Adversarially Robust and Interpretable Magecart Malware Detection

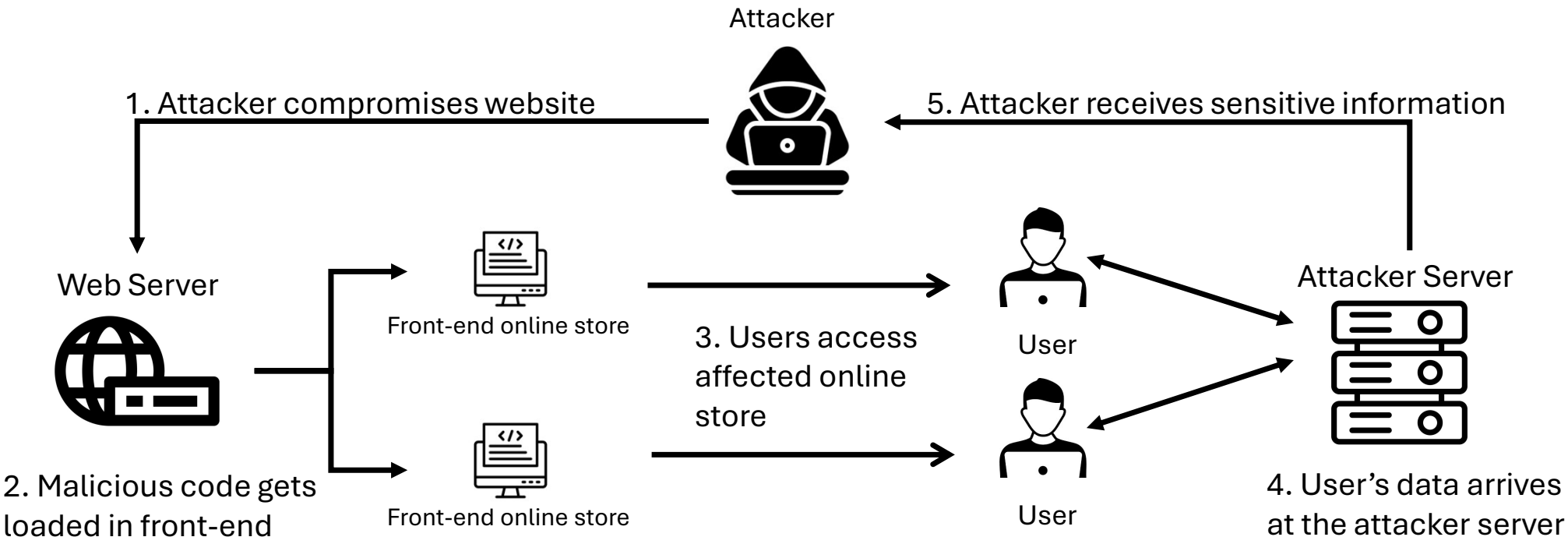
Pedro Pereira, José Gouveia, João Vitorino, Eva Maia, Isabel Praça

NCA 2025
23 rd IEEE International Symposium on Network Computing and Applications
Lisbon, Portugal, November 05-07

Context and Motivation

- **JavaScript malware** on the client side is increasingly exploited to **bypass traditional defenses** and compromise user data.
- This type of compromise puts **sensitive user information** at risk and significantly reduces **user trust** in web applications and online services.
- There is a growing need for a **lightweight, robust** and **explainable approach** that can detect **malicious scripts** in real time.
- Combining **robust machine learning** models with **explainable AI** allow us to **detect malicious attacks** and understand why it is an actual attack.

Context and Motivation - What is a Magecart Attack?



Proposed Solution

1

Machine learning (ML) models
+
Behavior Deterministic Finite Automaton (DFA)

2

Behavioral Feature Engineering

3

Adversarial Robustness

4

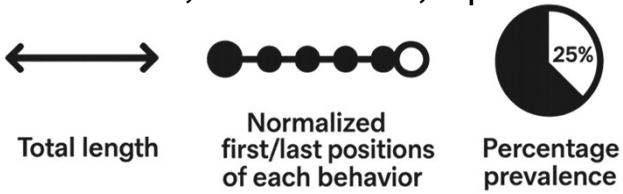
Explainable AI (XAI)

