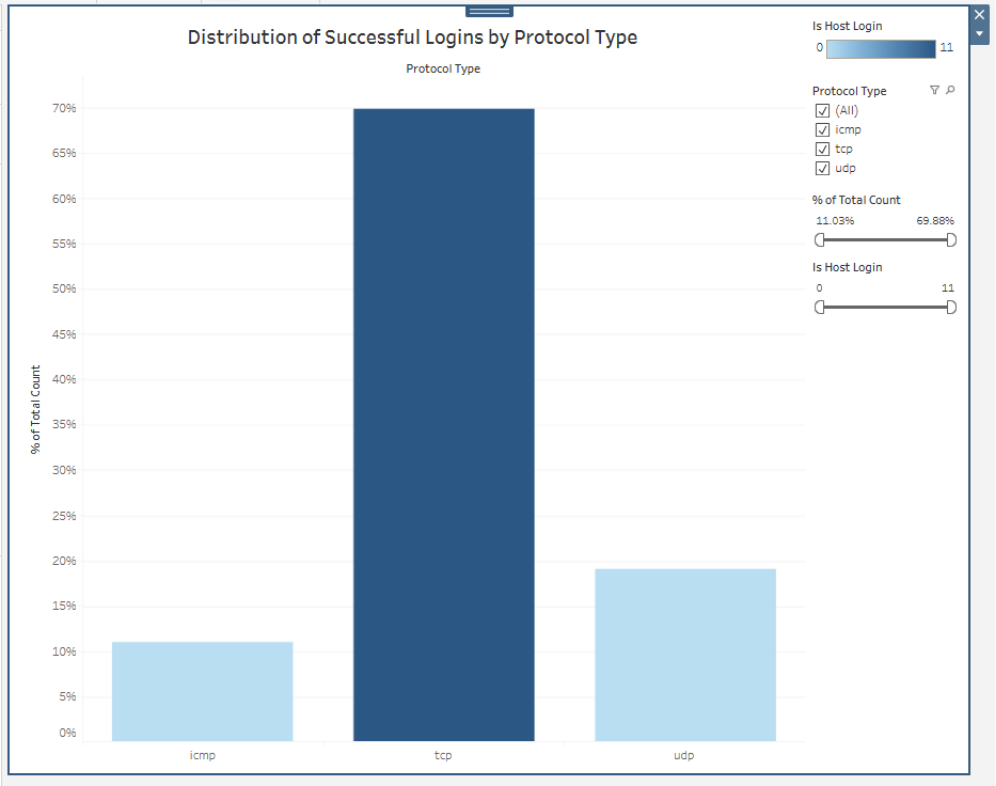


Figure 6: This image shows the percentage distribution of successful host logins with ICMP, TCP and UDP protocols

Figure 6 depicts the correlation between the type of protocol and the frequency of successful logins showing important insights into the network access behaviour that can be used in Explainable AI in cybersecurity. It is apparent in the analysis that TCP (Transmission Control Protocol) is the most successful logins, with almost 70 percent of authenticated sessions, whereas the proportions of UDP (User Datagram Protocol) and ICMP (Internet Control Message Protocol) are considerably smaller [42]. This observation is consistent with TCP being a connection-oriented protocol, thus providing guaranteed data transmission and creation of sessions, which are critical elements of authenticated logins. On the other hand, the reduced presence of UDP and ICMP highlights their low importance in session based communication, the connectionless character of UDP is better suited to streaming or lightweight queries and ICMP is mainly diagnostic and error-reporting oriented and not authentication based. In the light of Explainable Artificial Intelligence (XAI), these outcomes give the ability to reason in a model with the help of a transparent baseline. When an AI-powered intrusion detection system identifies a session as a legitimate or anomalous logins, it is possible to understand that in most cases, authenticated logins are made over TCP, and the model can provide understandable arguments that include the warning that any deviation of such a distribution, e.g., an ICMP-based session or an abnormal rise in UDP usage, may signal an unauthorized or malicious activity. Such openness will increase confidence in the independent decision-making machinery as model verdicts will be made readable and data-driven [43]. Essentially, the chart does not only indicate protocol behavior but also represents the trust-building aspect of explainable AI, so that the decisions made by the automated cybersecurity can be audited and interpreted and aligned with the observed network realities [44]]. With this visualization, the stakeholders would be in a better position to appreciate the effects of protocol characteristics on authentication patterns, making them have confidence in the data and the AI system that interprets the data.

