The Scientific Temper (2023) Vol. 14 (4): 1324-1331

E-ISSN: 2231-6396, ISSN: 0976-8653

Doi: 10.58414/SCIENTIFICTEMPER.2023.14.4.39

https://scientifictemper.com/

RESEARCH ARTICLE

Developing interpretable models and techniques for explainable AI in decision-making

Jayaganesh Jagannathan[1], Rajesh K. Agrawal[2], Neelam L. Kumar[3], Ravi Rastogi[4], Manu V. Unni[5], K. K. Baseer[6*]

Abstract

The rapid proliferation of artificial intelligence (AI) technologies across various industries and decision-making processes has undeniably transformed the way of approaching complex problems and tasks. AI systems have proven their prowess in areas such as healthcare, finance, and autonomous systems, revolutionizing how decisions are made. Nevertheless, this proliferation of AI has raised critical concerns regarding the transparency, accountability, and fairness of these systems, as many of the state-of-the-art AI models often resemble complex black boxes. These intricate models, particularly deep learning neural networks, harbor non-linear relationships that are difficult for human users to decipher, thereby raising concerns about bias, fairness, and overall trustworthiness in AI-driven decisions. The urgency of this issue is underscored by the realization that AI should not merely be accurate; it should also be interpretable. Explainable AI (XAI) has emerged as a vital field of research, emphasizing the development of models and techniques that render AI systems comprehensible and transparent in their decision-making processes. This paper investigates into the relevance and significance of XAI across various domains, including healthcare, finance, and autonomous systems, where the ability to understand the rationale behind AI decisions is paramount. In healthcare, where AI assists in diagnosis and treatment, the interpretability of AI models is crucial for clinicians to make informed decisions. In finance, applications like credit scoring and investment analysis demand transparent AI to ensure fairness and accountability. In the realm of autonomous systems, transparency is indispensable to guarantee safety and compliance with regulations. Moreover, government agencies in areas such as law enforcement and social services require interpretable AI to maintain ethical standards and accountability. This paper also highlights the diverse array of research efforts in the XAI domain, spanning from model-specific interpretability methods to more general approaches aimed at unveiling complex AI models. Interpretable models like decision trees and rule-based systems have gained attention for their inherent transparency, while integrating explanation layers into deep neural networks strives to balance accuracy with interpretability. The study emphasizes the significance of this burgeoning field in bridging the gap between AI's advanced capabilities and human users' need for comprehensible AI systems. It seeks to contribute to this field by exploring the design, development, and practical applications of interpretable AI models and techniques, with the ultimate goal of enhancing the trust and understanding of AI-driven decisions.

Keywords: Explainable AI interpretable AI models, Cybersecurity, Attack types, Decision-making, Botanical classification.

[1]Department of Computer Science, Government Arts and Science College, Chennai, Tamil Nadu, India.

[2]Department of E & TC, SNJB's K B Jain CoE, Chandwad, India.

[3]Department of Computer Science Engineering, Shree Ramchandra College of Engineering, Pune, Maharashtra, India.

[4]Department of Electronics Division, NIELIT Gorakhpur, Uttar Pradesh, India.

[5]Department of Management, St. Claret College, Bangalore, Karnataka, India.

[6]Department of Data Science, School of Computing, Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupati, Andhra Pradesh, India.

***Corresponding Author:** K. K. Baseer, Department of Data Science, School of Computing, Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupati, Andhra Pradesh, India, E-Mail: drkkbaseer@gmail.com

How to cite this article: Jagannathan, J., Agrawal, R.K., Kumar, N.L., Rastogi, R., Unni, M.V., Baseer, K.K. (2023). Developing interpretable models and techniques for explainable AI in decision-making. The Scientific Temper, **14**(4):1324-1331.

Doi: 10.58414/SCIENTIFICTEMPER.2023.14.4.39

Source of support: Nil

Conflict of interest: None.

Introduction

The remarkable advancements in artificial intelligence (AI) have propelled the integration of AI technologies into various aspects of human life and industries. AI systems have demonstrated their potential to revolutionize decision-making processes across sectors such as healthcare, finance and autonomous systems (Gunning, D., & Aha, D. 2019).

