Feature Effect Visualization in Cybersecurity: A Study of PDP and ICE

Clarence Bostic*
clarence.bostic@my.hamptonu.edu
Computer Science Department, Hampton University
Hampton, Virginia, USA

Janett Walters-Williams†
janett.williams@hamptonu.edu
Computer Science Department, Hampton University
Hampton, Virginia, USA

ABSTRACT

The integration of Artificial Intelligence (AI) into cybersecurity has significantly developed advanced threat detection and analysis. However, due to the deep learning nature, the inherent opacity that comes with these "black box" models creates doubts in the decisions of incident investigation. Explainable Artificial Intelligence (XAI) is the backbone of the future of this transparency gap, by utilizing visualization tools to make these decisions more interpretable. This paper examines two feature-visualization tools in cybersecurity: Partial Dependence Plots (PDPs) and Individual Conditional Expectation (ICE) plots. We analyse the differences between PDPs, which are global explanations, by averaging the effects, and ICE plots, which offer local instance-level insights, to show heterogeneous attack patterns. By evaluating these methods with the focus of improving cybersecurity intrusion detection and malware analysis. This study highlights the necessary balance between clarity and depth to enhance operational reliability in AI-driven security systems.

CCS CONCEPTS

• **Computing methodologies** → *Reasoning about belief and knowledge*; **Feature selection**; *Information extraction*; Probabilistic reasoning.

KEYWORDS

Explainable Artificial Intelligence, Partial Dependence Plot (PDP), Individual Conditional Expectation (ICE), Cybersecurity

ACM Reference Format:
Clarence Bostic and Janett Walters-Williams. 2026. Feature Effect Visualization in Cybersecurity: A Study of PDP and ICE. In *Proceedings of ACM Symposium 2026, March 26–29, 2026, Orangeburg, SC*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Artificial Intelligence (AI) has become deeply embedded in different sectors, including cybersecurity, healthcare, finance, engineering, and public policy. The introduction of advanced Deep Learning (DL) models has significantly improved capabilities such as threat detection, anomaly analysis, automated response, and real-time decision-making. However, while these systems achieve high predictive accuracy, their internal reasoning processes are often opaque—even to experts, creating a major challenge called the "black box" problem that creates a critical dilemma: *"AI is trusted to make critical decisions, yet the logic behind those decisions is difficult to understand, validate, audit, or regulate"* [10].

In cybersecurity, particularly within critical infrastructure environments, this opacity introduces serious risks. Analysts may see that activity was flagged as malicious, but may not understand why. This weakens trust in automated defences, complicates incident investigation, obscures accountability, and makes it harder to detect bias, data poisoning, adversarial manipulation, or model drift after updates. Unlike traditional rule-based "white-box" systems, AI-driven defences often trade interpretability for performance, creating tension between efficiency and the non-negotiable requirements of reliability, auditability, and compliance.

In response, the field of Explainable Artificial Intelligence (XAI) has emerged to address this transparency gap. XAI seeks to "open the black box" by making AI decisions understandable without sacrificing performance. Its objectives are both technical and ethical: building stakeholder trust, enabling debugging and validation, detecting and mitigating bias, strengthening accountability, and satisfying regulatory demands for explainability.

Major initiatives, such as DARPA's XAI program and NIST's explainability principles, underscore the growing recognition that AI systems must be not only powerful, but also understandable, trustworthy, and accountable [3]. As AI continues to expand into critical real-world applications, explainability is no longer optional—it is essential to ensure that advanced systems strengthen security and decision-making rather than introduce new vulnerabilities.

2 UNDERSTANDING XAI

XAI has become the backbone for improving understanding, trust, and operational reliability in AI-driven cybersecurity systems. In the ever-increasing use of machine learning models in cybersecurity, these models have grown to reach all aspects of cybersecurity; intrusion detection, automated incident response, but the opacity of these models limits cyber-analysts' confidence in just "accepting" the content produced, making analysts need to check over twice on the output just for confidence. Research in a cybersecurity contexts emphasizes that explainability over an machine learning output can improve human understanding of alerts, it would support validation of a model behaviour, and enhances adoption in real-world security operations centers. [4, 5]. Within XAI, there are numerous

