Explainable Artificial Intelligence as a Tool for Function Extraction Attacks

*Ka-Kyung Kim[a], Joon-Seok Kim[a], Ieck-Chae Euom[a]**
[a] Chonnam National University System Security Research Center, 77 Yongbong-ro 138beon-gil, Gwangju, 61186
**Corresponding author: iceuom@jnu.ac.kr*

***Keywords :** Explainable Artificial Intelligence, Model Extraction Attack, Industrial Control System, Adversarial Attack, AI-Powered System

1. Introduction

The introduction of artificial intelligence (AI) technology in industrial control systems enhances efficiency and safety throughout the entire lifecycle. However, AI inherently involves unresolved risks due to its uncertainty and complexity. In safety-critical environments, a more differentiated approach is necessary to mitigate AI risks compared to other environments.

European Union (EU) has classified "safety-related AI systems used in the management and operation of critical infrastructure" as high-risk and imposed strict transparency requirements on AI models in its first comprehensive AI regulation, AI Act (Regulation (EU) 2024/1689), which came into effect on August 1, 2024 [1].

In South Korea, the "Basic Act on the Development of Artificial Intelligence and the Establishment of a Trust-Based Framework [2]" scheduled to take effect on January 22, 2026, designates nuclear facilities as high-risk AI areas under Article 2(4) and imposes responsibilities on businesses related to high-risk AI under Article 34.

To mitigate the "black-box" problem caused by the uncertainty and complexity of AI, research is being conducted to ensure "Explainability" [3]. The AI Risk Management Framework (AI RMF, NIST AI 100-1[4]) published by the National Institute of Standards and Technology (NIST) in the United States also explicitly states that "Explainable and Interpretable" must be ensured.

Explainability involves a structure that provides additional output of the rationale and internal information behind AI model decisions. However, there is a possibility that attackers could exploit explainability to access AI models and misuse it. As the network connectivity of industrial control systems increases, the number of points where attackers can gain access also increases, thereby raising the likelihood of such exploitation.

This study discusses adversarial attack techniques that exploit explainability to extract the functionality of local AI models. Experiments were conducted to confirm how well the proxy models generated for extracting the functionality of target models mimic the existing models and infer sensitive internal information of the models.

2. Background

Explainable artificial intelligence primarily employs methods such as SHAP (SHapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanation), Counterfactual Explanation, Attention Mechanism, and gradient-based explanations. Among these, research on SHAP and LIME, which can provide proxy interpretations regardless of the model, is the most widespread.

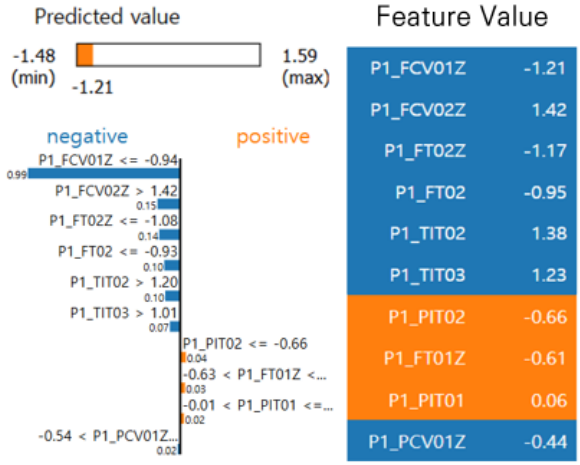

Figure 1 LIME Explanation

In this study, LIME is selected and used as it is a relatively computationally efficient XAI method. LIME is a technique designed to generate local explanations by creating a target model. It works by approximating nonlinear patterns with local linear models to generate explanations. The goal is to find the linear model that best explains the input data 'a' and provide explanations for the model predictions. The LIME method randomly generates similar data points located near a specific data point, then creates a linear model based on the generated data points. By observing changes in the model's output, the contribution of each feature can be calculated.

LIME explains the decision-making process of locally interpreted deep learning models by providing feature contributions, class classification results, and local decision boundaries. In Figure 1, "Predicted value" indicates that the range of the corresponding data

point is between -1.48 and 1.59. The features under the label “Negative” contribute to lowering the model prediction, while those under the label “Positive” contribute to raising the model prediction. The inequalities associated with each feature indicate the conditions that must be met for the feature to be negative or positive in the model prediction.

For example, feature P1_FCV01Z is a feature that contributes to identifying the data point as normal when its value is -0.94 or lower. The table for ‘Feature Value’ shows the values and colors associated with the analyzed data points. Furthermore, since LIME generates a linear model as a local proxy model, the underlying function of the linear model can also be extracted with additional implementation effort.