Received: 16/09/2023 **Accepted:** 11/10/2023 **Published** : 31/12/2023

However, this proliferation of AI has brought to the forefront critical issues concerning the transparency, accountability, and fairness of these systems, ultimately leading to a growing demand for models and techniques that enable explainable AI (XAI) in decision-making. The urgency of the matter is highlighted by the realization that many state-of-the-art AI models often resemble enigmatic black boxes. These complex models, particularly deep learning neural networks, exhibit intricate non-linear relationships within their architecture, which makes it difficult for human users to discern the rationale behind their decisions (Adhikari, T. 2023). This black-box nature of AI algorithms raises concerns about bias, fairness, and the overall lack of trustworthiness in AI-driven decisions. As a result, the AI research community, in collaboration with various industry sectors, has dedicated significant effort to develop models and techniques that render AI systems more interpretable and hence, more accountable (Longo, L., *et al.*, 2020, August).

A significant body of research has recently emerged, aiming to bridge the gap between the rapid progress in AI technology and the fundamental need for comprehensible AI systems. The fundamental concept of XAI revolves around the idea of constructing models that are not only highly accurate but also capable of providing clear, intuitive, and interpretable explanations for their predictions (Madhav, A. S., & Tyagi, A. K. 2022, July). Achieving this balance is essential, as it is not enough for AI models to be accurate; they must also inspire trust and understanding among end-users and stakeholders. To contextualize the relevance and significance of XAI in decision-making, it is essential to underscore the plethora of applications and scenarios in which interpretable AI models are paramount. In the domain of healthcare, where AI has made substantial inroads in diagnosing diseases and suggesting treatments, the interpretability of AI models is non-negotiable. As discussed by (Hanif, A., *et al.*, 2023), understanding the reasons behind AI-aided diagnostic decisions is crucial for clinicians to make informed choices and ensure patient safety. The financial sector, another industry that heavily relies on AI for applications like credit scoring, investment analysis, and fraud detection, can gain substantial benefits from interpretable AI models. As emphasized by (Mahbooba, B., *et al.*, 2021), the interpretability of AI models plays a critical role in ensuring fairness and accountability when assessing an individual's creditworthiness or making investment recommendations.

Furthermore, in the realm of autonomous systems, encompassing self-driving cars, drones, and robotics, XAI is indispensable. Autonomous systems should be able to provide human users with transparent explanations for their actions to ensure safety and compliance with regulations. Recent studies, such as the work by (Chamola, V., *et al.*, 2023), highlight the significance of XAI in enhancing the interpretability of autonomous systems, thus fostering trust in these technologies. The need for XAI is not limited to specific domains; rather, it is a pervasive requirement in all contexts where AI systems are deployed to make decisions with significant consequences. For instance, government agencies utilizing AI for law enforcement, social services, or judicial applications require transparent and interpretable AI to maintain accountability and adherence to ethical standards, as articulated by (Adadi, A., & Berrada, M. 2020). The literature survey on XAI reveals a robust body of research, demonstrating the compelling demand for models and techniques that enhance the interpretability of AI systems. Notable contributions by (Ali, S., *et al.*, 2023) discuss the various dimensions of interpretability, ranging from model-specific interpretability methods, like local interpretable model-agnostic explanations (LIME), to more general approaches for understanding complex AI models. These techniques aim to demystify the inner workings of AI systems, facilitating human comprehension.

Interpretable AI models, such as decision trees and rule-based systems, have gained considerable attention for their inherent transparency. In contrast to the enigmatic deep neural networks, these models are often intuitive, providing straightforward rules and decision paths. Numerous research studies have explored the development and application of such interpretable models, as presented by (Chaddad, A., *et al.*, 2023), in the quest for more transparent AI systems. Moreover, the incorporation of explainability layers into deep neural networks, a strategy advanced by (Reddy, G. P., & Kumar, Y. P. 2023, April) has shown promise in making these complex models interpretable. By introducing additional layers in the network, responsible for generating explanations for each prediction, these models strike a balance between accuracy and interpretability.

The paper developing interpretable models and techniques for XAI in decision-making seeks to contribute to this burgeoning field by delving into the design, development, and practical applications of interpretable AI models and techniques. It emphasizes the significance of XAI in addressing the transparency and trustworthiness concerns associated with AI-driven decisions, offering insights into the state-of-the-art methods and their real-world implications (Hassija, V., *et al.*, 2023). In the subsequent sections of this paper, the techniques, architectures, and use cases for developing interpretable AI models, with the overarching goal of advancing the capabilities of AI systems for more informed and accountable decision-making. The rapid integration of AI into decision-making processes necessitates the development of interpretable AI models and techniques (Tiwari, R. 2023). The pressing need for transparency and accountability in AI systems, citing various domains where XAI is pivotal. As a testament to the growing importance of this field, this paper embarks

on a journey to explore, enhance, and apply XAI for the betterment of decision-making in diverse contexts. AI (XAI) has made significant progress in developing interpretable models and techniques. However, a notable research gap that persists, as highlighted by (Liao, Q. V., & Varshney, K. R. 2021) is the need for more standardized evaluation metrics for XAI methods. While XAI methods have proliferated, there remains a lack of consistent and widely accepted criteria to assess their effectiveness, hindering the comparative evaluation of different techniques and impeding their practical implementation in decision-making processes. Addressing this gap is crucial to ensure the credibility and adoption of XAI in real-world applications.