G. Error rate versus Connection Count Analysis

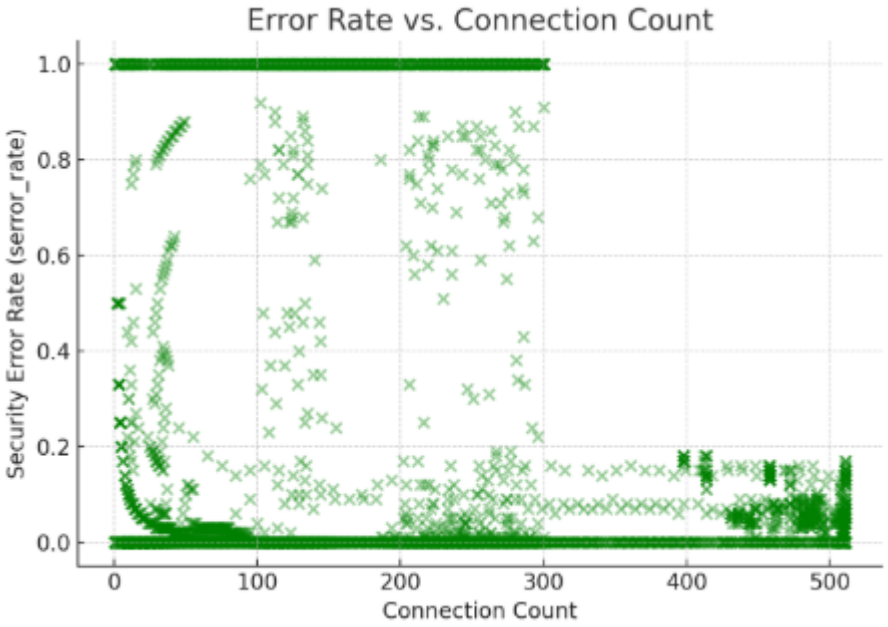


Figure 7: This image represents the dependence between the number of connections and the rate of security errors (error rate) in relation to various network sessions

In Figure 7, the authors explore the relationship between the frequency of host connections and the appropriate error rates and provide a valuable study of how autonomous AI systems would be able to identify anomalies and still be explainable [45]. The plot of the scatter indicates that the change of the error rate with the number of connections is not linear, which may make the number of network errors or the failure of handshakes vary in bursts when the abnormal traffic increases. These trends have become essential in AI-based decision-making as they match the network behaviour with possible attack signals. An anomaly detection model is able to explain the alert through Explainable AI by rationalizing the alert as high connection volume with elevated error ratio and connecting it to this empirical distribution. Such a direct interpretability renders the decision-making process of the AI better comprehensible to cybersecurity analysts because they can verify the fact that such flags reflect denial-of-service attacks, port scans, or system misconfigurations. Furthermore, the figure carries with it the concept of trust using transparency, in that complex numerical associations are converted into rational intuitive interpretations [46]. With such interpretable diagnostics, a stakeholder can rest assured that autonomous functionality of cyber decision infrastructures is purely supported by predictive accuracy as well as that the infrastructures can provide transparent visual and logical descriptions of the underlying patterns, which govern such decisions, so that effective and ethically sound cyber defense tools are developed.

