International Journal of Safety and Security Engineering
Vol. 16, No. 2, February, 2026, pp. 393-402
Journal homepage: http://iieta.org/journals/ijsse

Comparative Performance of Class Imbalance Treatments in Random Forest for Social Engineering Classification

Ratih Hafsarah Maharrani[1*], Laura Sari[2], Oman Somantri[1]

[1] Cyber Security Engineering Department, Cilacap State Polytechnic, Cilacap 53212, Indonesia
[2] Informatics Engineering information, Cilacap State Polytechnic, Cilacap 53212, Indonesia

Corresponding Author Email: ratih.hafsarah@pnc.ac.id

https://doi.org/10.18280/ijsse.160214

Received: 10 November 2025
Revised: 1 February 2026
Accepted: 19 February 2026
Available online: 28 February 2026

Keywords:
social engineering, critical infrastructure, random forest, class imbalance, synthetic minority over-sampling technique

ABSTRACT

Social engineering (SE) attacks remain a significant threat to critical infrastructure (CI), yet their detection in real-world datasets is inherently challenging due to class imbalance, reporting bias, and the risk of label leakage. This study develops a leak-free benchmarking protocol for SE detection using the European Repository of Cyber Incidents (EuRepoC) dataset. A structured preprocessing pipeline is applied to ensure data consistency, prevent leakage, and focus specifically on CI-relevant incidents. Three random forest (RF)-based strategies are evaluated: class-weighted RF, RF with SMOTE, and balanced random forest (BRF). Experiments are conducted using a stratified train, test split with SMOTE applied exclusively to the training data. The models achieve moderate but consistent performance, with Precision–Recall – AUC (PR-AUC) around 0.73, Receiver Operating Characteristic – Area Under Curve (ROC-AUC) around 0.67, and F1-scores between 0.63 and 0.70 at a fixed threshold of 0.50. Threshold sensitivity analysis shows that lower thresholds improve recall at the cost of increased false positives, whereas the default threshold provides a more balanced trade-off suitable for operational settings. Robustness experiments across multiple test sizes and random seeds indicate stable model behavior, suggesting that the proposed pipeline is not overly sensitive to data partitioning. Overall, the findings emphasize the importance of leak-free evaluation, careful threshold selection, and stability analysis in SE detection for CI environments. The proposed framework provides a reproducible and realistic baseline for future cybersecurity research.

1. INTRODUCTION

Critical infrastructure (CI) forms the backbone of modern public services and economic activity, encompassing the energy, telecommunications, transportation, healthcare, and other essential service sectors. Although increased digitalization through the integration of information technology and operational technology (IT/OT) enhances operational efficiency, it simultaneously expands the attack surface that can be exploited by threat actors, particularly through social engineering (SE) [1, 2]. SE refers to psychological manipulation techniques used to gain unauthorized access to systems, information, or physical facilities by exploiting human vulnerabilities distinct from purely technical attacks that rely on exploiting hardware or software weaknesses [3, 4]. In the CI context, SE attacks such as spear-phishing, pretexting, baiting, or tailgating can have severe consequences, ranging from operational disruption to public safety risks [5]. The 2021 Colonial Pipeline incident illustrates how a credential-based compromise can trigger large-scale disruption, highlighting the need for robust and reliable SE detection mechanisms [5].

Recent advancements in attack techniques, including the use of artificial intelligence technologies such as deepfakes and social media analytics have made SE vectors increasingly personalized and more difficult to detect [1, 3, 4]. However, research on SE detection in CI remains limited, primarily due to the lack of representative datasets. National security sensitivities often make organizations reluctant to disclose incident data [6, 7], forcing researchers to rely on generic public datasets that do not adequately capture the characteristics of SE attacks in CI environments. The European Repository of Cyber Incidents (EuRepoC) provides a structured alternative with standardized taxonomies and multi-sector coverage. Nevertheless, EuRepoC is not without limitations, including underreporting, media bias toward high-profile incidents, and variability in documentation detail, all of which may affect model generalization [8].

Beyond data availability, machine learning-based SE detection also faces methodological challenges related to label construction, feature leakage, and evaluation bias. In particular, rule-based label definitions derived from incident descriptors may introduce semantic overlap with input features, potentially leading to overly optimistic performance estimates if leakage is not carefully controlled [9, 10]. Furthermore, the distribution of SE and non-SE incidents may vary depending on dataset filtering and labeling strategies, rather than being inherently imbalanced, which necessitates

empirical validation instead of assumptions. To address these challenges, prior work has proposed several imbalance-handling techniques, including class weighting, synthetic oversampling methods such as the synthetic minority over-sampling technique (SMOTE) [11], and Balanced Random Forest (BRF), which integrates resampling within the ensemble learning process [12]. However, the effectiveness of these methods is highly dependent on preprocessing design, encoding strategies, and the evaluation protocol, particularly in heterogeneous tabular datasets such as EuRepoC.

Evaluation metrics also play a critical role in imbalanced cybersecurity classification tasks. While Receiver Operating Characteristic – Area Under Curve (ROC-AUC) is widely used, it may provide overly optimistic assessments under skewed class distributions. Precision–Recall (PR) analysis has been shown to provide a more informative evaluation in such settings, as it directly captures the trade-off between false positives and false negatives [13]. Consequently, PR-AUC and F1-score are more appropriate for assessing SE detection performance in CI contexts, where missed attacks (false negatives) can have severe operational consequences.

This study addresses these challenges by proposing a leak-free and reproducible benchmarking framework for evaluating Random Forest (RF)-based approaches to SE detection in CI using the EuRepoC dataset. The proposed pipeline includes deterministic preprocessing, explicit rule-based label construction, and strict feature selection to prevent leakage from post-incident attributes. Unlike conventional approaches that rely solely on threshold optimization, this study distinguishes between main performance evaluation using a fixed decision threshold and sensitivity analysis across a threshold grid, allowing for a more transparent interpretation of model behavior. The evaluation protocol further incorporates robustness analysis across multiple test splits and random seeds. Three imbalance-handling strategies are compared: class-weighted RF with SMOTE, and BRF, to analyze their behavior under realistic data conditions.

Rather than emphasizing absolute performance gains, this study focuses on providing a systematic and transparent evaluation of model behavior under varying data and decision conditions. In particular, we examine precision–recall trade-offs, threshold sensitivity, and performance stability across different data partitions. By explicitly addressing leakage risks, evaluation bias, and dataset limitations, this work contributes a reproducible and operationally grounded framework for SE detection in CI environments, supporting both research reproducibility and practical deployment considerations.