*Undergraduate student
†Student's advisor

Conference acronym 'XX, March 26–29, 2026, Orangeburg, SC

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

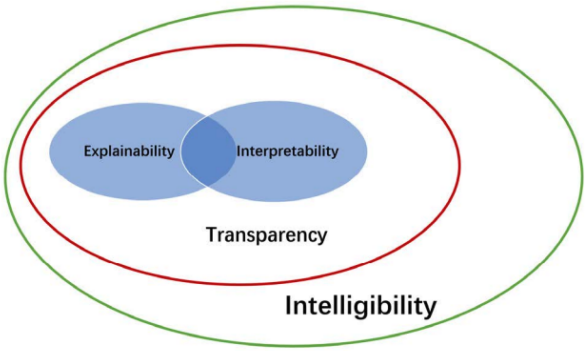

Figure 1: A Venn Diagram showing the connections between words in XAI domains [12] .

concepts and phrase that are used to characterize XAI: intelligibility, explainability, transparency and interpretability. Intelligibility and Explainability are explained similarly as concepts, and on **Figure 1**, you can see the relationship between these terms.

In most recent years though, the definition for interpretability has changed into an term information extraction, not necessarily providing explanations [12]. Explainability on the other hand, has not changed; it is the ability for an AI system to explain the internal decisions and decision-making, making them more human-digestible. Transparency refers to the degree to which users can see and understand how a model processes data and reaches decisions, including insight into its structure, logic, and feature influence. Interpretability focuses on how users understand the meaning of a model's decisions and outputs. Together, these principles aim to reduce the trade-off between predictive accuracy and explainability that characterizes many modern AI systems. In 2020, the National Institute of Standards and Technology (NIST) formalized the need for XAI by introducing four core principles: **Explanation** (systems must provide evidence or reasoning for outputs), **Meaningful** (explanations must be understandable to the intended user), **Explanation Accuracy** (explanations must faithfully reflect the system's actual processes), and **Knowledge Limits** (systems must recognize when outputs are unreliable or outside their design scope). These principles reinforce the role of XAI in cybersecurity governance and accountability[4].

2.1 XAI Methods

Over time, a comprehensive taxonomy of XAI methods has emerged, particularly within cybersecurity applications. A central distinction exists between true transparency (intrinsic interpretability)—models that are understandable by design, such as decision trees, rule-based systems, and linear models—and post-hoc explainability, which applies additional techniques to clarify complex black-box models after training. These methods may be **model-specific**, designed for a particular architecture, or **model-agnostic**, applicable across different models without access to internal parameters. They can also be categorized as **intrinsic** (interpretable by structure) or **extrinsic/post-hoc** (requiring explanation tools after training), and further divided into **local explanations**, which clarify individual predictions, or **global explanations**, which describe overall model behaviour. Common post-hoc techniques in cybersecurity include LIME, SHAP, global surrogate models, feature attribution methods, and counterfactual explanations.

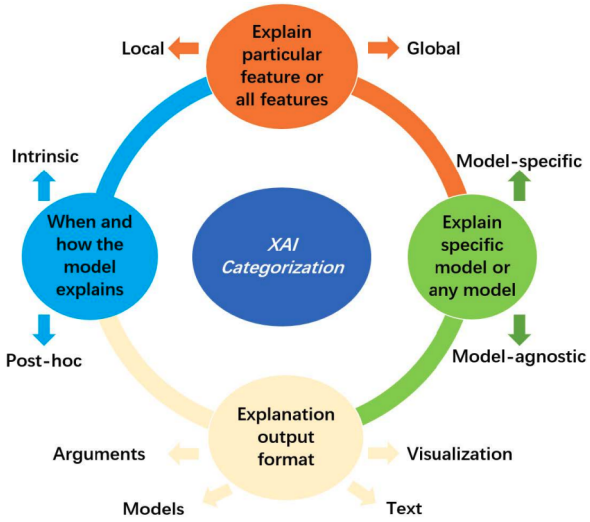

Figure 2: An overview diagram showing the categorization of XAI in different aspects [12] .

2.2 Interpretability Differences

Explanation methods can also be grouped by the different outcomes of interpretability, which help differentiate how these techniques provide insight into model behaviour [6].

The most common one being **feature explainablity**, feature explainbility can come from many different methods but namely Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations(SHAP), the goal of this method is to describe decision in text form to help describe decisions telling you what exactly is output and how it got to the conclusion.

Another category and the focus of this paper is **feature summary statistics** (global, model-agnostic). These types of techniques generate quantitative summaries of how each feature affects model predictions, including feature importance measures and statistics that capture interaction strength between variables. Another important category is **feature summary visualizations**(model-agnostic), which provide qualitative insights that may not be meaningfully represented in tabular form. These visual tools aim to reveal what a model has learned. Included in this category are Partial Dependence Plots (PDPs) that illustrate the marginal effect of a feature on the predicted outcome for both linear and non-linear relationships. PDP variants include Individual Conditional Expectation (ICE) plots, which show prediction effects at the individual instance level, and Accumulated Local Effects (ALE) plots, which offer more reliable interpretations when features are correlated.

