2024 - NS - 9B9-D35 - Explainability of Machine Learning in Intrusion Detection Systems - zhangy42@rose-hulman.edu - Yujie Zhang

Indiana Statewide Cybersecurity

Yujie Zhang, Xianshun Jiang
Advised by Dr. Mohammad Noureddine

Explainability of Machine Learning in Intrusion Detection Systems

Motivation

- Increasing Network Security Threat
- Huge economic loss due to attacks

Potential problems of machine learning models in IDS (Intrusion Detection Systems)

- Semantic Gap
- Shortcut Learning
- High Cost of Errors

Definition of Explainability

- The degree to which a human can understand the cause of a decision

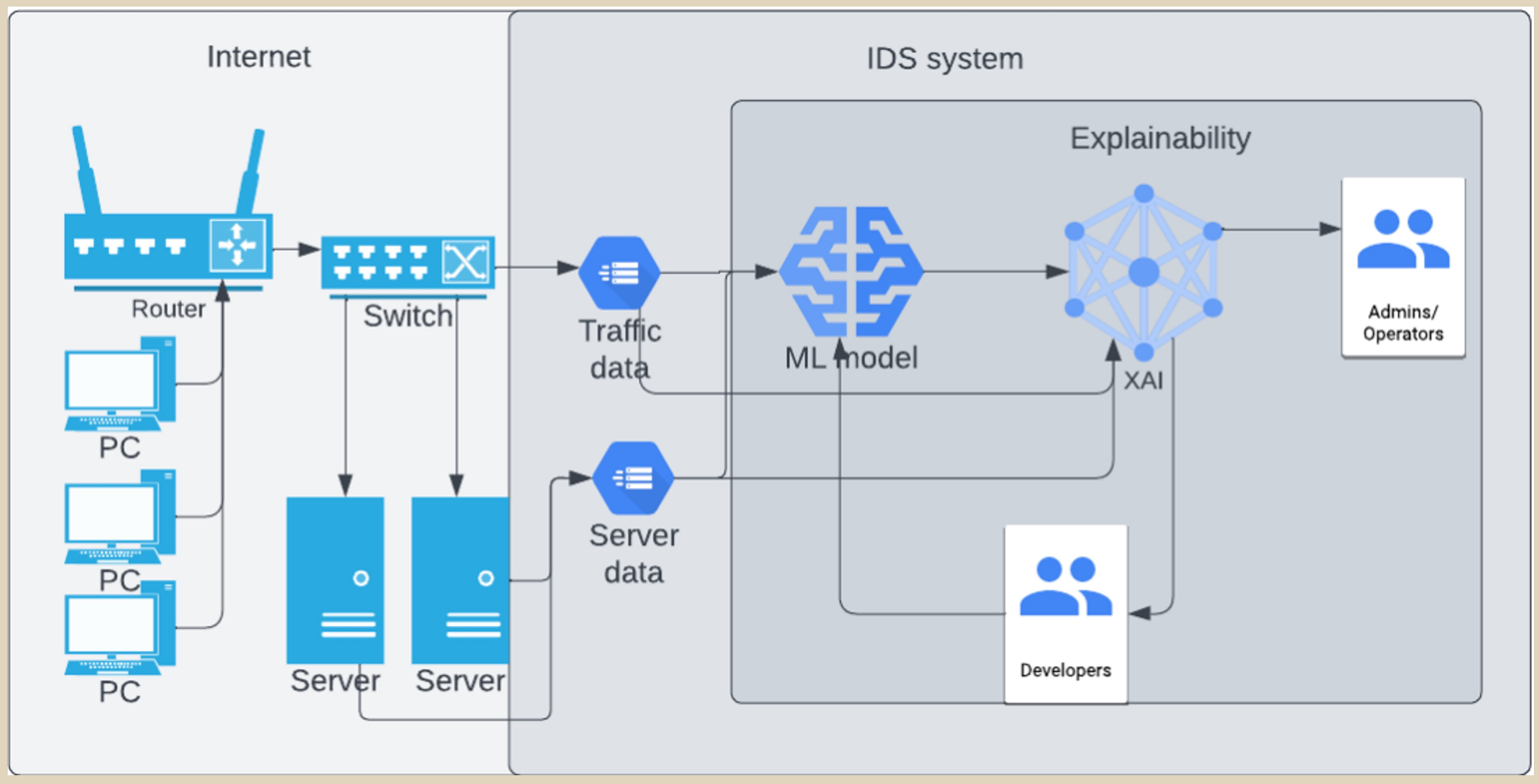


Figure 1: Our Pipeline of Implementing XAI on Network Environment

SHAP and shapley value distributions

- An Idea from game theory
- Calculate contributions of each feature to the prediction
- Explain the output of models locally with visualized reports