2. METHODOLOGY

2.1 Literature review

Research on cyberattack detection in CI has historically focused on technical intrusion vectors, including distributed denial-of-service (DDoS) and software vulnerability exploitation. However, recent developments indicate a substantive shift toward SE, where adversaries manipulate human psychology rather than exploiting system vulnerabilities [3]. Although increasingly relevant, machine learning-based SE detection in CI remains underexplored. Prior studies commonly rely on general-purpose datasets such as NSL-KDD and UNSW-NB15, which do not reflect the behavioral characteristics or attack patterns of SE incidents within real-world CI environments [14, 15]. As a result, a substantial gap remains between contemporary threat patterns and existing detection capabilities.

A central methodological challenge in cybersecurity datasets is class imbalance, where malicious incidents represent only a small fraction of the data. Numerous efforts have attempted to address this issue using RF in combination with resampling techniques such as the Synthetic Minority Oversampling Technique (SMOTE). Abdelhamid and Desai [16], for example, demonstrated that SMOTE improves the performance of several ensemble models on the BoT-IoT dataset; however, the study remains focused on IoT botnets and does not examine model stability under varying imbalance conditions.

Bagui et al. [17] further investigated the configuration of SMOTE on UNSW-NB15, identifying that a modest oversampling ratio (10%) and k-values between 3–5 yield optimal performance. Importantly, they highlight that excessive oversampling can induce overfitting, suggesting that SMOTE cannot be applied uniformly across domains without systematic validation. Khan et al. [18] compared multiple resampling techniques, including SMOTE, ADASYN, and random undersampling on RF models, finding that SMOTE offers the most consistent improvement in minority-class detection. Nevertheless, these approaches are primarily evaluated on numerical or low-cardinality datasets and do not address challenges associated with high-dimensional categorical features, such as those present in EuRepoC.

More advanced ensemble approaches have been proposed to improve robustness. Ismail et al. [19] introduced the Weighted Score Selector (WSS) for Wireless Sensor Networks, achieving lower false alarm rates and faster processing than traditional RF. However, this technique is domain-specific and not suited to incident-level analysis found in EuRepoC. Ahmad et al. [20] combined RF, Gradient Boosting, and Neural Networks within a stacking ensemble enhanced by SMOTE-ENN to detect multi-class attacks on NSL-KDD and CICIDS2017. Although this method improved classification performance, it demands large datasets and high computational resources, making it less suitable for EuRepoC, which features a moderate dataset size, heterogeneous attributes, and operational requirements for lightweight and interpretable models suitable for SIEM/SOC integration.

From these studies, several key limitations emerge. First, existing research relies heavily on outdated or generic datasets that do not capture SE behavior within CI environments, leading to limited real-world relevance. Second, none of the reviewed works explicitly adopts a leakage-aware or leak-free experimental design, even though data leakage can significantly distort performance estimation. Third, no standardized evaluation protocol systematically examines model stability under variations in threshold selection, test-set size, or random seed. Fourth, prior studies rarely address practical considerations such as alert fatigue caused by false positives or the operational risks associated with false negatives in CI settings. Finally, limited attention has been given to the appropriateness of evaluation metrics, particularly the role of Precision–Recall analysis in imbalanced cybersecurity tasks.

Given these limitations, this study positions itself not as an introduction of new algorithms but as a provider of systematic, reproducible, and leakage-aware benchmarking framework for detecting SE incidents in CI using the EuRepoC dataset.

Unlike previous works, this study constructs an explicit taxonomy-aligned preprocessing pipeline, builds the *is_SE* label from documented rules, and applies feature selection designed specifically to prevent data leakage. Furthermore, this research conducts empirical stability assessments through threshold sensitivity analysis, test-set proportion variation, and multi-seed evaluation, offering deeper insight into the robustness of RF-based approaches under realistic imbalance conditions.

Finally, this study provides operational interpretation of precision–recall trade-offs, directly relevant to SOC/SIEM environments, including how minimizing false positives helps prevent alert fatigue, and how reducing false negatives mitigates the risk of missed SE incidents in CI operations. By evaluating three RF variants, class_weight, RF+SMOTE, and BRF, this work identifies the most stable and computationally efficient approach for CI use cases. The proposed framework thus fills a notable gap in the literature by offering a reproducible and empirically grounded evaluation protocol that bridges academic findings and real-world requirements in SE detection for CI environments.

2.2 Methods

This section presents a leakage-aware, reproducible, and systematically controlled methodological pipeline designed to benchmark RF-based approaches for SE detection in CI. The workflow follows a strictly controlled sequence comprising dataset preprocessing, label construction, leakage prevention, modeling, evaluation, and robustness testing, enabling independent replication and ensuring reliable performance estimation. Figure 1 provides an overview of the stepwise process.

2.2.1 Dataset description

This study employs the EuRepoC, a structured and curated cyber-incident database that documents attacker profiles, initial access vectors, sectoral targets, and operational impacts across European Union member states. The subset used in this research focuses exclusively on incidents associated with the CI sectors, as identified and filtered using EuRepoC's official codebook and sector taxonomy.

The dataset comprises heterogeneous variables, including categorical, ordinal, and multi-level attributes, such as incident_type, receiver_category, initiator_category, MITRE initial_access, functional_impact_code, and related attributes. While this richness makes EuRepoC highly suitable for evaluating machine-learning models on tabular data, it also introduces several methodological challenges:

(1) potential class distribution imbalance depending on label construction and filtering strategy;
(2) underreporting and media-driven bias, as EuRepoC relies on publicly accessible incident disclosures; and
(3) inconsistent documentation granularity across CI sectors, which may hinder the model's ability to generalize to unseen or emerging patterns.

Additionally, EuRepoC contains heterogeneous categorical representations, missing values, and varying temporal coverage, which require explicit preprocessing and careful evaluation design. These limitations necessitate a rigorously designed preprocessing workflow and a fully leak-free evaluation protocol to avoid inflated performance estimates and to ensure reproducibility. Although EuRepoC is one of the most comprehensive cyber-incident repositories currently available, its structural constraints, such as noisy entries, incomplete metadata, and reporting inconsistencies, require systematic data cleaning, taxonomy harmonization, and controlled validation procedures. Accordingly, the methodology in this study incorporates deterministic preprocessing rules, explicit leakage-prevention mechanisms, and robustness analyses to mitigate these challenges and strengthen the external validity of the findings.