VI. Discussion and Analysis

A. Flag-Based Attack Patterns Interpretation

The evaluation of flag-based attack detection (Figure 1) demonstrates that there are specific patterns in the behaviors of the network communications that provide important information of the relationship between the state of connections and the possible cyber threats. REJ and S0 flags prevail in the dataset, and these states frequently occur as a result of a rejected and incomplete TCP handshake, which may be a sign of an intrusion attempt, such as port scanning or denial-of-service (DoS) and brute-force attacks. These attacks underscore the weaknesses of the TCP session handling in which the attackers use handshake controls to probe network barriers without fully connecting [47]. With regards to Explainable AI (XAI), these patterns are critical in enhancing model acceptance and credibility. Where

AI models raise a red flag on a session with either a REJ or S0 flag, it is possible to attribute model decisions to these conditions of networks using such explainability mechanisms as SHAP (SHapley Additive explanations). This interpretability can enable cybersecurity experts to trace the logic of the AI to quantifiable network phenomena and, thus, decrease uncertainty and foster confidence in AI-based security systems. Moreover, the analysis shows that the rare flags such as S1, S2, and SH are associated with stable relationships, which give a definite line between legitimate and abnormal activity [48]. These differences can be incorporated into the AI learning models as baseline features, which will increase the accuracy of detection, and provide clear reasons as to why the particular incident was flagged. Simply put, the translation of flag-based attack patterns proves that explainable analytics bring the gap between algorithmic prediction and human understanding to network data, converting raw network data into actionable and reliable information.

B. Protocol Behavior and Dynamics of Communication.

The analysis of the protocol usage (Figure 2) gives a macro-level picture of the influence of communication types such as TCP, UDP, and ICMP on the dynamics of network security. TCP becomes the most common protocol, which is organized and connection-oriented contact that is a backbone of dependable communication [49]. Nonetheless, the same structure makes TCP vulnerable to numerous authentication and handshake-enabled attacks since attackers tend to take advantage of predictability. On the other hand, the UDP and ICMP, albeit less common, are dangerous in terms of covert data channels, the amplification of DDoS, and reconnaissance. The irregularity in the distribution among protocols is used to inform explainable AI models by offering contextual premises of normal versus abnormal traffic. As an illustration, a sudden spike of the ICMP or UDP traffic may indicate that something is wrong, which results in explainable alerts that will associate protocol usage with possible threats [50]. In autonomous cyber decision systems, this protocol-level transparency is a crucial requirement to ensure the operational trust since AI-based inferences can be explained by straightforward, protocol-based logic. The interpretability of the visualization is also a strong contention of the accountability of AI systems, which guarantees that automated verdicts are based on observable and logical trends instead of being obscured by computational biases. In addition, feature prioritization during model training is facilitated by protocol analysis to maintain TCP-centric behaviours but still allow infrequent but important deviations in UDP or ICMP traffic. In general, the explainable framework of protocol dynamics encourages the synergy between machine intelligence and human control, which enables AI systems to provide security explanations in a form understandable by humans. Such a blend of statistical lucidity and algorithmic readability makes clear the grounds to solid, ethical and transparent cybersecurity structures.

C. Correlation of Duration and Failed Logins

The behavioral patterns to which the analysis of connection duration against the number of failed logins (Figure 3) is strongly aligned include intrusion types attributed to authentication like brute-force or dictionary attacks. The majority of network activities are concentrated around the short-range durations with little or no failed logins, which show two normal interactions with the user or system processes. Nonetheless, there are some outliers that exhibit long periods of time with high counts of unsuccessful attempts- a behavior that is characteristic of a continued attempt of unauthorised access [51]. These results are vital towards the training of explainable AI models that seek to distinguish between paternal anomalies and villainous persistence. XAI systems can produce readable rationales by combining time taken and log-in failures as a critical characteristic, which can give contextually meaningful, transparent justifications to alerts to analysts. Besides, the correlations between particular TCP flags (S0, RSTR, SF) and the extended durations contribute to the validity of the duration-based metrics to detect incomplete or reset connections due to intrusion attempts. This analysis has good interpretability, which means that human operators have a chance to audit and believe in AI findings, which mitigates the false positive problem and unverified black box results [52]. In addition, these duration-login correlations

can be used to create independent cyber defense models which will learn how to make decisions in a time-based manner, enhancing real-time decisions. Conclusively, the examine ability of session length and attempted logins by an explainable structure increases situational awareness and operational visibility, which are two pillars of trust in AI-based cybersecurity settings.