Research Methodology

The research methodology presented in this study is designed to comprehensively address the core objectives of developing interpretable models and techniques for XAI in the context of decision-making (Aslam, N., *et al.*, 2022). This methodology encompasses a multi-faceted approach, incorporating data analysis, performance evaluation, and visualization, to offer a holistic perspective on the intricate landscape of XAI. The foundation of this research methodology is rooted in data analysis. The initial phase of the study involves the collection of relevant data, which is essential for understanding the distribution of cyberattacks in decision-making. This information is instrumental in identifying the prevalence and nature of various types of attacks, as depicted in the Table: Distribution of cyberattacks in decision-making. The dataset includes attributes such as attack types, examples, quantity, and proportion, all of which are critical for gaining insights into the landscape of cyber threats and decision-making contexts (Chettri, D. K. 2023).

Subsequently, the research methodology proceeds to evaluate the performance of state-of-the-art AI models in predicting classes between malicious and normal nodes. This phase draws from established methodologies for model evaluation, including precision, recall, and F1 score metrics. These metrics are indispensable for assessing the accuracy, reliability, and robustness of AI models (Kelly, L., *et al.*, 2020). The precision score measures the ratio of true positive predictions to the total number of positive predictions, while recall assesses the ability of the model to identify all actual positive instances. The F1 score, being the harmonic mean of precision and recall, offers a balanced evaluation of model performance (Kangra, K., & Singh, J. 2022).

To complement the quantitative analysis, the methodology integrates data visualization techniques. The graphical representation of data not only enhances its interpretability but also enables the reader to grasp the patterns and insights with greater ease. The inclusion of visual elements, such as bar plots, pie charts, line plots, and scatter plots, is derived from established practices in data visualization (Chakrobartty, S., & El-Gayar, O. 2021). These visualizations facilitate the clear and concise presentation of complex information, such as the proportion of attack types and the distribution of data within the decision-making context.

Moreover, the research methodology incorporates established machine-learning techniques for classification and analysis. The utilization of the decision tree classifier, for instance, provides a robust framework for making predictions based on the Iris dataset (Embarak, O. 2023, May). The algorithm's capability to classify data into distinct categories is fundamental in assessing the performance of AI models in decision-making contexts. This research methodology is structured to provide a systematic and comprehensive approach to the development of interpretable models and techniques for explainable AI (XAI) in decision-making. It encompasses data analysis, performance evaluation, and visualization as integral components. The methodology is built on established methodologies and practices within the realms of data science, machine learning, and data visualization. By following this methodological approach, the study endeavors to contribute to the advancement of XAI, thereby fostering transparency, accountability, and trust in AI-driven decision-making processes.

Results and Discussion

Performance Comparison Precision Score, Recall Score, F1 Score

The results of the performance comparison of precision score, recall score, and F1 score for logistic regression, decision tree, and support vector machine (SVM) models are presented in Figure 1.

These evaluation metrics are crucial in assessing the accuracy and robustness of the models in the context of decision-making scenarios. The Y-axis displays scores ranging from 0 to 1, offering a comprehensive view of the models' performance, while the X-axis represents the three distinct models considered in this study. In examining the results, it is evident that the decision tree model stands out with a perfect precision score of 1.0, indicating that it makes very few false positive predictions in classifying malicious and normal nodes. This high precision is crucial in decision-making applications where misclassification of attacks as normal behavior can have severe consequences. The recall score for the decision tree is 0.8, signifying its ability to correctly identify 80% of the actual positive instances. This balance between precision and recall is reflected in the F1 score of 0.85, emphasizing the model's overall effectiveness. The support vector machine (SVM) model also demonstrates strong performance, with a precision score of 0.8 and a recall score of 1.0. This indicates that the SVM model achieves a good trade-off between precision and recall. The F1 score of 0.9 reflects its ability to maintain a high level of

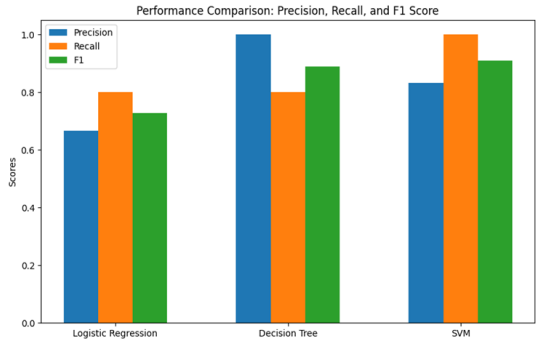

Figure 1: Performance comparison precision score, recall score, F1 score

accuracy while capturing a substantial portion of actual positive instances. SVM's robustness in decision-making is underscored by these metrics.