3. Functional Extraction Attacks Through Explainable Artificial Intelligence Exploitation

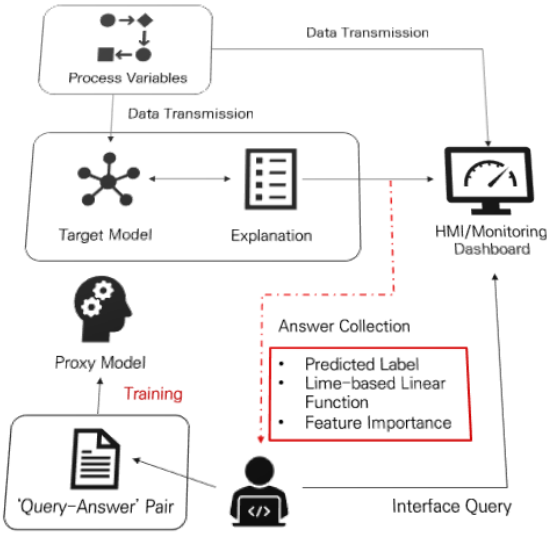

Figure 2 Structures of Explainability Exploits and Proxy Model Learning

The proposed method is written from the attacker's perspective. Attacker analyzes information about the target environment. If the target is a video-based AI system, the attacker sets up a proxy model using image processing algorithms. If the target is an operational data processing AI system, the attacker sets up a proxy model using time series data processing algorithms.

The attacker intentionally manipulates the data input into the AI model and collects the AI model's prediction results and the explanations provided by XAI that change according to the manipulated data. The attacker trains the proxy model using “query-response” pairs, where the manipulated data is set as the ‘query’ and the AI model's prediction results and explanations are set as the “response.” As a result, the attacker successfully replicates the AI model by exploiting explainability.

Following this, the attacker evaluates the performance and fidelity of the proxy model, refines it, and uses it as a robust foundation for conducting further attacks. Fidelity is set as an evaluation metric to determine how similar the prediction results are to those of the existing target model. While performance evaluation is conducted by comparing the actual labels with the model's predicted labels, fidelity evaluation determines whether the prediction results of the target model and the proxy model match. The higher the fidelity, the higher the degree of emulation of the proxy model.

4. Case Study

4.1 Datasets

Using the publicly available HAI control system dataset [10], we experimented with an adversarial attack method based on explainability exploitation for replicating the AI model proposed in this paper.

The HAI dataset was collected in an environment where industrial control system test beds from GE, EMERSON, and FESTO were integrated with a HIL simulator. The testbed includes turbines, boilers, water treatment processes, and HIL simulation. In this case study, we used the latest version, HAI 23.05, and utilized a total of 87 variables, excluding timestamps.

4.2 Function Extract Target Model Building

To build a function extraction target model, this study constructed “LSTM-AE” by combining Long Short-Term Memory (LSTM), which excels at time series data modeling and generation processing, with an Autoencoder (AE).

After training the target model with 896,400 operational data points collected during normal operation, the model was configured to predict anomalies every second. Explanations from LIME connected to the target model are also generated every second.

4.3 Model Function Extract Scenario

Usually, explainability is provided to operators or monitors. Therefore, assume that LIME-based explanations are provided in the main control room, HMI, or remote monitoring software. Additionally, systems combining AI require additional infrastructure such as data pipelines, model development-deployment-operation servers, and external APIs. In other words, the AI infrastructure itself is located in the computing environment, which can introduce various attack vectors.

An attacker is assumed to have gained access to the internal network through traditional cyber-attacks and obtained access to the process data pipeline that feeds into the target model. The attacker intentionally

manipulates the data, and the AI model performs operational monitoring based on the manipulated data. In this scenario, LIME techniques are applied to provide explanations for the target model alongside the operational monitoring data on the driver's dashboard.

The attacker manipulates data as “queries,” and LIME provides explanations as “responses,” constructing a dataset of “query-response” pairs. The “responses” include variable-specific contributions provided by LIME and whether the target model detected normal or abnormal conditions. Assuming realistic security policies, information about the anomaly probability predicted by the target model is not output.

4.4 Building, Training, and Evaluating the Fidelity of Proxy Models

This study assumes that the attacker chose an approach utilizing a proxy model based on a deep network architecture for high-dimensional pattern learning to construct the proxy model. LSTM, GAN, and Transformer architectures were used in the proxy model construction experiments.