Data Preprocessing

JavaScript execution logs

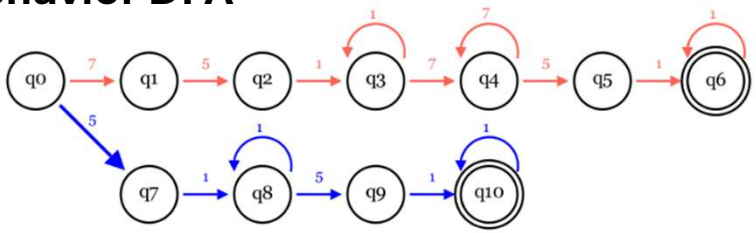
Sequence-based

"Set Callback", "Add Event Handler", "Acces Input, Access DOM Element Attribute", "Send Data, Update DOM Element"



Total Length: 4
First_SetCallback: 0.0
Last_SetCallback: 0.0
...
SetCallbackPrcntg: 25%

Behavior DFA



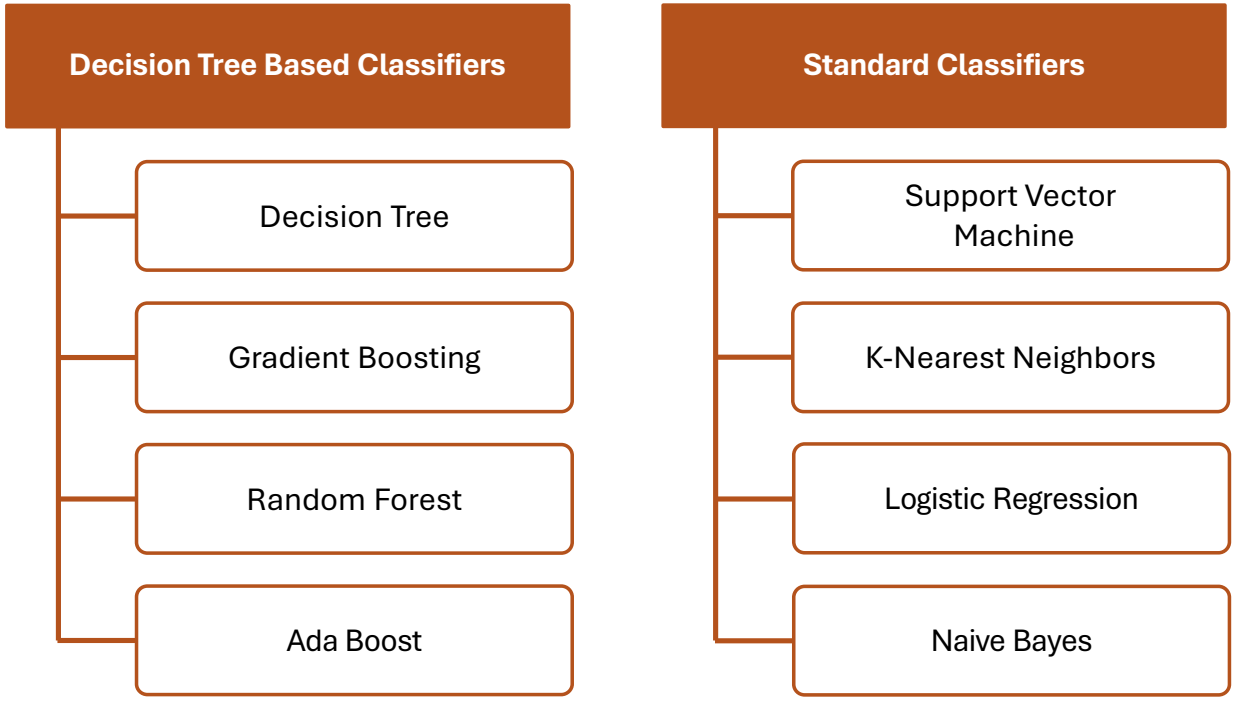
Classification: 75%

Contextual

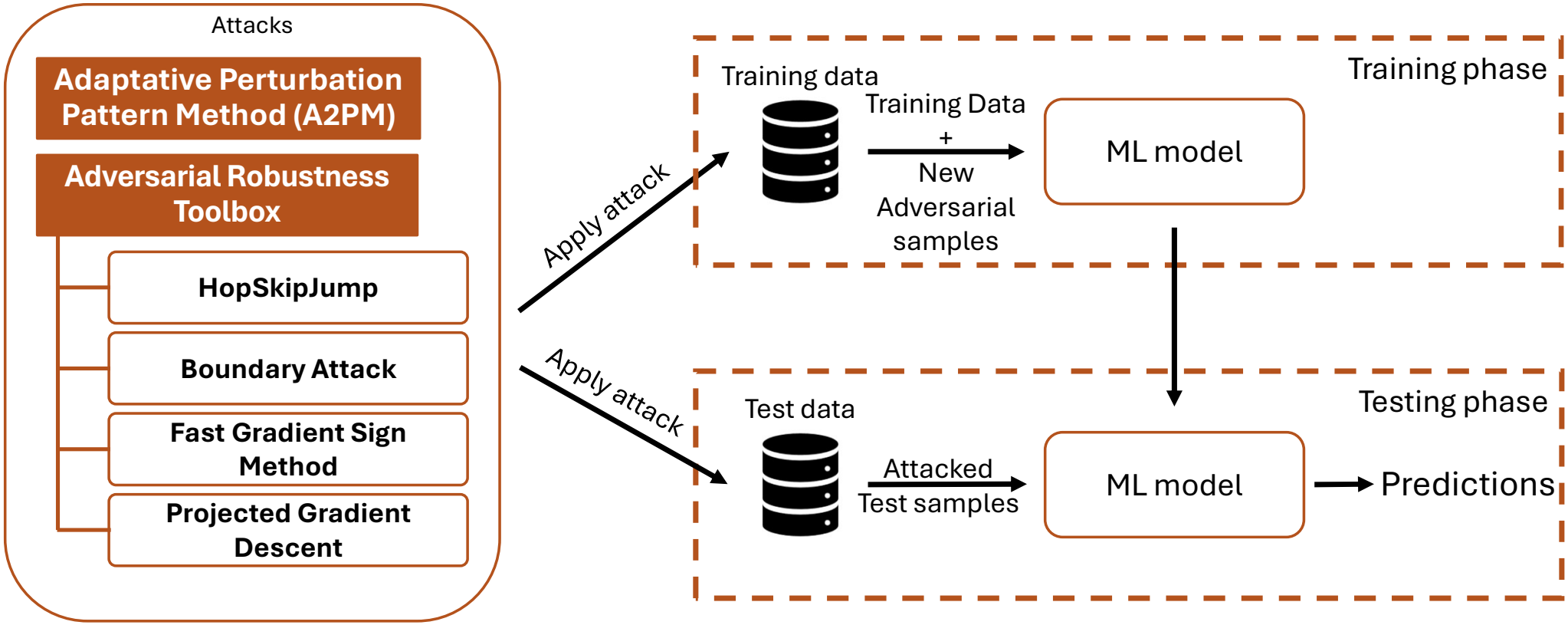
Average user clicks preceding an action

Binary indicator for dynamically injected scripts