Figure 1. Proposed method framework

2.2.2 Data preprocessing and leak-free label construction

A strict, sequential, and leak-free preprocessing protocol was implemented to prevent inflated performance estimates, particularly those caused by semantic overlap between features and label definitions, an issue often overlooked in prior SE detection studies. All preprocessing steps were performed before train–test splitting, except for oversampling, which was restricted to the training partition to avoid data

leakage.

(1) Canonization and Normalization

All categorical variables were normalized using deterministic mapping rules derived from the official EuRepoC codebook. This process ensured canonical consistency across:

1. receiver_category and receiver_subcategory,
2. incident_type (including explicit preservation of the doxing tag),
3. initiator_category, and
4. MITRE initial_access and impact labels.

Normalization removed spelling inconsistencies, merged semantically equivalent categories, standardized hierarchical taxonomies, and eliminated redundant or ambiguous entries.

(2) Ordinal Transformation

Several EuRepoC attributes represent ordered impact scales rather than nominal categories. To preserve their ordinal semantics, the following fields were transformed into numeric values using rule-based mappings:

1. data_theft_code,
2. disruption_code,
3. hijacking_code, and
4. functional_impact_code.

This transformation ensures compatibility with RF-based models while maintaining the internal progression of severity levels.

(3) Leak-Free Construction of the is_SE Label

The target variable is_SE was constructed using explicit rules tied strictly to incident-stage or pre-incident indicators, avoiding reliance on post-incident outcomes. A record was labeled SE only if one or more of the following criteria were satisfied:

1. MITRE initial_access indicates phishing or quishing vectors.
2. incident_type includes doxing or hijacking with misuse.
3. user_interaction = required, implying a human-driven exploitation pathway.

To reduce optimistic bias, ambiguous or weakly supported cases were conservatively assigned to the non-SE class. This conservative labeling strategy reflects realistic operational constraints, where complete ground truth is rarely available.

(4) Stratified Train–Test Split

To maintain proportional representation of SE and non-SE cases, a stratified hold-out split was applied using five test-set proportions:

test_size ∈ {0.10, 0.15, 0.20, 0.25, 0.30}

Each configuration was repeated over 30 random seeds, enabling variance estimation and robustness analysis across different data partitions. All steps of preprocessing, including cleaning, canonicalization, feature engineering, and label construction, were carried out prior to splitting. SMOTE was applied only to the training set, strictly preventing information leakage into the testing process.

2.2.3 Feature selection and leakage prevention

Feature selection followed an explicit leak-prevention policy to ensure that no variable containing post-incident, outcome-related, or label-defining information could influence model predictions. In particular, features directly involved in label construction—such as incident_type, MITRE initial_access, and user_interaction—were excluded to prevent semantic leakage between input features and the target variable.

Features reflecting outcomes, impacts, forensic findings, or incident resolution were also removed, as they would not be available at real-time detection points in CI operations. The remaining features consisted exclusively of attributes observable before or at the moment of incident initiation, ensuring operational validity.

A complete list of excluded leakage-prone features is provided in the supplementary material to support reproducibility and transparency.

2.2.4 Modeling approaches

Three modeling approaches were evaluated, all based on the RF architecture due to its robustness, suitability for heterogeneous tabular data, and interpretability advantages [14]. RF combines predictions from multiple decision trees built using bootstrap samples and random feature subsets; final predictions are obtained via majority voting for classification tasks [14, 15].

(1) RF with class weighting

The class_weight mechanism applies a larger penalty to misclassification of minority-class samples, making the model inherently more sensitive to SE incidents [16, 21]. This approach requires no synthetic data and is computationally efficient for SOC/SIEM deployment.

(2) SMOTE + RF

The SMOTE addresses class imbalance by generating synthetic samples through linear interpolation between a minority sample and its nearest neighbors [22]. The generation process is expressed as Eq. (1):

$$Y' = Y^i + (Y^j - Y^i) * \gamma \quad (1)$$

where, Y^i represents a minority class sample, Y^j is a randomly selected neighbor, and γ is a random value in the range [0, 1] [23]. RF was then trained on the oversampled training set, leveraging its robustness for complex and high-dimensional data [24]. In this study, SMOTE is applied after categorical encoding, which may introduce limitations due to interpolation in high-dimensional feature space. This aspect is explicitly considered in the analysis of synthetic sample validity.

(3) BRF

BRF performs bootstrap sampling with internal undersampling of the majority class in each tree, ensuring balanced training subsets and reducing variance under severe imbalance conditions [12, 25, 26].

2.2.5 Evaluation protocol, threshold tuning, and robustness analysis

(1) Fixed threshold and sensitivity analysis

The primary evaluation is conducted using a fixed classification threshold of 0.50 to ensure comparability and avoid bias from threshold optimization. Additionally, a sensitivity analysis is performed using a threshold grid in the range [0.30, 0.70] with a step size of 0.01 to examine the stability of model performance across decision thresholds [27].

(2) Evaluation metrics

To evaluate model performance under class imbalance, this study employs F1-score, ROC-AUC, and PR-AUC. Precision, recall, and F1-score. Let TP, TN, FP, and FN denote True Positive, True Negative, False Positive, and False Negative, respectively. Precision, recall, and the F1-score are defined as shown in Eqs. (2)-(4) [28].

$$Precission = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - score = 2 \times \frac{Precission \times Recall}{Precission + Recall} \quad (4)$$

The F1-score is emphasized over accuracy because it balances false positives and false negatives. Accuracy is reported only as a complementary metric. PR-AUC is particularly emphasized, as it provides a more informative evaluation than ROC-AUC under imbalanced conditions, capturing the trade-off between precision and recall [27]. It is included for completeness and is defined in Eq. (5):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100\% \quad (5)$$

The ROC-AUC quantifies the model's discriminative capability by evaluating the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) [27, 29]. Mathematically, ROC-AUC is expressed as Eq. (6):

$$ROC - AUC = \int_0^1 TPR(FPR^{-1}(x))dx \quad (6)$$

For highly imbalanced datasets—such as SE vs. non-SE classification, the PR-AUC (Precision–Recall AUC) provides a more representative assessment of minority-class performance [30, 31]. PR-AUC can be defined as Eq. (7):

$$\theta = \int_{-\infty}^{\infty} Prec(c)dP(Y \leq c) \quad (7)$$

or equivalently in Eq. (8),

$$\theta = \int_{-\infty}^{\infty} P(D = 1 \mid Z > c)dP(Y \leq c) \quad (8)$$

with the probabilistic relationship between the positive and negative classes (Eq. (9)) defined as:

$$P(D = 1|Z > c) = \frac{\pi P(Y > c)}{\pi P(Y > c) + (1 - \pi)P(X > c)} \quad (9)$$

where, $\pi = P(D = 1)$ denotes the prior probability of the positive class [32].