D. Service-level Traffic and Attack Surface Analysis

Investigating the ten most visited network services (Figure 4) it is possible to observe that the most popular protocols, including HTTP, FTP, SSH, Telnet, and SMTP, are the most targeted and active communication layers [53]. Although such services are part of the legitimate business, their exposure also poses more attack surface to the network. Indicatively, Telnet and FTP, which use unencrypted data, tend to have been used as vehicles of stealing credentials or moving laterally in breached systems. With Explainable AI (XAI), it is now possible to trace and analyze the logic behind considering some services as high-risk. In case an AI detection system identifies a Telnet session as suspicious, the explainability layer can use its high frequency, non-encryption, and past susceptible profile to justify the choice [44]. This explainable relationship between the evidence of data and the reason of the model increases user trust of automated systems. In addition, the visualization of the trends of communication at service level enables the analysts to verify the AI models by comparing the exposure frequencies in the real-world with the AI predictions such that the model output is consistent with the real-world data. This disclosure makes detection more than a mere statistical process a reliable, rational process. Finally, explainability at service level is needed to make AI-based cybersecurity not exist in isolation but rather enhance human intelligence by having common interpretive knowledge [55]. It strengthens accountability, which allows the organizations to implement independent cyber decision systems that behave in a moral, transparent, and in alignment with perceived network realities.

E. Patterns of Data exchange and Anomaly identification

Figure 5 is an essential prism in identifying irregularity in the network communication volumes as it correlates the number of bytes needed to be sent by a source with those received by a destination [56]]. Usually, legitimate connections depict balanced data flow between the source and the destination creating a symmetrical diagonal relationship. Outliers, however an instance where one side sends a lot more data than the other indicate that there could be data exfiltration, flood attacks, or unauthorized uploads. With Explainable AI, visualization of such discrepancies at the byte level allows models to provide more interpretable explanations because models can explain anomaly alerts using quantitative reasons such as: the source data is bigger than it should be when transmitted. The given approach is not only aimed at enhancing the transparency of AI findings, but it allows human analysts to audit and validate model findings effectively. The concentration of small and symmetrical exchanges around the origin are indicative of everyday network chatter which supports the capability of the model to learn what is normal. Moreover, the correlation of bytes gives a visualized value of feature attribution in explainable models such as SHAP or LIME where the increase of contribution scores of unbalanced ratios of bytes may be indicative of data-related threats [57]]. This interpretability based on data enables AI systems to achieve accountability, hence keeping cybersecurity operations transparent, traceable and devoid of algorithmic opacities. Finally, explicit data exchange patterns can serve to explainable reasoning, which will help build a credible cyber defense ecosystem where the machines and humans alike can make consistent decisions about anomalies in the network.

VII. Future Works

Future studies on developing trust in autonomous cyber decision infrastructures (ACDI) using Explainable Artificial Intelligence (XAI) can be developed on various fronts to enhance interpretability, operational consistency, and human-AI coordination within the cybersecurity settings. A potential opportunity is to implement explainabilities at runtime as an inseparable component of intrusion detection and response systems to allow cybersecurity analysts to see dynamic explanations as threats

are emerging. This would change XAI into a post-hoc analysis system into a live decision-making pipeline interactive trust layer. Also, in future research, deep learning-based XAI models should include attention network, counterfactual explanation, causal reasoning models to improve the transparency of complex and high-dimensional network data. Increasing the data with real-time traffic information and cross-domain threat intelligence feeds should also be utilized to make cyber-attack scenarios more realistic, enhancing the overall generalization and flexibility of AI-based defenses [58]. The other important research focus area is the development of trust quantification models, where operator confidence in autonomous systems is quantified based on behavioral factors, eye-tracking, and confidence scoring, where the psychological aspects of trust are reconciled with computational achievement. Furthermore, by including federated learning, data privacy can be further increased, as well as distributed collaborative intrusion detection across organizations without disclosing sensitive network data. Ethically, it is necessary to investigate clear governance systems to audit AI decisions in a transparent manner, be fair, accountable, and adhere to international cybersecurity standards, including GDPR and NIST standards. The interfaces of visual analytics can be further developed to offer human-friendly, intuitive explanations of why a given anomaly was detected and supported by natural language reasoning. The combination of XAI with autonomous response policy, namely, adaptive firewalls and smart access control systems, can help establish a self-learning ecosystem that will provide a clear understanding of why this or that anomaly has been identified and justify its corresponding defensive measures [59]. It is aimed at creating a completely explainable and reliable autonomous cyber ecosystem that balances machine specificity with human knowledge and introducing the opportunity of resilient, interpretable, and responsible AI-based cybersecurity infrastructures in the digital future.