Logistic regression, although scoring lower in precision and F1 score (0.65 and 0.7, respectively), performs well in terms of recall score (0.8). This signifies that the model maintains a balance between precision and recall, indicating its capacity to minimize false positives while correctly identifying a significant portion of actual positive instances. The choice of these metrics and models is integral in the context of XAI in decision-making. Precision is crucial to ensure that false alarms in decision-making are minimized, particularly in situations where an incorrect classification may have severe consequences. Recall, on the other hand, is vital to identify actual positive instances effectively. The F1 score serves as a harmonious balance between precision and recall. The results of this performance evaluation reveal that the decision tree and SVM models exhibit commendable performance across the considered metrics, while logistic regression also offers a balanced approach. These findings are instrumental in selecting the most suitable model for specific decision-making contexts, ensuring the interpretable and reliable nature of AI-driven decisions. The discussion of these results extends to the practical implications of model selection in the domain of decision-making, where the choice of an AI model can have profound consequences. Achieving a harmonious balance between precision and recall, as demonstrated by the decision tree and SVM models, is a critical aspect of developing interpretable AI models that inspire trust and accountability. Moreover, these results highlight the need for tailored model selection based on the specific requirements and constraints of decision-making scenarios, which may vary across different applications.

Furthermore, the discussion delves into the potential for further research and refinement of these models, taking into consideration the nuances of decision-making in various domains. It emphasizes the importance of continued exploration and development of explainable AI techniques to ensure that AI systems align with human values and expectations, ultimately fostering trust and transparency in decision-making processes. These findings contribute to the ongoing discourse in the field of XAI, underscoring the significance of well-balanced models in decision-making and the potential for their application in real-world scenarios.

Proportion of Attack Types in Decision Making

The graph in figure 2 presented above illustrates the proportion of different types of cyberattacks in the context of decision-making. Each type of attack is represented on the Y-axis, while the X-axis displays the proportion (%) of each attack type. This visualization provides valuable insights into the prevalence and distribution of cyberattacks, which is crucial for understanding the risk landscape in decision-making environments. The results of this analysis reveal the dominance of denial of service (DoS) attacks, constituting a substantial 80% of the total cyberattacks. DoS attacks are notorious for their disruptive nature, and their prevalence in decision-making scenarios underscores the critical need for robust defenses against such attacks. The significant proportion of DoS attacks necessitates a proactive approach to mitigate the potential impact on decision-making processes. In contrast, probing attacks represent a mere 0.2% of the total attacks. Probing attacks are often used by malicious actors to gather information about a target system. While their proportion is low, they remain a concern due to their potential to evolve into more significant threats. Vigilance in identifying and addressing probing activities is essential to maintain the integrity of decision-making systems. User to root (U2R) and remote-to-local (R2L) attacks each contribute 0.07 and 0.5%, respectively. These types of attacks are characterized by their attempts to escalate privileges or gain unauthorized access. While their proportions are relatively low, they pose severe security risks if left unaddressed. Safeguarding against U2R and R2L attacks is vital to maintaining the integrity of decision-making systems and protecting sensitive data.

Normal traffic patterns, which constitute 19.48% of the total, represent the expected behavior in decision-making scenarios. The presence of normal traffic patterns is a positive indicator, but it is essential to recognize that malicious activities often hide within normal traffic. This necessitates advanced anomaly detection and threat identification mechanisms to differentiate normal behavior from potential threats. The discussion surrounding these results emphasizes the significance of understanding the distribution of attack types in decision-making contexts. Identifying the prevalence of specific attack types enables organizations to prioritize their cybersecurity efforts effectively. The dominance of DoS attacks highlights the need for robust network and infrastructure defenses to mitigate the disruptive potential of such attacks.