(3) Robustness analysis

A comprehensive robustness analysis was performed to evaluate model stability under different data conditions. This included:

1. variations in test size,
2. repetitions across 30 random seeds,
3. positive sample adequacy checks (ensuring ≥ 60 SE samples per test fold), and
4. analysis of performance variance across splits.

These procedures ensure that performance is not inflated due to dataset leakage or favorable random partitions, aligning with the reviewer's request for a reproducible and leak-free experimental design.

Implementation details are summarized as follows. All categorical features were encoded using one-hot encoding prior to model training. To address class imbalance, SMOTE was applied exclusively to the training data after encoding, ensuring valid synthetic sample generation. The RF models were configured with 500 trees (n_estimators = 500) and class_weight = balanced, while SMOTE used k_neighbors = 5. The BRF implementation follows the imbalanced-learn library defaults. Finally, all features involved in label construction were explicitly excluded from the model to ensure a leak-free setup.

3. RESULT AND DISCUSSION

3.1 Initial dataset

The dataset employed in this study is the EuRepoC, a curated repository that compiles reports of cyber incidents across European Union member states and associated regions [33]. EuRepoC provides labeled incident records that have been used in prior work to analyze attack patterns and to build machine learning-based detection models, particularly in the domains of phishing and SE [34, 35]. At the same time, EuRepoC exhibits several structural limitations that are important for the interpretation of our results. First, incident reporting is voluntary and partly media-driven, which leads to under-reporting of less visible events and a bias toward high-profile cases [35]. Second, attacker behavior evolves rapidly, especially for variants such as targeted spear-phishing and multi-stage SE, so static labels may not fully capture current tactics. Third, the level of detail differs across entries, with some incidents richly annotated and others described only in coarse terms.

In addition to these known limitations, a systematic dataset characterization was performed to assess feature completeness, data types, and variability across records. The dataset contains both structured and semi-structured attributes, including categorical variables (e.g., incident type, country), temporal fields (e.g., start date), and descriptive metadata. Several features exhibit missing values with varying degrees of sparsity, reflecting inconsistent reporting practices across incidents. This heterogeneity introduces potential bias if not handled carefully during preprocessing. These factors imply that a model trained on EuRepoC CI incidents may overfit to well-documented patterns and be less sensitive to under-reported or covert SE campaigns.

To mitigate these issues, we apply a structured preprocessing pipeline, harmonize the incident taxonomy, and restrict our analysis to features that are directly relevant to SE detection in CI. Political indicators, state-actor attributes, and broader cyber-conflict variables are excluded to reduce confounding effects [36].

Furthermore, temporal attributes are treated with caution to avoid potential data leakage. Although time-related fields are available in the dataset, they are not used in a way that would allow future information to influence model training. Instead, they are leveraged only for consistency checks and, where applicable, for temporal validation strategies. Feature selection is additionally guided by data quality and completeness considerations, ensuring that variables with excessive missingness or ambiguous definitions do not introduce instability into the learning process.

A high-level overview of the EuRepoC schema and the subset of variables used in this work is shown in Figure 2. A

more detailed discussion of residual biases and their implications for deployment in real CI environments is provided in the Discussion and Limitations section. The dataset consists of 3,414 records and 85 features, including categorical, temporal, and textual attributes. Several features contain missing values with varying completeness levels.

```
RangeIndex: 3414 entries, 0 to 3413
Data columns (total 10 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   incident_id                 3414 non-null   object
 1   name                        3414 non-null   object
 2   description                 3414 non-null   object
 3   start_date                  3322 non-null   object
 4   end_date                    1541 non-null   object
 5   inclusion_criterion         3402 non-null   object
 6   inclusion_criterion_subcode 1475 non-null   object
 7   source_disclosure           3348 non-null   object
 8   incident_type               3411 non-null   object
 9   receiver_name               2222 non-null   object
dtypes: object(10)
memory usage: 266.8+ KB
```

Figure 2. Sample initial dataset structure

3.2 Pre-processing stage

Prior to the modeling stage, the EuRepoC subset underwent a leak-free preprocessing pipeline designed to ensure data quality and consistency. The process consists of five stages: data cleaning and canonization, ordinal mapping, construction of the *is_SE* label, leakage-aware feature selection, and stratified train–test splitting.

In the cleaning stage, key categorical variables (e.g., *receiver_category*, *incident_type*, and MITRE attributes) were normalized based on the EuRepoC codebook to resolve inconsistencies and harmonize labels. This step focuses solely on standardization and does not remove any records, resulting in no change in dataset size. Ordinal mapping then converts impact-related variables into ordered numerical representations while preserving their severity structure.

The binary target label *is_SE* is defined using explicit rules based on phishing-related access vectors, incident types, and user interaction. To avoid optimistic bias, incomplete or ambiguous cases are conservatively labeled as non-SE.

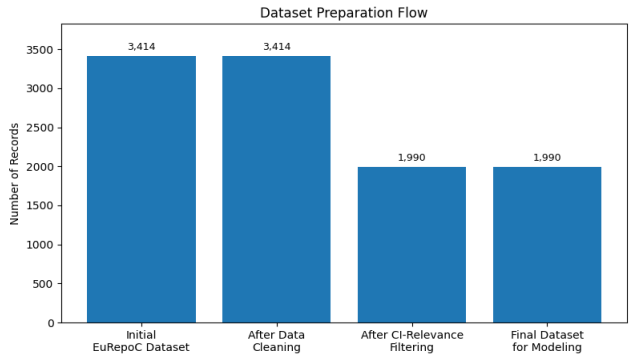


Figure 3. Dataset flow

A domain-specific filtering step is subsequently applied to retain only CI related incidents, reducing the dataset from 3,414 to 1,990 records. Feature selection is then performed to remove leakage-prone variables, without affecting the number of samples.