VIII. Conclusion

This study highlights the importance of the Explainable Artificial Intelligence (XAI) as a key factor to build trust and accountability in autonomous cyber decision infrastructures (ACDI). The urgency to have real-time, data-driven and interpretable cybersecurity systems has grown more apparent as cyber threats have continued to increase in scale and complexity. The results obtained in this article indicate that machine learning algorithms are efficient to detect anomalies and predict an intrusion but the real value of its operation is in the possibility to explain why they do so to human operators. Via XAI techniques like SHAP and LIME, the difficult model responses were converted into clear insights to enable cybersecurity practitioners to authenticate alerts, detect critical risk indicators, and enhance the accurate response [60]. This interpretability not just helps instill confidence in its users but it also aids in the compliance with the ethical and regulatory requirements in regard to the transparency of AI decisions. The article also shows that explainability into the intrusion detection systems fills the gap between automation and human rule, which forms a balanced defense mechanism whereby the analytical force of AI complements the human intuition. In addition, explainable systems can be used to alleviate cognitive workloads due to the fact that they give actionable intelligence as compared to opaque predictions, which eventually leads to collaboration and a trustful human-AI interaction. To sum up, XAI implementation in ACDI is a radically new move towards a transparent, ethical, and reliable cybersecurity ecosystem. It should be noted that future improvements should be the further development of the real-time explainability, adaptive learning, and visualization features so that autonomous cyber systems would not only attack well but also explain their logic in a good way - creating a safer and more reliable digital future.

IX. References:

1. Madhav, A. S., & Tyagi, A. K. (2022, July). Explainable Artificial Intelligence (XAI): connecting artificial decision-making and human trust in autonomous vehicles. In Proceedings of Third International Conference on Computing, Communications, and Cyber-Security: IC4S 2021 (pp. 123-136). Singapore: Springer Nature Singapore.