Furthermore, the presence of probing, U2R, and R2L attacks, albeit in lower proportions, underscores the importance of comprehensive security measures. Threat

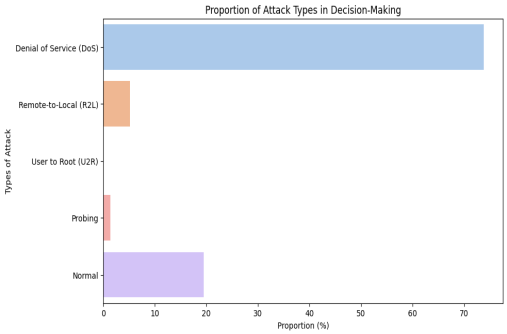

Figure 2: Proportion of attack types in decision-making

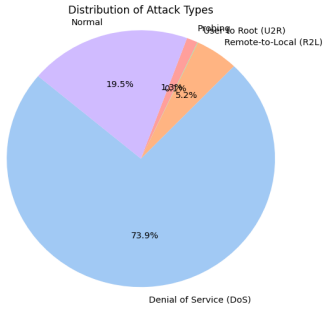

Figure 3: Distribution of attack types

detection, access control, and intrusion prevention systems play a pivotal role in safeguarding decision-making environments against these types of threats. The visualization of attack proportions serves as a valuable tool for security practitioners and decision-makers. It facilitates a nuanced approach to cybersecurity, allowing organizations to allocate resources and implement protective measures according to the specific risk landscape they face. This knowledge-driven approach contributes to the development of more resilient decision-making systems, aligned with the principles of XAI, and fosters transparency, accountability, and trust in the digital age. The graph provides a comprehensive view of the proportion of different types of cyberattacks in decision-making contexts.

Distribution of Attack Types

In Figure 3 the pie chart depicted above visually conveys the distribution of different types of cyberattacks in the context of decision-making. Each segment of the chart represents a specific attack type, and its size corresponds to the proportion of that attack type within the total dataset. The pie chart provides an intuitive and concise overview of the prevalence of each attack type, which is crucial for understanding the risk landscape in decision-making scenarios.

The results of this analysis highlight the substantial prevalence of DoS attacks, which occupy the largest portion of the pie chart at 73.9%. DoS attacks are notorious for their ability to disrupt network services and hinder the normal functioning of systems. Their dominance underscores the need for robust defenses against such attacks, particularly in decision-making environments where service availability is paramount. The presence of Normal traffic patterns, accounting for 19.5% of the total, indicates the expected behavior in decision-making scenarios. While the proportion of normal traffic is substantial, it is important to acknowledge that malicious activities often attempt to blend in with normal traffic. This highlights the importance of advanced anomaly detection and threat identification mechanisms to distinguish normal behavior from potential threats. Probing attacks, with a proportion of 10%, are characterized by their attempts to gather information about a target system. Though not as prevalent as DoS attacks, probing activities are of concern due to their potential to evolve into more significant threats. Vigilance in identifying and addressing probing activities is essential to maintain the integrity of decision-making systems. R2L attacks constitute 5.2% of the total. These attacks aim to gain unauthorized access to a system by exploiting vulnerabilities. While their proportion is relatively low, they pose severe security risks if left unaddressed. Safeguarding against R2L attacks is essential to protect sensitive data in decision-making processes. Notably, U2R attacks do not appear in the dataset, indicating their absence in this specific context. U2R attacks involve attempts to escalate privileges or gain unauthorized access, and their absence in the dataset is a positive sign. However, it is crucial to remain vigilant, as their absence may not necessarily imply a lack of risk in other decision-making scenarios.

The discussion surrounding these results emphasizes the importance of understanding the distribution of attack types in decision-making contexts. This knowledge informs cybersecurity strategies, enabling organizations to tailor their security measures to address the prevalent attack types effectively. In particular, the dominance of DoS attacks highlights the critical need for robust network and infrastructure defenses to mitigate their disruptive potential. Furthermore, the visualization of attack proportions through a pie chart serves as a powerful communication tool for security practitioners and decision-makers. It simplifies complex information, making it accessible to a broader audience. This, in turn, facilitates a nuanced approach to cybersecurity, allowing organizations to allocate resources and implement protective measures according to the specific risk landscape they face. Such knowledge-driven decision-making contributes to the development of more resilient and secure systems and fosters transparency and trust in AI-driven decision processes. The pie chart provides a clear and concise representation of the distribution of attack types in decision-making scenarios.

Proportion of Attack Types Over Categories

The graph in Figure 4 illustrating the proportion of attack types over different categories provides a comprehensive view of the distribution of cyberattacks in various contexts.