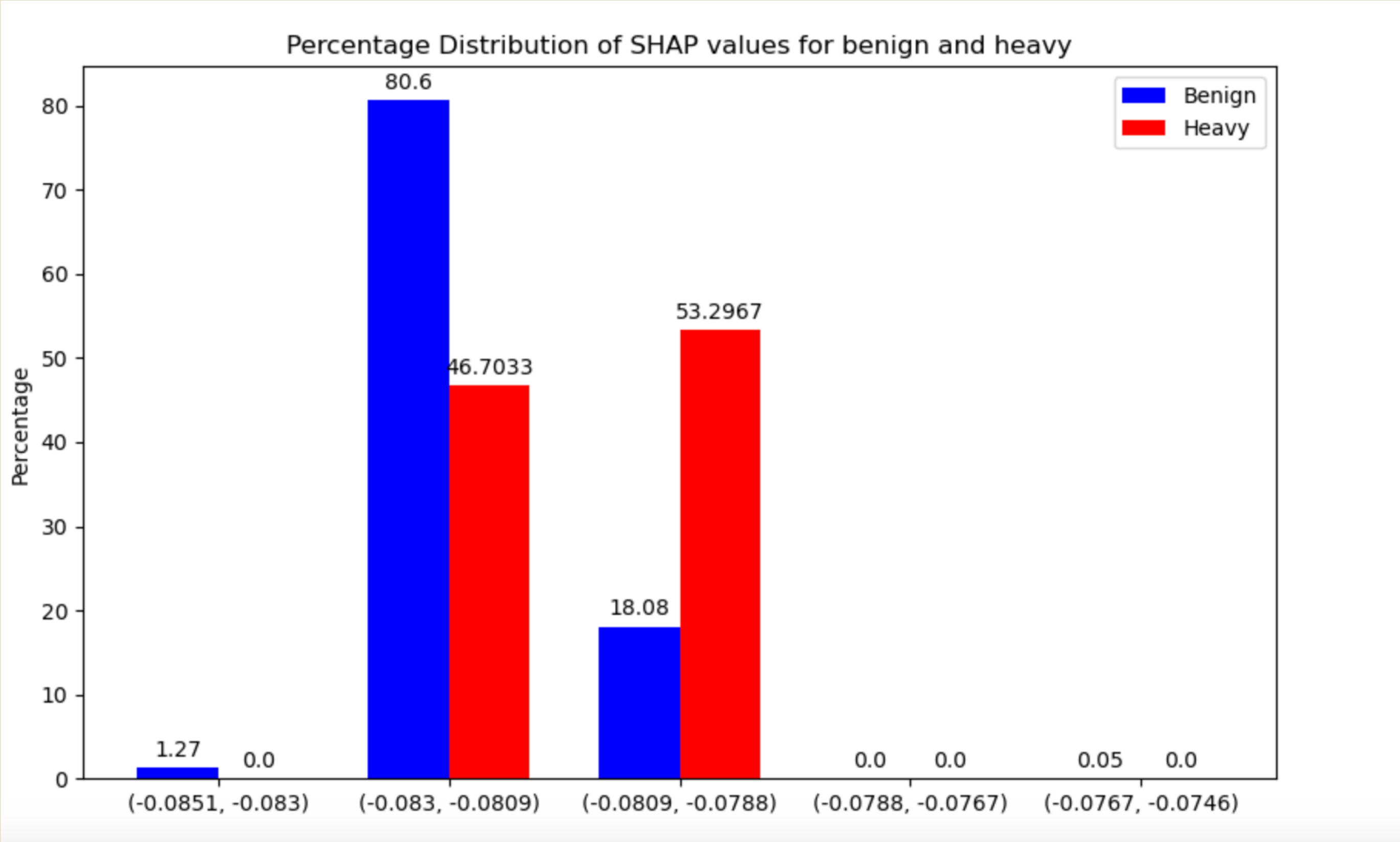
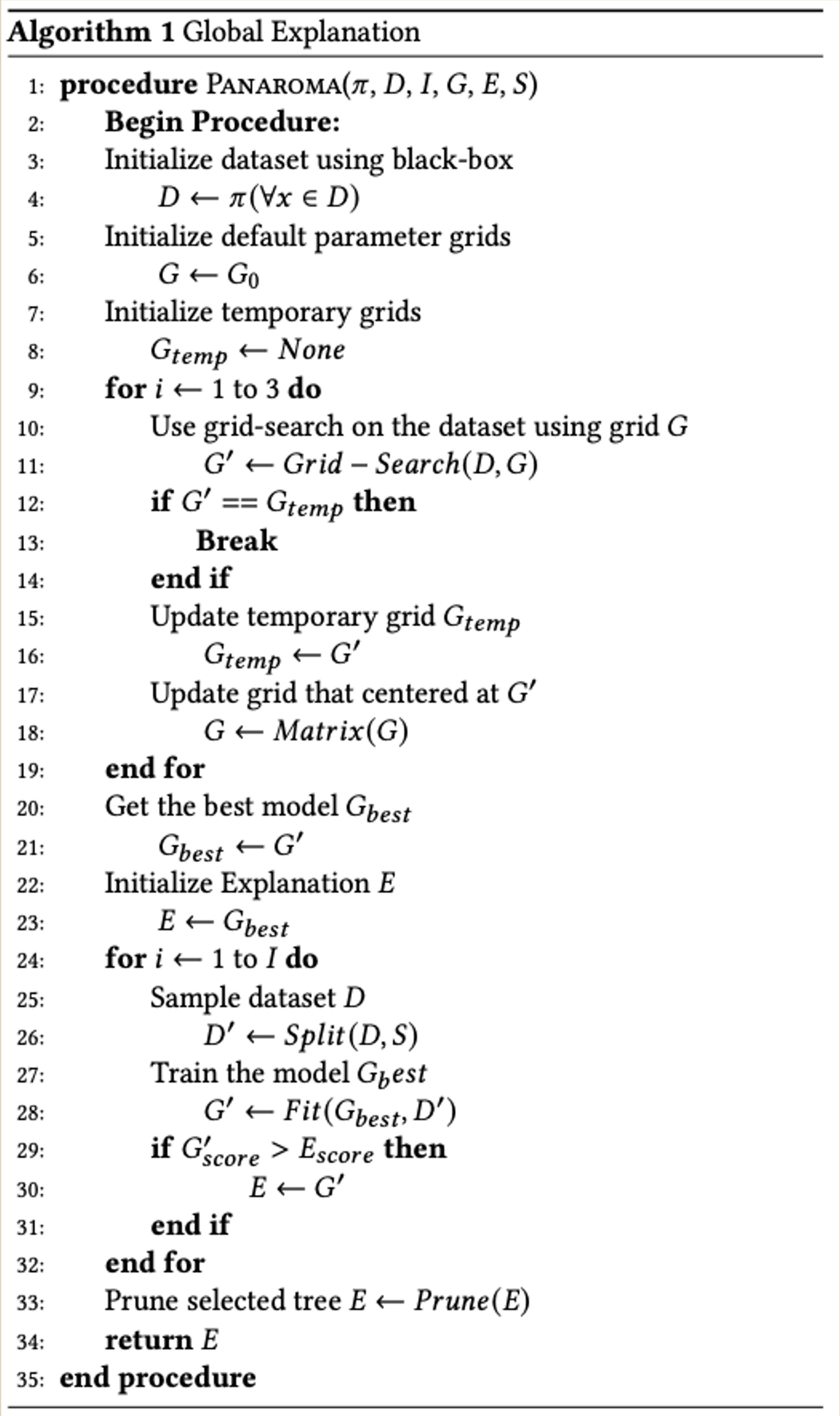