2. Mia, L. (2020). Enhancing Trust and Accountability in Autonomous Cyber Defense Systems Through Explainable AI. Available at SSRN 5140431.
3. Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. IEEe Access, 10, 93104-93139.
4. Mahbooba, B., Timilsina, M., Sahal, R., & Serrano, M. (2021). Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. Complexity, 2021(1), 6634811.
5. Kabir, M. H., Hasan, K. F., Hasan, M. K., & Ansari, K. (2022). Explainable artificial intelligence for smart city application: A secure and trusted platform. In Explainable artificial intelligence for cyber security: next generation artificial intelligence (pp. 241-263). Cham: Springer International Publishing.
6. Agoro, H., & Gray, R. (2021). The Impact of Explainable AI on User Trust in Autonomous Cyber Defense.
7. Hullurappa, M. (2022). The Role of Explainable AI in Building Public Trust: A Study of AI-Driven Public Policy Decisions. International Transactions in Artificial Intelligence, 6.
8. Saeed, Y., & Khan, S. (2022). Comprehensive Cyber Defense: Leveraging AI and Machine Learning in Cloud and Network Infrastructure Protection.
9. Nassar, M., Salah, K., Ur Rehman, M. H., & Svetinovic, D. (2020). Blockchain for explainable and trustworthy artificial intelligence. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(1), e1340.
10. Hayes, B., & Moniz, M. (2020, June). Trustworthy human-centered automation through explainable ai and high-fidelity simulation. In International Conference on Applied Human Factors and Ergonomics (pp. 3-9). Cham: Springer International Publishing.
11. Hernandez, C. S., Ayo, S., & Panagiotakopoulos, D. (2021, October). An explainable artificial intelligence (xAI) framework for improving trust in automated ATM tools. In 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC) (pp. 1-10). IEEE.
12. Flammini, F., Alcaraz, C., Bellini, E., Marrone, S., Lopez, J., & Bondavalli, A. (2022). Towards trustworthy autonomous systems: Taxonomies and future perspectives. IEEE Transactions on Emerging Topics in Computing, 12(2), 601-614.
13. Ejeofobiri, C. K., Adelere, M. A., & Shonubi, J. A. (2022). Developing adaptive cybersecurity architectures using Zero Trust models and AI-powered threat detection algorithms. Int J Comput Appl Technol Res, 11(12), 607-621.
14. He, H., Gray, J., Cangelosi, A., Meng, Q., McGinnity, T. M., & Mehnen, J. (2021). The challenges and opportunities of human-centered AI for trustworthy robots and autonomous systems. IEEE Transactions on Cognitive and Developmental Systems, 14(4), 1398-1412.
15. Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., & Seale, M. (2022). Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities. IEEE Access, 10, 112392-112415.
16. Tiwari, S., Sarma, W., & Srivastava, A. (2022). Integrating artificial intelligence with zero trust architecture: Enhancing adaptive security in modern cyber threat landscape. International Journal of Research and Analytical Reviews, 9, 712-728.
17. Renda, A., Ducange, P., Marcelloni, F., Sabella, D., Filippou, M. C., Nardini, G., ... & Baltar, L. G.

(2022). Federated learning of explainable AI models in 6G systems: Towards secure and automated vehicle networking. Information, 13(8), 395.

18. Oseni, A., Moustafa, N., Creech, G., Sohrabi, N., Strelzoff, A., Tari, Z., & Linkov, I. (2022). An explainable deep learning framework for resilient intrusion detection in IoT-enabled transportation networks. IEEE Transactions on Intelligent Transportation Systems, 24(1), 1000-1014.
19. Procopiou, A., & Chen, T. M. (2021). Explainable ai in machine/deep learning for intrusion detection in intelligent transportation systems for smart cities. In Explainable Artificial Intelligence for Smart Cities (pp. 297-321). CRC Press.
20. Wang, S., Qureshi, M. A., Miralles-Pechuan, L., Huynh-The, T., Gadekallu, T. R., & Liyanage, M. (2021). [20]. Applications of explainable AI for 6G: Technical aspects, use cases, and research challenges. arXiv preprint arXiv:2112.04698.
21. Chamberlain, L. B., Davis, L. E., Stanley, M., & Gattoni, B. R. (2020, May). Automated decision systems for cybersecurity and infrastructure security. In 2020 IEEE Security and Privacy Workshops (SPW) (pp. 196-201). IEEE.
22. Mylrea, M., Nielsen, M., John, J., & Abbaszadeh, M. (2021). Digital twin industrial immune system: AI-driven cybersecurity for critical infrastructures. In Systems Engineering and Artificial Intelligence (pp. 197-212). Cham: Springer International Publishing.
23. Srivastava, G., Jhaveri, R. H., Bhattacharya, S., Pandya, S., Maddikunta, P. K. R., Yenduri, G., ... & Gadekallu, T. R. (2022). XAI for cybersecurity: state of the art, challenges, open issues and future directions. arXiv preprint arXiv:2206.03585.
24. Kanak, A., Ergün, S., Atalay, A. S., Persi, S., & Karcı, A. E. H. (2022, October). A review and strategic approach for the transition towards third-wave trustworthy and explainable ai in connected, cooperative and automated mobility (CCAM). In 2022 27th Asia Pacific Conference on Communications (APCC) (pp. 108-113). IEEE.
25. Gholami, A., Torkzaban, N., & Baras, J. S. (2021). On the importance of trust in next-generation networked cps systems: An ai perspective. arXiv preprint arXiv:2104.07853.
26. Alix, C., Lafond, D., Mattioli, J., De Heer, J., Chattington, M., & Robic, P. O. (2021, June). Empowering adaptive human autonomy collaboration with artificial intelligence. In 2021 16th International Conference of System of Systems Engineering (SoSE) (pp. 126-131). IEEE.
27. Maathuis, C. (2022, June). On the road to designing responsible AI systems in military cyber operations. In European Conference on Cyber Warfare and Security (Vol. 21, No. 1, pp. 170-177).
28. Jagatheesaperumal, S. K., Pham, Q. V., Ruby, R., Yang, Z., Xu, C., & Zhang, Z. (2022). Explainable AI over the Internet of Things (IoT): Overview, state-of-the-art and future directions. IEEE Open Journal of the Communications Society, 3, 2106-2136.
29. Roque, A., & Damodaran, S. K. (2022). Explainable AI for security of human-interactive robots. International Journal of Human–Computer Interaction, 38(18-20), 1789-1807.
30. Ahmed, I., Jeon, G., & Piccialli, F. (2022). From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. IEEE transactions on industrial informatics, 18(8), 5031-5042.
31. Adamson, G. (2020). Explainable Artificial Intelligence (XAI): A reason to believe?. Law Context: A Socio-Legal J., 37, 23.
32. Mohammed, R. (2022). Artificial intelligence-driven robotics for autonomous vehicle navigation and safety. NEXG AI Review of America, 3(1), 21-47.