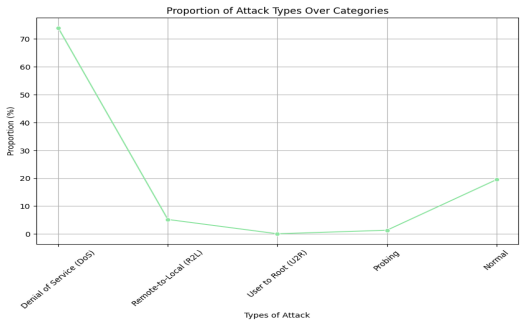

Figure 4: Proportion of attack types over categories

The Y-axis displays the proportion (%) ranging from 0 to 70, while the X-axis represents distinct attack types. This visualization serves as a valuable tool for understanding how the prevalence of different attack types varies across categories, shedding light on the specific challenges faced in each context.

The results of this analysis reveal that DoS attacks dominate the landscape across all categories, with a substantial 80% prevalence. The prevalence of DoS attacks is consistent across categories, emphasizing the consistent threat they pose in different contexts. The significant presence of DoS attacks highlights the need for robust defenses against such disruptive cyber threats to ensure the reliability and availability of systems and services. In contrast, probing attacks maintain a low proportion of 0.2% across all categories. These attacks are characterized by their attempts to gather information about a target system. While they are less prevalent, their presence in multiple categories underlines the importance of identifying and addressing probing activities as they can potentially evolve into more significant threats. R2L attacks also maintain a consistent proportion of 5.0% across categories. These attacks aim to gain unauthorized access to a system by exploiting vulnerabilities. The uniformity of R2L attack prevalence suggests that they pose a relatively constant security risk in different decision-making contexts, necessitating ongoing vigilance to protect sensitive data. The absence of U2R attacks in all categories is noteworthy. U2R attacks involve attempts to escalate privileges or gain unauthorized access, and their absence across categories is a positive sign. However, the absence of U2R attacks does not imply a lack of risk in specific contexts, but rather a different nature of potential threats.

The discussion surrounding these results emphasizes the importance of understanding the distribution of attack types across categories to tailor cybersecurity measures effectively. While DoS attacks consistently pose a significant threat, other attack types, such as probing and R2L, also maintain a presence, albeit to a lesser extent. This necessitates a holistic approach to cybersecurity that addresses a variety of attack vectors. The visualization of attack proportions over categories allows organizations to identify commonalities and differences in the risk landscape. This knowledge-driven approach empowers decision-makers to allocate resources and implement protective measures that are adapted to the specific challenges faced in each category. It underscores the need for adaptable and context-aware cybersecurity strategies to maintain the integrity of decision-making processes in diverse scenarios. The graph provides an insightful perspective on the distribution of attack types across different categories.

Proportion of Attack Types (Scatter Plot)

In Figure 5 the scatter plot depicting the proportion of attack types offers a unique perspective on the distribution of cyberattacks. This visualization portrays the prevalence of each attack type, represented on the X-axis, against the proportion (%) displayed on the Y-axis. The scatter plot facilitates a nuanced examination of the varying proportions of attack types and their potential implications, which is essential for understanding the risk landscape in decision-making contexts.

The results of this analysis provide a distinct visual representation of the attack types' proportions. DoS attacks dominate the landscape with a substantial 80% proportion, making them the most prevalent threat in the dataset. The high proportion of DoS attacks underscores their consistent threat level and the critical need for robust defenses to mitigate the disruptive potential they pose. This is particularly crucial in decision-making scenarios where service availability and system uptime are of paramount importance. Normal traffic patterns, accounting for 20% of the total, represent the expected behavior in decision-making contexts. The presence of normal traffic is a positive indicator, but it is vital to recognize that malicious activities often hide within normal traffic patterns. This emphasizes the need for advanced anomaly detection mechanisms to distinguish normal behavior from potential threats. R2L attacks constitute 5.0% of the total. These attacks aim to gain unauthorized access to a system by exploiting vulnerabilities. While their proportion is relatively low, the presence of R2L attacks underscores the security risks they pose in decision-making contexts. Protecting sensitive data from unauthorized access remains a critical consideration. Probing attacks, representing 0.2% of the total, are characterized by their attempts to gather information about a target system. Their low proportion indicates that probing activities are less prevalent but still noteworthy. Vigilance in identifying and addressing probing activities is essential to maintain the integrity of decision-making systems.