Finally, the dataset is split into training and test sets using stratified sampling. Overall, the preprocessing pipeline introduces a single major reduction step at the CI filtering stage, as illustrated in Figure 3.

3.3 Handling class imbalance

After preprocessing and CI filtering, the dataset shows an imbalanced distribution between SE and Non-SE incidents. The full dataset contains 2,092 SE and 1,322 Non-SE cases, indicating a higher prevalence of SE incidents in the selected subset.

To preserve this distribution, a stratified hold-out split is applied. This results in 1,673 SE and 1,058 Non-SE samples in the training set, and 419 SE and 264 Non-SE samples in the test set, maintaining consistent class proportions across both subsets.

To improve the model's ability to capture minority patterns, the SMOTE is applied exclusively to the training data. SMOTE generates synthetic samples by interpolating between neighboring instances in the feature space, which has been widely used to enhance minority-class learning in imbalanced cybersecurity datasets [22-24]. After resampling, the training set becomes balanced at 1,673 SE and 1,673 Non-SE samples. The test set is intentionally left unchanged to ensure realistic evaluation and to avoid information leakage. The distribution of SE and Non-SE instances across the full dataset, the training set (before and after SMOTE), and the test set is illustrated in Figure 4.

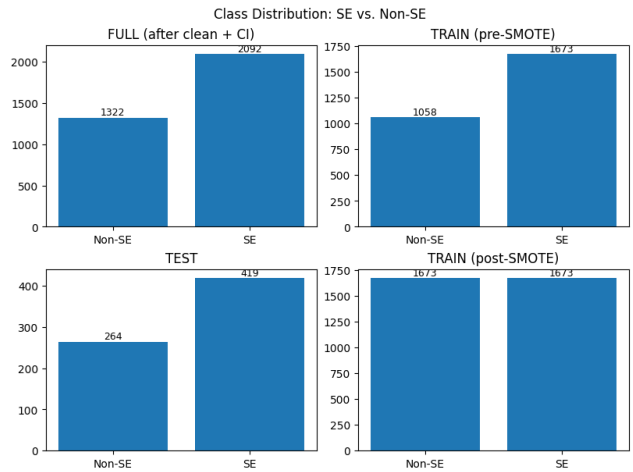


Figure 4. Distribution of FULL classes, Train pre-SMOTE, Test, Train post-SMOTE

3.4 Modeling

The modeling stage, illustrated in Figure 5, evaluates three RF-based approaches for handling class imbalance: (i) RF (class_weight), RF with class weighting, where higher penalties are assigned to minority-class errors [16, 21]; (ii) RF + SMOTE, where the model is trained on a balanced training set; (iii) BRF, which applies random undersampling within each tree to maintain class balance.

All models are trained using the same leak-free feature set and evaluated on the untouched test set, with a fixed decision threshold of 0.50 to ensure a fair comparison. RF is selected due to its robustness on heterogeneous tabular data and its ability to reduce overfitting through ensemble averaging [14]. The final prediction is obtained through majority voting across decision trees [26].

$$RF_{(x)} = \frac{1}{T}\ \sigma\ _{n\ =\ 1}^{T} C_n(x) \quad (10)$$

Figure 5 presents the confusion matrices for all models. The results highlight distinct trade-offs between the three approaches. The RF + SMOTE model improves the detection of SE incidents, achieving the highest number of true positives (292), but at the cost of increased false positives (119). In contrast, the BRF produces fewer false positives (71) but also misses more SE cases, as reflected in its higher number of false negatives (191). The standard RF with class weighting shows a more balanced behavior, but with comparatively lower detection performance.

Rather than emphasizing a single best model, these results illustrate the practical trade-off between recall and precision when handling imbalanced cybersecurity data. This comparison provides insight into how different balancing strategies affect detection performance under consistent evaluation settings.

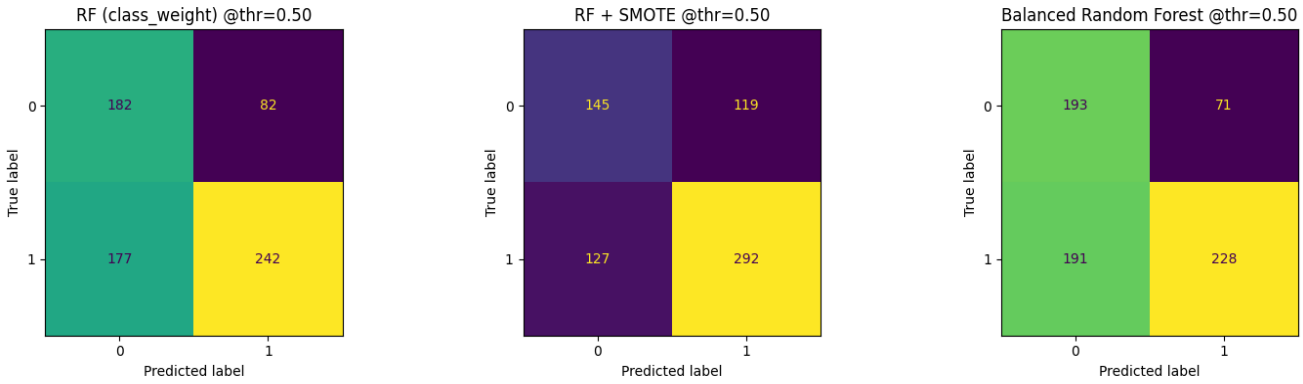
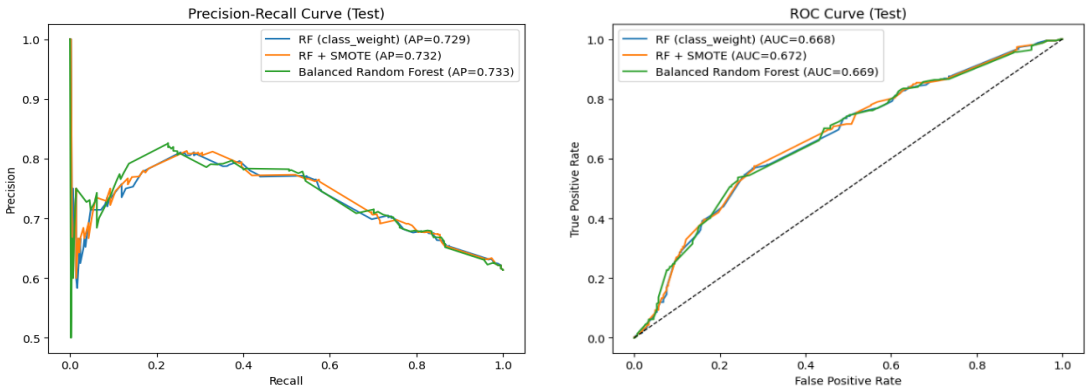