33. Arisdakessian, S., Wahab, O. A., Mourad, A., Otrok, H., & Guizani, M. (2022). A survey on IoT intrusion detection: Federated learning, game theory, social psychology, and explainable AI as future directions. IEEE Internet of Things Journal, 10(5), 4059-4092.

34. Hussain, Z., & Khan, S. (2022). AI and Cloud Security Synergies: Building Resilient Information and Network Security Ecosystems.

35. Vigano, L., & Magazzeni, D. (2020, September). Explainable security. In 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW) (pp. 293-300). IEEE.

36. Roy, S., Li, J., Pandey, V., & Bai, Y. (2022, August). An explainable deep neural framework for trustworthy network intrusion detection. In 2022 10th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud) (pp. 25-30). IEEE.

37. Sharma, D. K., Mishra, J., Singh, A., Govil, R., Srivastava, G., & Lin, J. C. W. (2022). Explainable artificial intelligence for cybersecurity. Computers and Electrical Engineering, 103, 108356.

38. Embarak, O. (2021). Explainable artificial intelligence for services exchange in smart cities. In Explainable Artificial Intelligence for Smart Cities (pp. 13-30). CRC Press.

39. Veitch, E., & Alsos, O. A. (2021). Human-centered explainable artificial intelligence for marine autonomous surface vehicles. Journal of Marine Science and Engineering, 9(11), 1227.

40. Chennam, K. K., Mudrakola, S., Maheswari, V. U., Aluvalu, R., & Rao, K. G. (2022). Black box models for eXplainable artificial intelligence. In Explainable AI: foundations, methodologies and applications (pp. 1-24). Cham: Springer International Publishing.

41. Mishra, S. (2020). The Age of Explainable AI: Improving trust and transparency in AI models. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 1(4), 41-51.

42. Taj, I., & Zaman, N. (2022). Towards industrial revolution 5.0 and explainable artificial intelligence: Challenges and opportunities. International Journal of Computing and Digital Systems, 12(1), 295-320.

43. Islam, M. U., Mozaharul Mottalib, M., Hassan, M., Alam, Z. I., Zobaed, S. M., & Fazle Rabby, M. (2022). The past, present, and prospective future of xai: A comprehensive review. Explainable Artificial Intelligence for Cyber Security: Next Generation Artificial Intelligence, 1-29.

44. Saraswat, D., Bhattacharya, P., Verma, A., Prasad, V. K., Tanwar, S., Sharma, G., ... & Sharma, R. (2022). Explainable AI for healthcare 5.0: opportunities and challenges. IEEe Access, 10, 84486-84517.