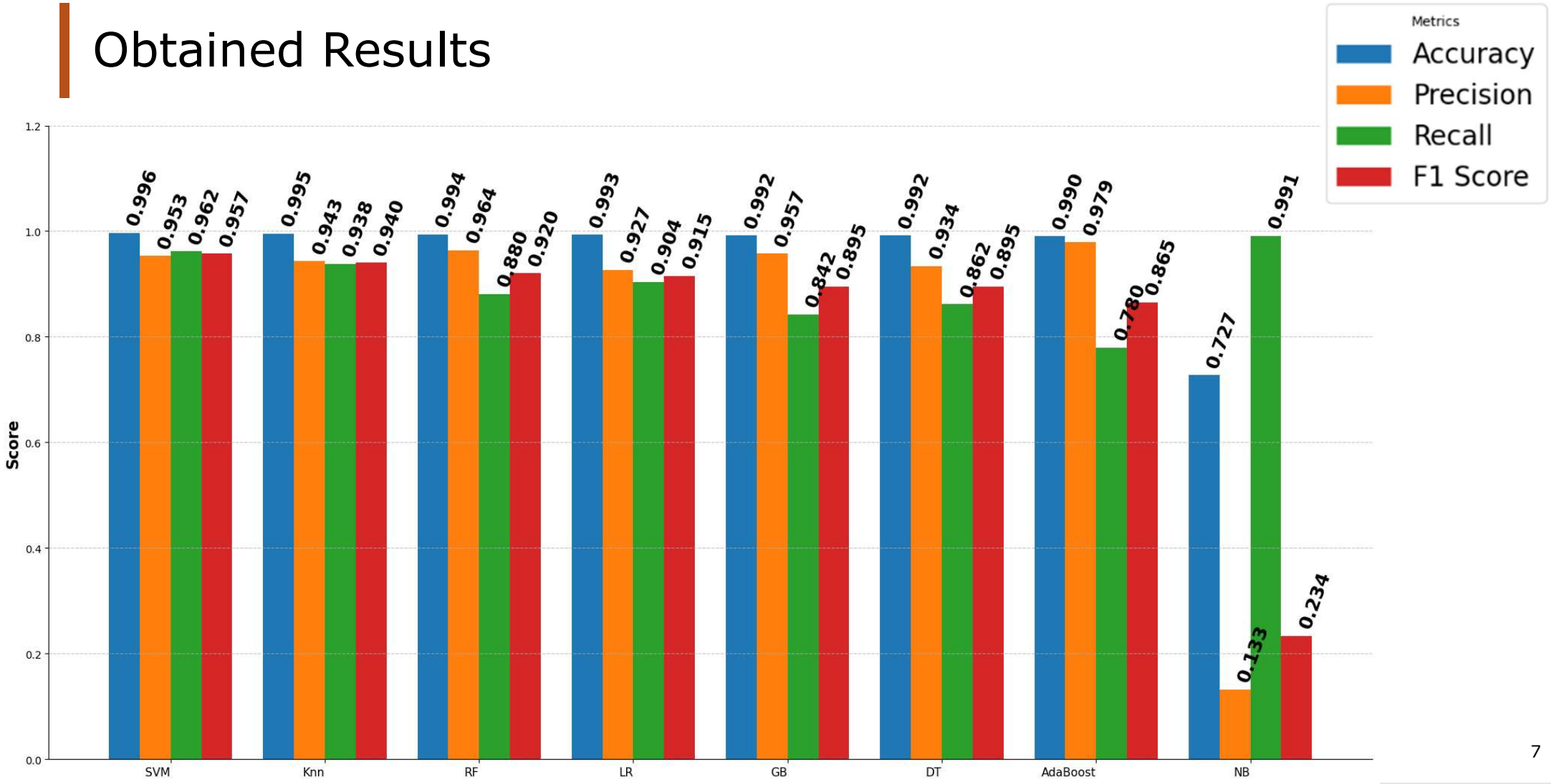
Machine Learning Models



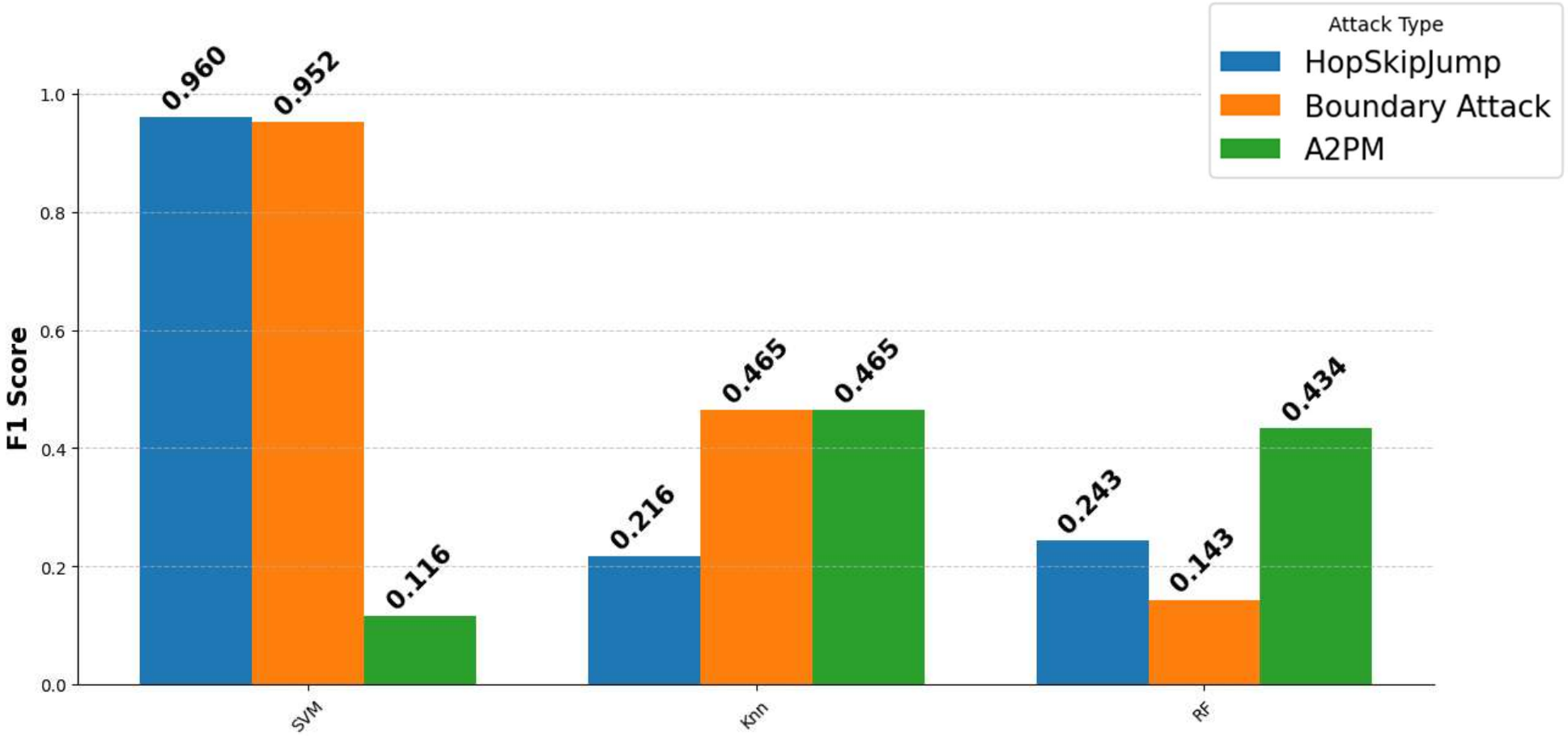
Adversarial Robustness Evaluation



Obtained Results

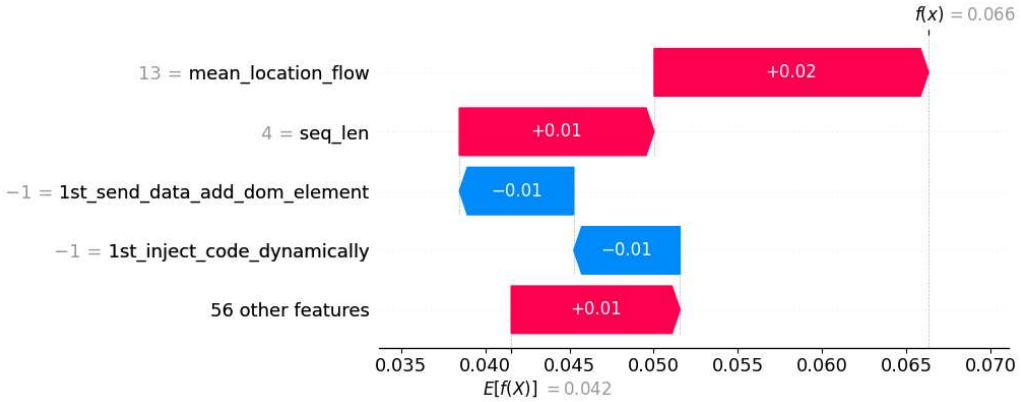


Obtained Results - Robustness

