1. Discuss how the explanation provided by your chosen XAI method(s) can improve the efficiency and workflow of industrial security analysts.

In the context of the CIC EV Charger Attack Dataset 2024, XAI methods like SHAP provide valuable insights into model predictions, enabling industrial security analysts to effectively monitor and secure EV charging infrastructure. The SHAP summary plot highlights time as a crucial feature, with nighttime periods (lower time values) correlating strongly with attack behaviors, aligning with real-world expectations that attacks are more likely when monitoring is reduced. This insight enables analysts to set time-based monitoring thresholds, increasing vigilance during higher-risk periods.

Additionally, the SHAP force plot emphasizes the role of power-related features such as shunt voltage, bus voltage, and current in distinguishing benign from attack behaviors. High shunt voltage and current are strongly associated with attacks, while bus voltage provides mixed signals. By understanding these feature interactions, analysts can pinpoint abnormal power consumption patterns, rapidly identifying potential threats. This transparency not only builds trust in the model but also guides more efficient resource allocation and decision-making, ultimately enhancing real-time threat detection and minimizing risks to EV charging infrastructure.

LIME provides instance-specific explanations by showing how features influence the model's prediction, either pushing it toward the attack class (red bars) or the benign class (green bars). In the LIME plots for the first 10 instances, time often pushes the prediction toward the attack class, as seen in 7 out of 10 cases with red bars, but not consistently, highlighting its context-dependent role. The weight on the x-axis represents the magnitude of each feature's contribution to the prediction, with larger absolute values indicating stronger influence. The y-axis intervals display the feature value ranges, revealing patterns such as lower time values (nighttime) being more frequently linked to attack predictions, while higher bus_voltage_V values often lean toward the benign class. These intervals provide interpretable thresholds that analysts can use to identify critical feature behaviors. By combining these localized insights with SHAP’s global patterns, analysts can better understand specific anomaly triggers and refine detection mechanisms tailored to varied scenarios in EV charging infrastructure.

(Detailed data analysis in XAI_all.ipynb)

2. Discuss the strengths and weaknesses of the selected XAI method(s).

Strengths of SHAP
SHAP excels at providing a comprehensive understanding of a model's behavior through its combination of summary and force plots, enabling both global and instance-level explanations. The summary plot shows the overall contribution and importance of each feature across the dataset, while the force plot reveals the contributions of specific features for individual predictions. Furthermore, SHAP includes tools like dependence plots, which

allow analysts to visualize the relationships and correlations between features and their impact on the model's predictions. Additionally, SHAP captures the distribution of feature importance, providing insights into how different feature values influence predictions, making it highly effective for detecting patterns across varying instances. These strengths significantly enhance the workflow of industrial security analysts by enabling them to identify critical features, understand global trends in attack behaviors, and focus on key variables for anomaly detection and resource allocation.

Weaknesses of SHAP

Despite its strengths, SHAP can be computationally expensive, particularly for complex models and large datasets, as it relies on Shapley values that require evaluating all possible feature subsets to estimate contributions. This high computational cost can limit its feasibility for real-time analysis, which is critical for industrial security analysts monitoring EV charging infrastructure. As a result, the background data is sampled to size 1000 in this case before applying the KernelExplainer. Additionally, while SHAP provides detailed and interpretable results, the sheer volume of information it generates can overwhelm analysts, making it challenging to prioritize actionable insights. For instance, in scenarios with numerous features, analysts might find it difficult to quickly focus on the most relevant factors affecting specific predictions. These limitations can reduce the method's practicality in time-sensitive workflows, particularly in environments requiring immediate threat detection.

Strengths of LIME

LIME's primary strength lies in its simplicity and clarity in generating local explanations for individual predictions. The method highlights the influence of specific feature values on a given instance by assigning interpretable weights, as seen in the LIME plots where feature contributions are clearly distinguished as pushing predictions toward either the attack or benign class. The feature value intervals on the y-axis also provide interpretable thresholds that analysts can use to assess specific conditions triggering certain predictions. LIME's focus on instance-level explanations allows security analysts to dissect individual cases and understand localized patterns of malicious behavior, which is particularly useful for investigating anomalies or false positives. Moreover, LIME's faster computation compared to SHAP makes it more suitable for real-time or near-real-time analysis, improving response times in high-stakes environments.

Weaknesses of LIME

A notable limitation of LIME is its reliance on local linear approximations, which may fail to capture complex, non-linear relationships in the underlying model. This can lead to explanations that oversimplify the true behavior of the model, potentially causing misinterpretation of feature importance. For example, while LIME plots effectively show how features like time or power_mW influence predictions, they do not provide a global view of feature importance or interactions, limiting the insights available to security analysts compared to SHAP's summary and dependence plots. Additionally, LIME's explanations can vary depending on the sampling process used to create local approximations, introducing potential inconsistencies in the results. These weaknesses may hinder the ability of industrial

security analysts to fully understand broader trends and ensure robust, reliable detection mechanisms for EV charging infrastructure.

3. Propose ideas to refine the explanations generated by the chosen XAI method(s), especially for one or more stakeholders with interest in industrial (ICS/CPS) security systems.

To refine the explanations generated by SHAP and LIME for stakeholders in industrial security systems (ICS/CPS), I propose integrating these XAI methods with the concept of Digital Twin, initially articulated in NASA's 2010 Roadmap Report by John Vickers, encompasses the replication of a 3D model from the physical realm to the simulated digital domain. It provides virtual, real-time replicas of physical assets, and embedding XAI insights into these can enhance interpretability and usability for engineers. This integration addresses a significant challenge: the statistical and abstract nature of XAI outputs can be difficult for non-expert stakeholders to interpret, potentially limiting their effectiveness in practical, high-stakes environments.

In this approach, the Digital Twin becomes a dynamic interface where XAI-generated explanations are visually and contextually represented. For example, in a predictive maintenance scenario, if SHAP analysis identifies that anomalous power readings are a key contributor to a model's prediction of an impending equipment failure, this insight is directly mapped onto the Digital Twin. Components or sensors with high influence are highlighted within the virtual environment, allowing stakeholders to visually identify critical areas of concern. Interactive features can provide detailed explanations, such as the magnitude of each feature's contribution to the prediction, ensuring that decision-makers understand not only the "what" but also the "why" behind AI alerts.

This integration significantly improves the interpretability of complex AI systems. Instead of relying on abstract plots and numerical outputs, operators see an intuitive representation of system status within a familiar context. The ability to connect XAI insights to specific processes fosters quicker comprehension and enables more targeted interventions, reducing downtime and enhancing operational efficiency. Moreover, this transparency builds trust in AI-driven systems, which is critical in safety-critical environments where decisions based on AI predictions can have significant consequences.

The combination of XAI and Digital Twin technologies also enhances training and onboarding processes. By visualizing the reasoning behind model predictions within a real-world context, new personnel can quickly grasp the interplay between AI insights and system operations. This reduces the learning curve and ensures consistent application of AI-driven recommendations. Furthermore, the alignment between XAI explanations and Digital Twin visualizations promotes collaboration across multidisciplinary teams, ensuring that engineers, data scientists, and operators have a shared understanding of system behaviors. Through these technologies, industrial stakeholders can improve system reliability,

optimize performance, and maintain convenient control in increasingly complex environments.