45. Sundaramurthy, S. K., Ravichandran, N., Inaganti, A. C., & Muppalaneni, R. (2022). The future of enterprise automation: Integrating AI in cybersecurity, cloud operations, and workforce analytics. Artificial Intelligence and Machine Learning Review, 3(2), 1-15.

46. Gaur, L., & Sahoo, B. M. (2022). Introduction to explainable AI and intelligent transportation. In Explainable artificial intelligence for intelligent transportation systems: Ethics and applications (pp. 1-25). Cham: Springer International Publishing.

47. Sayed-Mouchaweh, M. (2021). Explainable AI Within the Digital Transformation and Cyber Physical Systems. Springer International Publishing.

48. Tanikonda, A., Pandey, B. K., Peddinti, S. R., & Katragadda, S. R. (2022). Advanced AI-driven cybersecurity solutions for proactive threat detection and response in complex ecosystems. Journal of Science & Technology, 3(1).

49. Hamon, R., Junklewitz, H., & Sanchez, I. (2020). Robustness and explainability of artificial intelligence. Publications Office of the European Union, 207(40).

50. Müller, F., & Volkov, D. (2022). Cybersecurity Risks and Strategic Considerations for AI-Enabled Autonomous Weapons.

51. Barnard, P., Macaluso, I., Marchetti, N., & DaSilva, L. A. (2022, May). Resource reservation in sliced networks: An explainable artificial intelligence (XAI) approach. In ICC 2022-IEEE international conference on communications (pp. 1530-1535). IEEE.

52. Kandregula, N. (2020). Exploring Software-Defined Vehicles: A Comparative Analysis of AI and ML Models for Enhanced Autonomy and Performance.

53. Kandregula, N. (2020). Exploring Software-Defined Vehicles: A Comparative Analysis of AI and ML Models for Enhanced Autonomy and Performance.

54. Hagos, D. H., & Rawat, D. B. (2022). Recent advances in artificial intelligence and tactical autonomy: Current status, challenges, and perspectives. Sensors, 22(24), 9916.

55. Aramide, O. O. (2022). AI-Driven Cybersecurity: The Double-Edged Sword of Automation and Adversarial Threats. International Journal of Humanities and Information Technology, 4(04), 19-38.

56. Sundaramurthy, S. K., Ravichandran, N., Inaganti, A. C., & Muppalaneni, R. (2022). AI-powered operational resilience: Building secure, scalable, and intelligent enterprises. Artificial Intelligence and Machine Learning Review, 3(1), 1-10.

57. Kute, D. V., Pradhan, B., Shukla, N., & Alamri, A. (2021). Deep learning and explainable artificial intelligence techniques applied for detecting money laundering–a critical review. IEEE access, 9, 82300-82317.

58. Karie, N. M., Sahri, N. M. B., Yang, W., & Johnstone, M. N. (2022). Leveraging artificial intelligence capabilities for real-time monitoring of cybersecurity threats. In Explainable Artificial Intelligence for Cyber Security: Next Generation Artificial Intelligence (pp. 141-169). Cham: Springer International Publishing.

59. Luckey, D., Fritz, H., Legatiuk, D., Dragos, K., & Smarsly, K. (2020, July). Artificial intelligence techniques for smart city applications. In International Conference on Computing in Civil and Building Engineering (pp. 3-15). Cham: Springer International Publishing.

60. Kaikova, O., Terziyan, V., Tiihonen, T., Golovianko, M., Gryshko, S., & Titova, L. (2022). Hybrid threats against Industry 4.0: adversarial training of resilience. In E3S Web of Conferences. EDP Sciences.

61. Flournoy, M., Haines, A., & Chefitz, G. (2020). Building trust through testing. Center for Security and Emerging Technology.

62. Nyre-Yu, M., Morris, E. S., Moss, B. C., Smutz, C., & Smith, M. (2021). Considerations for Deploying xAI Tools in the Wild: Lessons Learned from xAI Deployment in a Cybersecurity Operations Setting (No. SAND2021-6069C). Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

63. Dataset link https://www.kaggle.com/datasets/sampadab17/network-intrusion-detection