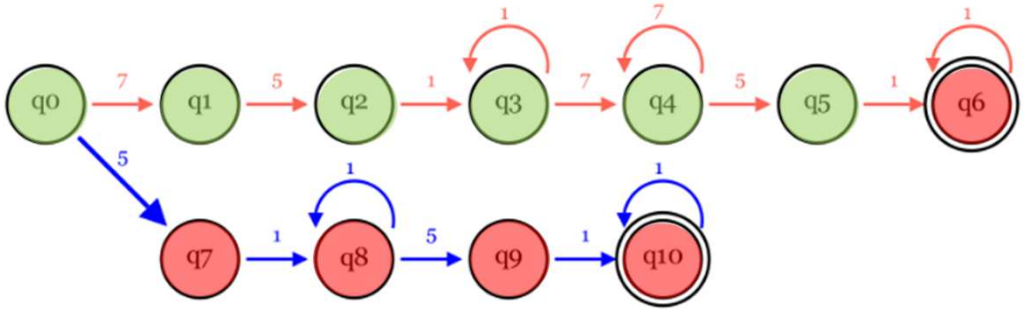


Interpretable Model Explanations

SHAP Values



Behavior DFA



LLM

Llama Scout 4

Obtained Results - Explainability

This script is classified as malicious with a high risk of approximately 99.66%.