Figure 5. Confusion matrix

Table 1. Evaluation of the proposed model

Model	PR-AUC	ROC-AUC	F1-Score	Accuracy	Precision	Recall
BRF	0.733	0.669	0.635	0.616	0.763	0.544
RF + SMOTE	0.732	0.672	0.704	0.640	0.710	0.697
RF (class_weight)	0.729	0.668	0.651	0.621	0.747	0.578



(a) (b)

Figure 6. Performance of evaluation model

3.5 ROC and PR-AUC curve analysis

Based on Table 1 and Figures 6(a) and (b), all three models demonstrate moderate but consistent discrimination capability, with PR-AUC values around 0.73 and ROC-AUC values around 0.67. While these values are lower than those typically reported in more controlled datasets, they reflect a more realistic evaluation setting using leak-free features and a fixed decision threshold. This indicates that the RF-based approaches are still able to distinguish SE from Non-SE incidents, albeit under more challenging and practical conditions [30, 31]. Among the three strategies, Among the three strategies, RF + SMOTE achieves the highest F1-score (0.704) at the fixed threshold of 0.50, indicating a better balance between precision and recall. In contrast, BRF exhibits higher precision (0.763) but lower recall (0.544), suggesting a more conservative behavior that reduces false positives at the expense of missing more SE incidents. The class-weighted RF lies between these two approaches, providing moderate precision and recall without additional resampling.

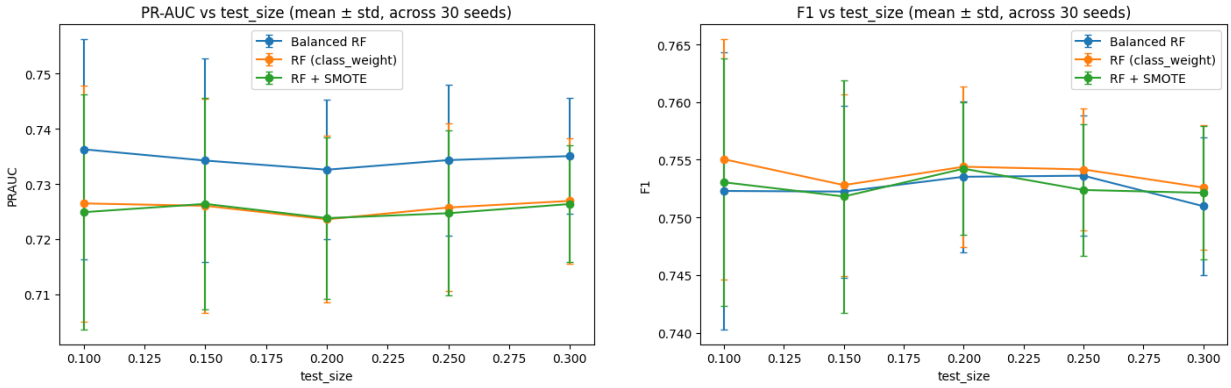
The Precision–Recall curves, while not concentrated in the extreme upper-right region, still show a stable trade-off between precision and recall across thresholds, whereas the ROC curves remain consistently above the diagonal baseline, indicating meaningful class separability. These results highlight that model performance is influenced not only by the learning algorithm but also by the choice of decision threshold

and imbalance-handling strategy.

From an operational CI perspective, these findings emphasize the importance of selecting models and thresholds based on deployment priorities. Models such as RF + SMOTE are more suitable when maximizing detection (recall) is critical, while BRF may be preferred when reducing false alarms (precision) is more important. This trade-off is particularly relevant in SOC/SIEM environments, where both missed incidents and excessive alerts can significantly impact operational effectiveness.

3.6 Robustness analysis

To characterize model stability, a key part of our contribution, we conducted a robustness analysis by varying the test set proportion from 10% to 30% and repeating the stratified split over multiple random seeds. Figure 7(a) and (b) show the mean and standard deviation of PR-AUC and F1-score across these configurations for RF (class_weight), RF + SMOTE, and BRF.



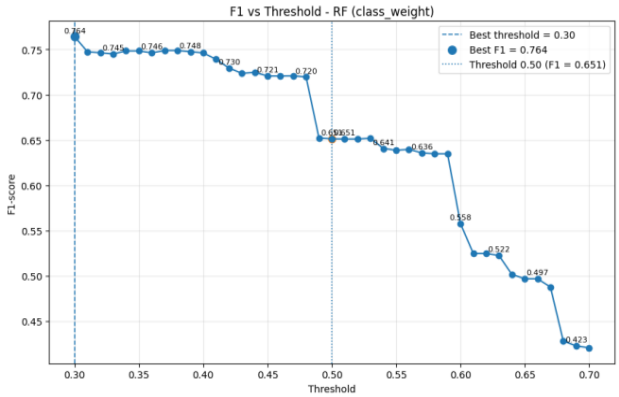
(a) (b)

Figure 7. Robustness analysis of test set variations on PR-AUC and F1-score

Overall, all three models show stable performance across different test sizes, with PR-AUC values consistently around 0.72–0.74 and F1-scores in the range of 0.75–0.76. The relatively small variation across configurations indicates that model performance is not highly sensitive to the choice of train–test split, suggesting reliable generalization within the EuRepoC CI subset.

Differences among the models are modest. BRF tends to achieve slightly higher PR-AUC values and exhibits relatively narrower error bars, indicating more stable performance across random splits. Meanwhile, RF (class_weight) and RF + SMOTE achieve comparable F1-scores, reflecting similar effectiveness in balancing precision and recall under varying data partitions.

From a practical standpoint, test sizes between 0.20 and 0.25 provide a reasonable balance, as they maintain stable evaluation metrics while ensuring a sufficient number of positive samples in the test set. This robustness analysis complements the single-split evaluation and provides additional evidence that the proposed RF-based approaches remain consistent under different data partitioning scenarios.