Together, these structured approaches enhance transparency and interpretability in AI-driven cybersecurity systems by making complex model behaviour more understandable and actionable.

2.3 XAI Motivations

Beyond methodological classification, these approaches support four widely recognized motivations for XAI: explain to justify (providing defensible reasoning behind security decisions), explain to control (enabling oversight, auditing, and governance), explain to improve (facilitating debugging, bias detection, and model refinement), and explain to discover (revealing new threat patterns or hidden feature relationships) [1]. Together, these structured methods aim to reduce the trade-off between predictive accuracy and explainability, ensuring that AI-powered cybersecurity systems remain not only high-performing but also accountable, auditable, controllable, and strategically trustworthy.

3 CHOSEN METHODS

3.1 Partial Dependence Plots (PDPs)

PDPs are widely used feature-effect visualization techniques within XAI, particularly in cybersecurity domains such as intrusion detection, malware analysis, and IoT threat monitoring. They provide a global, model-agnostic interpretation by estimating the marginal effect of one (or two) features on a model's predictions while averaging over the joint distribution of all other variables. In cybersecurity applications, this enables analysts to understand overall behavioural trends learned by a model. For example, how increasing failed login attempts, unusual port activity, or abnormal packet rates influence predicted compromise risk. As a result, PDPs are frequently used to support model validation, documentation, and governance efforts, helping justify automated alerts and improve transparency in security operations [2, 5, 9].

PDPs are widely used in cybersecurity as a global, model-agnostic interpretability technique for understanding how specific features influence machine learning predictions in applications such as intrusion detection, malware analysis, and IoT security monitoring. By estimating the marginal effect of key variables—such as packet rate, login attempts, or connection duration—on predicted threat likelihood, PDPs provide high-level transparency into model behavior and support analyst validation and governance efforts [2, 9]. However, because cybersecurity data often contains correlated features and complex attack interactions, PDPs may obscure subgroup-specific dynamics due to their averaging mechanism [5]. Despite this limitation, PDPs remain valuable for global model interpretation and strengthening transparency in AI-driven cybersecurity systems [7].

3.2 Individual Conditional Expectation (ICE)

ICE plots are a local, model-agnostic interpretability method used in XAI to understand how individual data instances influence a model's predictions. These plots generate a separate curve for each observation, illustrating how that specific instance's predicted outcome changes as a selected feature varies while all other features remain fixed [5]. This instance-level visualization enables analysts to examine prediction sensitivity and uncover heterogeneity that global summary methods may mask. In cybersecurity research, ICE plots are increasingly integrated into explainable intrusion detection and malware analysis frameworks to enhance transparency and operational trust. This means that ICE shows whether the model behaves consistently across cases—or whether it is driven by a subset of high-leverage instances that may correspond to a particular campaign, environment, or artifact of the dataset [2, 11]

In cybersecurity applications—such as intrusion detection, malware classification, and IoT security monitoring—ICE plots are particularly valuable because attack behaviors are often diverse and context-dependent. Different network segments, device types, or attack campaigns may exhibit distinct response patterns to the same feature. ICE enables analysts to visualize how individual sessions or threat instances respond to variables such as packet frequency, connection duration, or authentication attempts, thereby revealing subgroup-specific model behavior [9] Contemporary XAI surveys further emphasize that local explanation techniques like ICE improve debugging, drift detection, and model validation in AI-driven cybersecurity systems, strengthening reliability and accountability in operational environments [7].

4 EXAMINING PDP AND ICE PERFORMANCES

4.1 Global versus Local Explanations

PDPs and ICE plots differ fundamentally in their scope of interpretation, and this distinction has important implications for how each model's behaviour is understood.