The Automata Model has identified a strong match to known malicious patterns, labeling it as MALIGN with a 100% match percentage, primarily due to behaviors such as setting callbacks, adding event handlers, accessing input and DOM elements, creating DOM elements, and sending data, which are commonly seen in malicious scripts.

The ML Model, which assesses risk based on specific features, also indicates a high risk, with features like adding DOM elements, sending data, and updating DOM elements contributing to the malicious classification, while safe behaviors like accessing known content or using standard features are not prominent; notably, there are no significant indications of benign activities.

Overall, the script appears to be harmful, with a high likelihood of performing malicious actions, and human review is recommended to understand its exact capabilities and mitigate potential threats.

Obtained Results - Explainability

Global Classification (Final Output)

This script is classified as malicious with a high risk of approximately 99.66%.

The Automata Model has identified a strong match to known malicious patterns, labeling it as MALIGN with a 100% match percentage, primarily due to behaviors such as setting callbacks, adding event handlers, accessing input and DOM elements, creating DOM elements, and sending data, which are commonly seen in malicious scripts.

The ML Model, which assesses risk based on specific features, also indicates a high risk, with features like adding DOM elements, sending data, and updating DOM elements contributing to the malicious classification, while safe behaviors like accessing known content or using standard features are not prominent; notably, there are no significant indications of benign activities.

Overall, the script appears to be harmful, with a high likelihood of performing malicious actions, and human review is recommended to understand its exact capabilities and mitigate potential threats.

Obtained Results - Explainability

Global Classification (Final Output)

Behavior DFA (Symbolic Reasoning Layer)

This script is classified as malicious with a high risk of approximately 99.66%.

The Automata Model has identified a strong match to known malicious patterns, labeling it as MALIGN with a 100% match percentage, primarily due to behaviors such as setting callbacks, adding event handlers, accessing input and DOM elements, creating DOM elements, and sending data, which are commonly seen in malicious scripts.

The ML Model, which assesses risk based on specific features, also indicates a high risk, with features like adding DOM elements, sending data, and updating DOM elements contributing to the malicious classification, while safe behaviors like accessing known content or using standard features are not prominent; notably, there are no significant indications of benign activities.

Overall, the script appears to be harmful, with a high likelihood of performing malicious actions, and human review is recommended to understand its exact capabilities and mitigate potential threats.

Obtained Results - Explainability

This script is classified as malicious with a high risk of approximately 99.66%.

The Automata Model has identified a strong match to known malicious patterns, labeling it as MALIGN with a 100% match percentage, primarily due to behaviors such as setting callbacks, adding event handlers, accessing input and DOM elements, creating DOM elements, and sending data, which are commonly seen in malicious scripts.

The ML Model, which assesses risk based on specific features, also indicates a high risk, with features like adding DOM elements, sending data, and updating DOM elements contributing to the malicious classification, while safe behaviors like accessing known content or using standard features are not prominent; notably, there are no significant indications of benign activities.

Overall, the script appears to be harmful, with a high likelihood of performing malicious actions, and human review is recommended to understand its exact capabilities and mitigate potential threats.

Global Classification (Final Output)

Behavior DFA (Symbolic Reasoning Layer)

Behavioral Pattern Evidence (Automata Output Details)

Obtained Results - Explainability

This script is classified as malicious with a high risk of approximately 99.66%.

The Automata Model has identified a strong match to known malicious patterns, labeling it as MALIGN with a 100% match percentage, primarily due to behaviors such as setting callbacks, adding event handlers, accessing input and DOM elements, creating DOM elements, and sending data, which are commonly seen in malicious scripts.

The ML Model, which assesses risk based on specific features, also indicates a high risk, with features like adding DOM elements, sending data, and updating DOM elements contributing to the malicious classification, while safe behaviors like accessing known content or using standard features are not prominent; notably, there are no significant indications of benign activities.