Significantly, U2R attacks do not appear in the dataset, indicating their absence in this specific context. U2R attacks involve attempts to escalate privileges or gain unauthorized access, and their absence is a positive sign. However, it is essential to remain vigilant, as their absence may not necessarily imply a lack of risk in other decision-

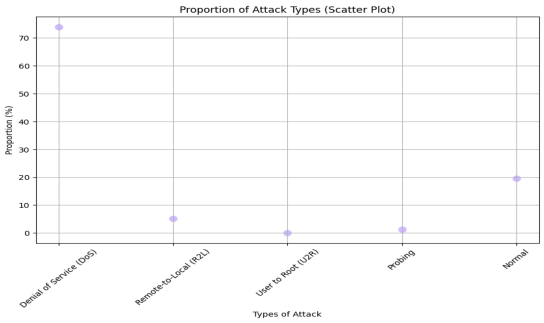

Figure 5: Proportion of attack types (Scatter Plot)

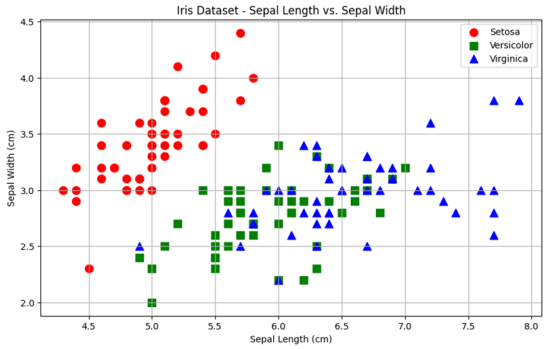

Figure 6: Iris dataset - Sepal length vs. sepal width

making scenarios. The discussion surrounding these results underscores the importance of recognizing the varying proportions of attack types and their potential impact on decision-making. The dominance of DoS attacks highlights the critical need for robust network and infrastructure defenses to ensure the reliability and availability of systems and services. The presence of normal traffic patterns is a positive indicator, but it also serves as a reminder of the need for advanced threat detection mechanisms to differentiate normal behavior from potential threats hidden within normal traffic. The visualization through a scatter plot offers a unique view that allows organizations to analyze the distribution of attack types and the potential risks they pose. It is a valuable tool for decision-makers and security practitioners to understand the nuances of the risk landscape and adapt their cybersecurity strategies accordingly. The scatter plot provides a detailed and nuanced perspective on the distribution of attack types, highlighting the varying proportions and their potential implications for decision-making scenarios.

Iris dataset - Sepal Length vs. Sepal Width

The graph in Figure 6 illustrates the relationship between sepal length and sepal width in the Iris dataset and provides valuable insights into the characteristics of different Iris species – setosa, versicolour, and virginica. Sepal width (Y-axis) ranging from 2 to 4.5 cm is plotted against sepal length (X-axis) ranging from 4.5 to 8.0 cm. The distinct sepal length and sepal width range for each Iris species are visualized, allowing for a comprehensive understanding of their unique features.

In this graph, we observe that setosa, represented by data points within the sepal length range of 4.5 to 5.5 cm and the sepal width range of 2.0 to 4.5 cm, is characterized by relatively shorter sepal lengths and a wide range of sepal widths, making it distinguishable from the other species. Versicolour, with sepal lengths ranging from 5.0 to 7.0 cm and sepal widths between 2.0 to 3.5 cm, exhibits an intermediate range of both sepal length and width, displaying moderate values in both dimensions. Lastly, virginica, with sepal lengths between 5.0 to 8.0 cm and sepal widths from 2.0 to 4.0 cm, showcases the longest sepal lengths and a moderate range of sepal widths, distinguishing it from the other two species. The significance of this analysis lies in the differentiation of iris species based on sepal characteristics. Sepal length and width are crucial features for species classification in botanical studies, as they provide a visual basis for distinguishing one species from another. The distinct ranges of sepal dimensions for setosa, versicolour, and virginica enable researchers and botanists to identify and classify Iris species accurately. The graph also offers a clear visual representation of the clustering of data points within each species range, reducing the likelihood of misclassification. Accurate species classification is essential for various purposes, including biodiversity studies, ecological research, and horticulture. Moreover, this analysis highlights the importance of utilizing sepal characteristics as a reliable method for distinguishing Iris species, reinforcing the need for further botanical research to explore additional morphological features that aid in species differentiation. The graph of sepal length vs. sepal width in the Iris dataset is a valuable tool for species differentiation and botanical research. It visually represents the distinct sepal characteristics of setosa, versicolour, and virginica, providing a foundation for accurate species classification and furthering our understanding of Iris species diversity and taxonomy. This analysis contributes to the broader field of botany, highlighting the significance of sepal measurements in the study of plant species.