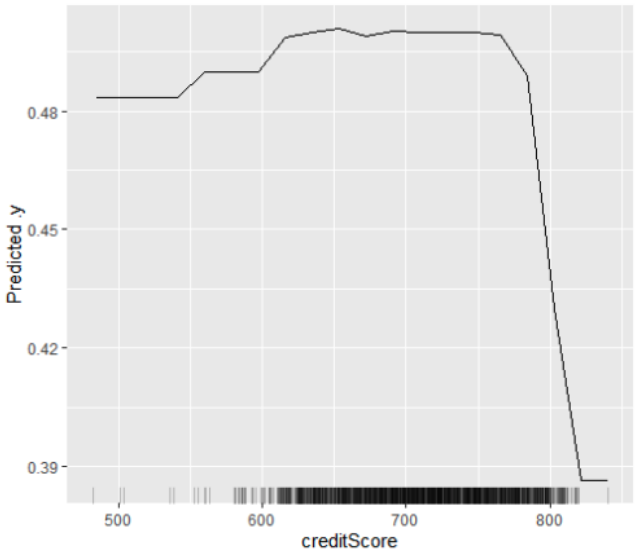

Figure 3: Partial Dependence Plot[8] .

4.1.1 PDPs. are best at providing a more general understanding provide a **global explanation** by estimating the average marginal effect of a feature on model predictions across the entire dataset. This means that the resulting curve reflects the overall trend the

model has learned, smoothing out individual variability. The advantage of this global view is clarity: PDPs offer a stable, easy-to-communicate summary of how a feature generally influences predictions, which is particularly useful for governance, reporting, and high-level validation. However, this averaging process can conceal heterogeneity. If different subgroups in the data respond differently to the same feature, PDPs may mask those variations, potentially leading to overconfident or oversimplified interpretations [7]. For example, **Figure 3** uses a PD plot to show the effect of a credit score on predictions.

4.1.2 ICE. plots, on the other hand, provide a **local explanation** by generating a separate curve for each instance, showing how that specific observation's prediction changes as a feature varies. This instance-level granularity allows analysts to detect whether the model behaves consistently across the population or whether distinct sub-patterns exist. The impact of this local perspective is significant: ICE can reveal interaction effects, subgroup behaviours, and hidden model sensitivities that global summaries overlook. For example, in cybersecurity contexts, different attack types may exhibit different response patterns to the same network feature—variations that a PDP would average away but ICE would clearly expose [11]. However, ICE also has limitations. It can only display one feature at a time, making multi-feature interaction analysis less straightforward without additional methods. Furthermore, similar to PDPs, ICE may produce invalid or unrealistic data points when the feature of interest is strongly correlated with another feature, since altering one feature independently can create combinations that do not naturally occur in the dataset [7]. Additionally, ICE plots can become visually cluttered when many instances are displayed, reducing interpretability without aggregation techniques. Below in **Figure 4** is an example of the clutter, as seen, which creates hard readability for the readers.

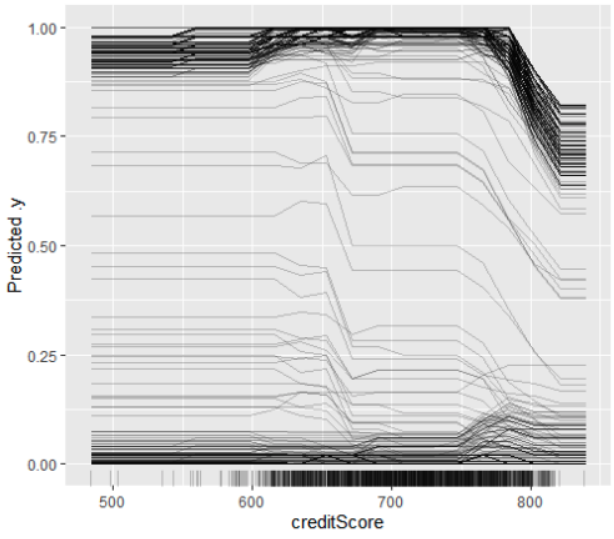

Figure 4: Partial Dependence Plot[8] .

4.1.3 Comparison: Overall, the global nature of PDPs makes them effective for summarizing broad model behaviour and supporting strategic oversight, while the local nature of ICE provides deeper diagnostic insight into model consistency and subgroup dynamics. The choice between them—and often the decision to use them together—depends on whether the goal is high-level explanation or detailed behavioural analysis.