Overall, the script appears to be harmful, with a high likelihood of performing malicious actions, and human review is recommended to understand its exact capabilities and mitigate potential threats.

Global Classification (Final Output)

Behavior DFA (Symbolic Reasoning Layer)

Behavioral Pattern Evidence (Automata Output Details)

Machine Learning Model (Statistical Reasoning Layer)

Obtained Results - Explainability

This script is classified as malicious with a high risk of approximately 99.66%.

The Automata Model has identified a strong match to known malicious patterns, labeling it as MALIGN with a 100% match percentage, primarily due to behaviors such as setting callbacks, adding event handlers, accessing input and DOM elements, creating DOM elements, and sending data, which are commonly seen in malicious scripts.

The ML Model, which assesses risk based on specific features, also indicates a high risk, with features like adding DOM elements, sending data, and updating DOM elements contributing to the malicious classification, while safe behaviors like accessing known content or using standard features are not prominent; notably, there are no significant indications of benign activities.

Overall, the script appears to be harmful, with a high likelihood of performing malicious actions, and human review is recommended to understand its exact capabilities and mitigate potential threats.

Global Classification (Final Output)

Behavior DFA (Symbolic Reasoning Layer)

Behavioral Pattern Evidence (Automata Output Details)

Machine Learning Model (Statistical Reasoning Layer)

SHAP Feature Attribution

Obtained Results - Explainability

This script is classified as malicious with a high risk of approximately 99.66%.

The Automata Model has identified a strong match to known malicious patterns, labeling it as MALIGN with a 100% match percentage, primarily due to behaviors such as setting callbacks, adding event handlers, accessing input and DOM elements, creating DOM elements, and sending data, which are commonly seen in malicious scripts.

The ML Model, which assesses risk based on specific features, also indicates a high risk, with features like adding DOM elements, sending data, and updating DOM elements contributing to the malicious classification, while safe behaviors like accessing known content or using standard features are not prominent; notably, there are no significant indications of benign activities.

Overall, the script appears to be harmful, with a high likelihood of performing malicious actions, and human review is recommended to understand its exact capabilities and mitigate potential threats.

Global Classification (Final Output)

Behavior DFA (Symbolic Reasoning Layer)

Behavioral Pattern Evidence (Automata Output Details)

Machine Learning Model (Statistical Reasoning Layer)

SHAP Feature Attribution

Absence of Benign Indicators

Obtained Results - Explainability

This script is classified as malicious with a high risk of approximately 99.66%.

The Automata Model has identified a strong match to known malicious patterns, labeling it as MALIGN with a 100% match percentage, primarily due to behaviors such as setting callbacks, adding event handlers, accessing input and DOM elements, creating DOM elements, and sending data, which are commonly seen in malicious scripts.

The ML Model, which assesses risk based on specific features, also indicates a high risk, with features like adding DOM elements, sending data, and updating DOM elements contributing to the malicious classification, while safe behaviors like accessing known content or using standard features are not prominent; notably, there are no significant indications of benign activities.

Overall, the script appears to be harmful, with a high likelihood of performing malicious actions, and human review is recommended to understand its exact capabilities and mitigate potential threats.

Global Classification (Final Output)

Behavior DFA (Symbolic Reasoning Layer)

Behavioral Pattern Evidence (Automata Output Details)

Machine Learning Model (Statistical Reasoning Layer)

SHAP Feature Attribution

Absence of Benign Indicators

LLM-Generated Human-Readable Summary

Main Conclusions

- A robust and interpretable system for Magecart detection **was successfully achieved** by integrating Machine Learning models with the Behavior DFA framework.
- The hybrid approach provides **strong robustness**, with models demonstrating resilience against a diverse set of adversarial evasion attacks.
- The system delivers **transparent explainability** by using an LLM to translate SHAP values and DFA insights into clear, natural language justifications for its decisions.
- The **SVM** was the star performer, consistently achieving the highest results, especially in Recall , and maintaining the best overall robustness against attacks.
- For **future work** we will test more advanced **Deep Learning architectures** (like **LSTMs** or **CNNs**).

Thank you

jiaga@isep.ipp.pt

This work was done and funded in the scope of the BEHAVIOR project (NORTE2030-FEDER-00576300 no. 14391). This work was also supported by UIDB/00760/2020.