Figure 3: SHAP value distribution histogram for a specific feature value for both benign and malicious records

- Calculate the SHAP value for every feature and values for all records
- Combine the result based on the feature names and feature values
- Conclude the distribution of the SHAP value
- Give a weight for each feature based on the distribution and fit a formula for the firewall

Data Preprocessing

- Remove features with only 1 unique value
- Drop records with Nan Values
- Build reference table for string types
  - Find unique value of each features
  - Build a table based on the unique value
- Replace the original string values with index on reference table



Algorithm 1 Global Explanation

```
procedure PANAROMA($\pi, D, I, G, E, S$)
    Begin Procedure:
    Initialize dataset using black-box
        $D \leftarrow \pi(\forall x \in D)$
    Initialize default parameter grids
        $G \leftarrow G_0$
    Initialize temporary grids
        $G_{temp} \leftarrow None$
    for $i \leftarrow 1$ to 3 do
        Use grid-search on the dataset using grid $G$
            $G' \leftarrow Grid - Search(D, G)$
        if $G' == G_{temp}$ then
            Break
        end if
        Update temporary grid $G_{temp}$
            $G_{temp} \leftarrow G'$
        Update grid that centered at $G'$
            $G \leftarrow Matrix(G)$
    end for
    Get the best model $G_{best}$
        $G_{best} \leftarrow G'$
    Initialize Explanation $E$
        $E \leftarrow G_{best}$
    for $i \leftarrow 1$ to $I$ do
        Sample dataset $D$
            $D' \leftarrow Split(D, S)$
        Train the model $G_best$
            $G' \leftarrow Fit(G_{best}, D')$
        if $G'_{score} > E_{score}$ then
                $E \leftarrow G'$
        end if
    end for
    Prune selected tree $E \leftarrow Prune(E)$
    return $E$
end procedure
```

Figure 2: Global Explanation Algorithm

Conclusion

- For maliciously predicted traffic, tell the network operators about the features and corresponding values that might cause the prediction result based on the distribution
- Provide a improved formula as the firewall rule based on the data

Future Work

- Self-update the formula for the firewall rule based on the new incoming traffics
- Implement the user interface