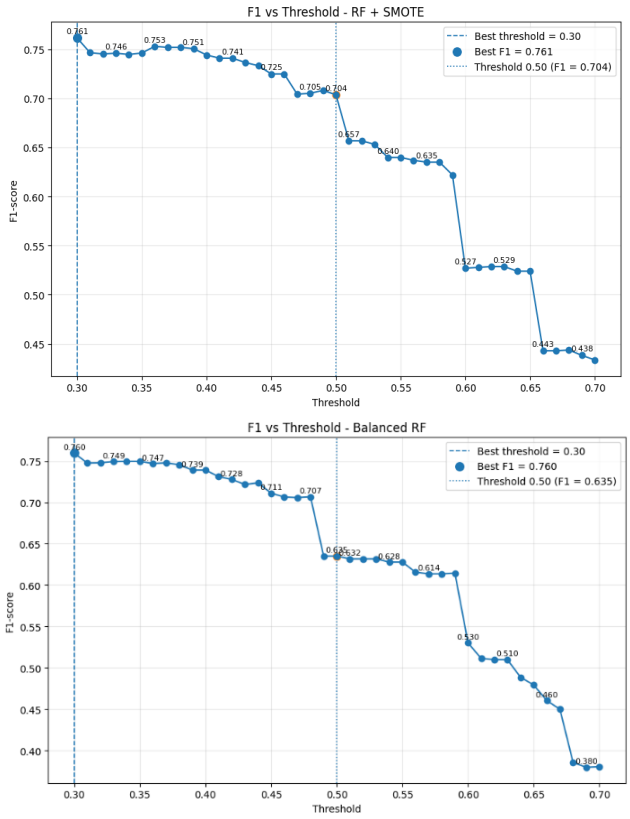


Figure 8. Threshold sensitivity analysis showing precision, recall, and F1-score across thresholds

3.7 Threshold sensitivity analysis

To further understand how the decision threshold affects model behavior, Figure 8 illustrates the relationship between precision, recall, and F1-score across thresholds ranging from 0.30 to 0.70 for all three models. Across all models, a

consistent pattern can be observed. At lower thresholds (around 0.30), recall is high, indicating that most SE incidents are successfully detected. However, this comes at the cost of lower precision, meaning more false positives are generated. As the threshold increases, precision gradually improves, while recall declines, reflecting a more conservative prediction strategy.

The F1-score reaches its maximum at lower thresholds (approximately 0.30), where the balance between precision and recall is optimized. For instance, RF + SMOTE achieves its highest F1-score of approximately 0.761 at this point. However, when using the fixed threshold of 0.50, the F1-score decreases (e.g., 0.704 for RF + SMOTE), reflecting a trade-off toward more balanced and realistic decision-making.

These findings highlight that model performance is highly dependent on the chosen threshold. While lower thresholds may maximize detection capability, they may not be suitable in operational environments due to increased false alarms. Conversely, a fixed threshold of 0.50 provides a more stable and interpretable operating point, aligning better with practical deployment scenarios.

4. CONCLUSIONS

This study develops a leak-free benchmarking protocol for detecting SE attacks in CI using the EuRepoC dataset. Three RF-based strategies: class-weighted RF, RF + SMOTE, and BRF are systematically compared in terms of performance, stability, and threshold behavior.

The models achieve moderate but consistent performance (PR-AUC ≈ 0.73, ROC-AUC ≈ 0.67, and F1 ≈ 0.63–0.70 at the fixed threshold of 0.50), indicating that SE detection in real-world CI datasets remains a challenging task. RF + SMOTE provides the highest F1-score at the default threshold, while BRF demonstrates more stable performance across different data splits, although performance differences between models are relatively small.

Threshold sensitivity analysis shows that lower thresholds (~ 0.30) increase recall at the expense of precision, whereas the default threshold (0.50) offers a more balanced and operationally practical trade-off. Robustness experiments across multiple test sizes and random seeds confirm that the evaluation pipeline remains stable and is not overly sensitive to data partitioning.

At the data level, the primary reduction occurs during CI-relevance filtering (from 3,414 to 1,990 records), while earlier preprocessing stages preserve all observations. The results should therefore be interpreted as realistic estimates under the constraints and biases of the EuRepoC dataset.

From a practical perspective, RF + SMOTE provides a balanced baseline when recall is prioritized, class-weighted RF offers a simpler and computationally efficient alternative, and BRF is preferable when stability is a key concern. Future work will extend this framework with additional models and temporal validation.

REFERENCES

[1] Adil, M.U., Ali, S., Haider, A., Javed, M.A., Khan, H. (2024). An enhanced analysis of social engineering in cyber security research challenges, countermeasures: A survey. The Asian Bulletin of Big Data Management, 4(4): 321-331. https://doi.org/10.62019/abbdm.v4i4.274

[2] Márton, Z., Rajnai, Z. (2024). The evolution and future of social engineering: Exploiting psychological vulnerabilities in the digital age [in Hungary]. Safety and Security Sciences Review, 6(4): 45-56. https://doi.org/10.12700/btsz.2024.6.4.45

[3] Sudha, R.P., Mahmood, A.H. (2024). Human factor in cybersecurity: Behavioral insights into phishing and social engineering attacks. Nanotechnology, 20(S15): 630-642. https://doi.org/10.62441/nano-ntp.vi.3556.

[4] Chowdhury, R.H., Mostafa, B. (2025). Cyber-physical systems for critical infrastructure protection: Developing advanced systems to secure energy grids, transportation networks, and water systems from cyber threats. Journal of Computer Science and Electrical Engineering, 7(1): 16-26. https://doi.org/10.61784/jcsee3027

[5] Mittal, M. (2024). Colonial pipeline cyberattack drives urgent reforms in cybersecurity and critical infrastructure resilience. International Journal of Oil, Gas and Coal Engineering, 12(5): 106-119. https://doi.org/10.11648/j.ogce.20241205.11

[6] Al Mamun, A., Al-Sahaf, H., Welch, I., Barcellos, M., Camtepe, S. (2024). Limitations of advanced persistent threat datasets: Insights for cybersecurity research. In 2024 34th International Telecommunication Networks and Applications Conference (ITNAC), Sydney, Australia, pp. 1-8. https://doi.org/10.1109/ITNAC62915.2024.10815148