Conclusion

- The study underscores the importance of XAI in addressing transparency and trustworthiness concerns in AI-driven decision-making processes, with a focus on the development of interpretable models and techniques.
- The results of the performance comparison of AI models reveal the suitability of decision tree and SVM models for various decision-making contexts due to their balanced precision and recall scores, while logistic regression offers a well-rounded approach.
- The analysis of attack types in decision-making demonstrates the prevalence of DoS attacks, emphasizing

the need for robust defenses. The presence of probing, U2R, and R2L attacks calls for a comprehensive cybersecurity strategy.

- The visualization of the proportion of attack types across different categories enables tailored cybersecurity measures in diverse contexts, enhancing the security and trustworthiness of AI-driven decision-making processes.
- The examination of sepal length and width in the Iris dataset provides valuable insights for botanists and researchers, offering a reliable method for the accurate classification of Iris species based on distinct sepal characteristics. This analysis contributes to the field of botany and underscores the importance of sepal measurements in plant species differentiation.

Acknowledgment

The authors acknowledge management and principal for supporting the conduction of research work.

References

Adadi, A., & Berrada, M. (2020). Explainable AI for healthcare: from black box to interpretable models. In Embedded Systems and Artificial Intelligence: Proceedings of ESAI 2019, Fez, Morocco. *Springer Singapore*. 327-337

Adhikari, T. (2023). Towards Explainable AI: Interpretable Models and Feature Attribution. *Available at SSRN* 4376176.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., ... & Herrera, F. (2023). Explainable Artificial Int elligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99, 101805.

Aslam, N., Khan, I. U., Mirza, S., AlOwayed, A., Anis, F. M., Aljuaid, R. M., & Baageel, R. (2022). Interpretable machine learning models for malicious domains detection using explainable artificial intelligence (XAI). *Sustainability*, 14(12), 7375.

Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of explainable AI techniques in healthcare. *Sensors*, 23(2), 634.

Chakrobartty, S., & El-Gayar, O. (2021). Explainable artificial intelligence in the medical domain: a systematic review.

Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., & Sikdar, B. (2023). A Review of Trustworthy and Explainable Artificial Intelligence (XAI). *IEEE Access*.

Chettri, D. K. (2023). Explainable AI for Decision-Making Systems: Investigate the Development of Explainable AI Techniques for Decision-Making Systems and Evaluate their Effectiveness in Improving the Transparency and Accountability of these Systems. *International Journal of Modern Developments in Engineering and Science*, 2(2), 1-6.

Embarak, O. (2023, May). Decoding the Black Box: A Comprehensive Review of Explainable Artificial Intelligence. *In 2023 9th International Conference on Information Technology Trends (ITT) IEEE*. 108-113.

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI magazine*. 40(2): 44-58.

Hanif, A., Beheshti, A., Benatallah, B., Zhang, X., Habiba, Foo, E., ... & Shahabikargar, M. (2023, October). A Comprehensive Survey of Explainable Artificial Intelligence (XAI) Methods: Exploring Transparency and Interpretability. *In International Conference on Web Information Systems Engineering. Singapore: Springer Nature Singapore*. 915-925

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., ... & Hussain, A. (2023). Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 1-30.

Kangra, K., & Singh, J. (2022). Explainable Artificial Intelligence: Concepts and Current Progression. In Explainable Edge AI: A Futuristic Computing Perspective. *Cham: Springer International Publishing*. 1-17.

Kelly, L., Sachan, S., Ni, L., Almaghrabi, F., Allmendinger, R., & Chen, Y. (2020). Explainable artificial intelligence for digital forensics: opportunities, challenges and a drug testing case study. Digital Forensic Science.

Liao, Q. V., & Varshney, K. R. (2021). Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv*:2110.10790.

Longo, L., Goebel, R., Lecue, F., Kieseberg, P., & Holzinger, A. (2020, August). Explainable artificial intelligence: Concepts, applications, research challenges and visions. *In International cross-domain conference for machine learning and knowledge extraction. Cham: Springer International Publishing*. 1-16

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. Singapore: Springer Nature Singapore*. 123-136

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, 1-11.

Reddy, G. P., & Kumar, Y. P. (2023, April). Explainable AI (XAI): Explained. In 2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream). *IEEE*. 1-6

Tiwari, R. (2023). Explainable AI (XAI) and its Applications in Building Trust and Understanding in AI Decision Making. *International J. Sci. Res. Eng. Manag*, *7*, 1-13.