5 RECOMMENDATIONS & FUTURE WORK

PDPs strengths are it scalability, being able visualize instances without a coded limit, making them effective summarizing broad model behaviours and support strategic oversights. Yet, they have a few draw backs of averaging out the data making each instance lose their individual depth. ICE, on the other hand, provides a deeper insight into models features consistently. But the main draw back is the lack of scalability, ICE plots can only display one instance [5]. Therefore our recommendation is combining both ICE and PDP into one Feature summary tool because ICE Plots and PDPs cover each others weakness, ICE Plots can be used to focus in on any one instance that seems awry when showed through PDPs, this would cover PDP' weakness of generalizing instances by focusing on them. The same would be for ICE, ICE struggling with scalability, it can be fixed with PDPs' ability to output multiple instances at the same time. This tool would be powerful, being able to visualize and explain decisions on a graph digestible for anyone we believe that this tool could be vital in advancing cybersecurity. Our suggestions of our future work is to further research aninto the combinations of PDP and ICE especially in cybersecurity. We believe that the niche of this combination in cybersecurity is a lost opportunity to make Explainability quick and accessible. Lastly, we want to further expand current XAI practices by integrating feature explainability with other interpretability models such as Local-Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations(SHAP), to provide a more in-depth analysis for the more experienced user as well.

6 CONCLUSION

Throughout this paper we explained both the what and why of why XAI,is the best next new step into cybersecurity. XAI's integration into the cyber world will increase efficiency by reducing the uncertainty, that AI decisions will lead to, with the black box models needed currently to produce results with depth. In this paper we covered, the important methodologies needed to help explain AI models decisions, Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) whose interpretability is focused in feature summary visualization giving detailed graphs to explain decisions. The recommendations focused on the potential combination of PDP and ICE into the same Explainable graph, giving the benefits of PDPs scalability to see all of the attacks, while also bringing in the benefits of ICEs more detailed examination on every instance making a dangerous combination by putting both together to cover each other weakness. Lastly we spoke about the future work that we plan to do by combining these feature summary visualizations with the feature explanations models, LIME and SHAP to provide both in-depth visual look for quick analysis and for general understanding and explanation models for a more in-depth look

and for the more experience cybersecurity analysts. Explainable AI, is truly the next step in the advancement into integrating artificial intelligence into cybersecurity. And feature summary visualization, will be the new way to make more digestible information for those not as deeply intertwined with cybersecurity.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[2] Osvaldo Arreche, Tanish R Guntur, Jack W Roberts, and Mustafa Abdallah. 2024. E-xai: Evaluating black-box explainable ai frameworks for network intrusion detection. *IEEE Access* 12 (2024), 23954–23988.

[3] Dmytro Batishchev and Motaz Saad. 2025. The Black Box Problem: AI Decision-Making in Critical Infrastructure and Its Implications.

[4] Nicola Capuano, Giuseppe Fenza, Vincenzo Loia, and Claudio Stanzione. 2022. Explainable artificial intelligence in cybersecurity: A survey. *Ieee Access* 10 (2022), 93575–93600.

[5] Fabien Charmet, Harry Chandra Tanuwidjaja, Solayman Ayoubi, Pierre-François Gimenez, Yufei Han, Houda Jmila, Gregory Blanc, Takeshi Takahashi, and Zonghua Zhang. 2022. Explainable artificial intelligence for cybersecurity: a literature survey. *Annals of Telecommunications* 77, 11 (2022), 789–812.

[6] Alexandre Duval. 2019. Explainable artificial intelligence (XAI). *MA4K9 Scholarly Report, Mathematics Institute, The University of Warwick* 4 (2019).

[7] Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. 2024. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation* 16, 1 (2024), 45–74.

[8] Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. 2021. Explainable Artificial Intelligence Approaches: A Survey. *arXiv preprint arXiv:2101.09429* (2021).

[9] Xavier Larriva-Novo, Luis Pérez Miguel, Victor A Villagra, Manuel Álvarez-Campana, Carmen Sanchez-Zas, and Óscar Jover. 2024. Post-Hoc Categorization Based on Explainable AI and Reinforcement Learning for Improved Intrusion Detection. *Applied Sciences* 14, 24 (2024), 11511.

[10] Xiaoming Liu, Danni Huang, Jingyu Yao, Jing Dong, Litong Song, Hui Wang, Chao Yao, and Weishen Chu. 2025. From Black Box to Glass Box: A Practical Review of Explainable Artificial Intelligence (XAI). *AI* 6, 11 (2025), 285.

[11] Subash Neupane, Jesse Ables, William Anderson, Sudip Mittal, Shahram Rahimi, Ioana Banicescu, and Maria Seale. 2022. Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities. *IEEE Access* 10 (2022), 112392–112415.

[12] Zhibo Zhang, Hussam Al Hamadi, Ernesto Damiani, Chan Yeob Yeun, and Fatma Taher. 2022. Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research. *IEEE Access* 10 (2022), 93104–93139. https://doi.org/10.1109/ACCESS.2022.3204051