By training the proxy model using “query-response” pairs, the attacker successfully replicates the existing target model. Queries represent input data, while responses include the target model’s predicted labels alongside LIME explanations such as feature importance, linear regression coefficients.

The target model and proxy model output ‘0’ for normal states and ‘1’ for abnormal states, and the accuracy of both models was evaluated. The predicted values from the target model and proxy model were output as ‘0’ for normal states and ‘1’ for abnormal states, and the accuracy of both models was evaluated.

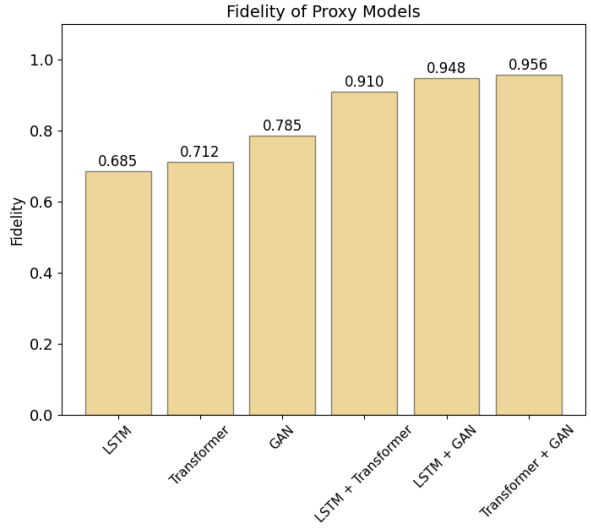

Figure 3 Fidelity of Proxy Models

Fidelity evaluation utilized predictions made on the entire set of 284,400 test data points collected under conditions where measurement replay attacks, false data injection attacks, and control setting manipulation attacks were introduced. The agreement between the target model's predictions and the proxy model's predictions was assessed. The results are presented in Figure 3.

While these results do not indicate a significant difference in model fidelity, it is evident that GAN, Transformer, and LSTM, in that order, learned meaningful patterns in the ‘query-response’ pairs more effectively. This holds true in terms of both single models and hybrid models.

5. Conclusion

This paper proposes an approach that utilizes XAI (explainable artificial intelligence) to solve the “black box” problem caused by the risks of ‘uncertainty’ and “complexity” in artificial intelligence. Using AI model information collected from explainability, we set up model replication scenarios and presented the results through experiments.

The replicated model can independently infer the decision boundaries of the target model and may leak sensitive information in an industrial control system environment. Additionally, the replicated model could serve as a foundation for developing more sophisticated and covert attacks. Therefore, AI systems introduced for cybersecurity, operational efficiency, and safety may introduce new security threats. AI can be used as a defense against attacks, but it can also be used as an attack tool.

Future research aims to explore methods for clearly inferring the decision boundaries of target models using replicated models, extracting inference-based process control capabilities, and detecting models trained on data with embedded backdoors.

ACKNOWLEDGMENT

The results of a study on the supported by Nuclear Safety Research Program through the Korea Foundation of Nuclear Safety (KoFONS) using the financial resource granted by the Nuclear Safety and Security Commission (NSSC) of the Republic of Korea (No.2106061, 50%) and was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2025-25394739, Development of Security Enhancement Technology for Industrial Control Systems Based on S/HBOM Supply Chain Protection, 50%)

REFERENCES

.

[1] European Union, “Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act),” Official Journal of the European Union, L 202, pp. 1–96, Jul. 12, 2024, doi: 10.2804/4225375.

[2] Republic of Korea, “Framework Act on the Development of Artificial Intelligence and the Establishment of Trust (AI Basic Act),” Act No. 20211, Jan. 21, 2025, effective Jan. 22, 2026. [Online]. Available: https://www.law.go.kr/lsInfoP.do?lsiSeq=268543.

[3] D. Gunning, E. Vorm, J. Y. Wang, and M. Turek, "DARPA's explainable AI (XAI) program: A retrospective," DARPA, Arlington County, VA, USA, Tech. Rep., 2021, doi: 10.22541/au.163699841.19031727/v1.
[4] National Institute of Standards and Technology, "Artificial intelligence risk management framework (AI RMF 1.0)," 2023. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.
[5] H.-K. Shin, W. Lee, S. Choi, J.-H. Yun, and B.-G. Min, "HAI security datasets," GitHub, 2023. [Online]. Available: https://github.com/icsdataset/hai.