[7] Nguyen, K., Pal, S., Jadidi, Z., Dorri, A., Jurdak, R. (2022). A blockchain-enabled incentivised framework for cyber threat intelligence sharing in ICS. In 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops), Pisa, Italy, pp. 261-266. https://doi.org/10.1109/PerComWorkshops53856.2022.9767226

[8] Shandilya, V., Simmons, C.B., Shiva, S. (2014). Use of attack graphs in security systems. Journal of Computer Networks and Communications, 2014(1): 818957. https://doi.org/10.1155/2014/818957

[9] Munaye, Y.Y., Molla, A., Belayneh, Y., Simegnew, B. (2024). Long short-term memory and synthetic minority over sampling technique-based network traffic classification. In 2024 International Conference on Information and Communication Technology for Development for Africa (ICT4DA), Bahir Dar, Ethiopia, pp. 120-124. https://doi.org/10.1109/ICT4DA62874.2024.10777078

[10] Rana, S., Kanji, R., Jain, S. (2024). Comprehensive analysis of oversampling techniques for addressing class imbalance employing machine learning models. Recent Advances in Computer Science and Communications, 19(2). https://doi.org/10.2174/0126662558347788241127051934

[11] Prasetya, J., Abdurakhman, A. (2023). Comparison of smote random forest and smote k-nearest neighbors classification analysis on imbalanced data. Media Statistika, 15(2): 198-208. https://doi.org/10.14710/medstat.15.2.198-208

[12] Sanjay Kumar, N.V., Krishna, N., Patil, K.A., Joy, S., Chithra, R.B., Raghavendra Patil, G.E. (2024). Imbalance dataset handling for classification using machine learning algorithm. Nanotechnology

Perceptions, 20(7): 777-787. https://doi.org/10.62441/nano-ntp.v20i7.3941
[13] Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J., Ahmad, F. (2021). Network intrusion detection system: A systematic study of machine learning and deep learning approaches. Transactions on Emerging Telecommunications Technologies, 32(1): e4150. https://doi.org/10.1002/ett.4150
[14] Ngo, G., Beard, R., Chandra, R. (2022). Evolutionary bagging for ensemble learning. Neurocomputing, 510: 1-14. https://doi.org/10.1016/j.neucom.2022.08.055
[15] Yang, Z., Li, Z., Du, X., Wang, F., Qiu, Y., Wei, J. (2025). Intelligent prediction model of a polymer fracture grouting effect based on an information acquisition optimizer-optimized random forest. Results in Engineering, 28: 107382. https://doi.org/10.1016/j.rineng.2025.107382
[16] Abdelhamid, M., Desai, A. (2024). Balancing the scales: A comprehensive study on tackling class imbalance in binary classification. arXiv preprint arXiv:2409.19751. https://doi.org/10.48550/arXiv.2409.19751
[17] Bagui, S.S., Mink, D., Bagui, S.C., Subramaniam, S. (2023). Determining resampling ratios using bsmote and SVM-SMOTE for identifying rare attacks in imbalanced cybersecurity data. Computers, 12(10): 204. https://doi.org/10.3390/computers12100204
[18] Khan, S.A., Kanagaraj, V., Kaya, E.B., Rahman, M.S., Quinn, L., Aslan, S. (2024). Cyber-attack monitoring and detection using machine learning techniques. In 2024 IEEE Future Networks World Forum (FNWF), Dubai, United Arab Emirates, pp. 946-951. https://doi.org/10.1109/FNWF63303.2024.11028778
[19] Ismail, S., El Mrabet, Z., Reza, H. (2022). An ensemble-based machine learning approach for cyber-attacks detection in wireless sensor networks. Applied Sciences, 13(1): 30. https://doi.org/10.3390/app13010030
[20] Ahmad, Z., Shahid Khan, A., Nisar, K., Haider, I., et al. (2021). Anomaly detection using deep neural network for IoT architecture. Applied Sciences, 11(15): 7050. https://doi.org/10.3390/app11157050
[21] Shiksha. (2022). Application of machine learning algorithms with balancing techniques for credit card fraud detection: A comparative analysis. In Fundamentals and Methods of Machine and Deep Learning: Algorithms, Tools and Applications, pp. 277-309. https://doi.org/10.1002/9781119821908.ch12
[22] Hairani, H., Widiyaningtyas, T., Prasetya, D.D. (2024). Addressing class imbalance of health data: A systematic literature review on modified synthetic minority oversampling technique (SMOTE) strategies. International Journal on Informatics Visualization, 8(3): 1310-1318. https://doi.org/10.62527/joiv.8.3.2283
[23] Hairani, H., Anggrawan, A., Priyanto, D. (2023). Improvement performance of the random forest method on unbalanced diabetes data classification using Smote-Tomek Link. International Journal on Informatics Visualization, 7(1): 258-264. https://doi.org/10.30630/joiv.7.1.1069
[24] Fulazzaky, T., Saefuddin, A., Soleh, A.M. (2024). Evaluating ensemble learning techniques for class imbalance in machine learning: A comparative analysis of balanced random forest, SMOTE-RF, SMOTEBoost, and RUSBoost. Scientific Journal of Informatics, 11(4): 969-980. https://doi.org/10.15294/sji.v11i4.15937
[25] Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2): 123-140. https://doi.org/10.1007/BF00058655
[26] Suthaharan, S. (2016). Random forest learning. In Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, pp. 273-288. https://doi.org/10.1007/978-1-4899-7641-3_11
[27] Abdennebi, A., Morsli, R., Kara, N., Ould-Slimane, H. (2024). Machine learning and large language models-based techniques for cyber threat detection: A comparative study. In 2024 7th Conference on Cloud and Internet of Things (CIoT), Montreal, QC, Canada, pp. 1-9. https://doi.org/10.1109/CIoT63799.2024.10756998
[28] Shanmugam, V., Razavi-Far, R., Hallaji, E. (2024). Addressing class imbalance in intrusion detection: A comprehensive evaluation of machine learning approaches. Electronics, 14(1): 69. https://doi.org/10.3390/electronics14010069
[29] de Arruda Botelho, M., Ata Baykara, C., Burak Ünal, A., Pfeifer, N., Akgün, M. (2025). Privacy-preserving AUC computation in distributed machine learning with PHT-meDIC. Plos Digital Health, 4(11): e0000753. https://doi.org/10.1371/journal.pdig.0000753
[30] Mallah, S., Delsouz Khaki, B., Davatgar, N., Scholten, T., et al. (2022). Predicting soil textural classes using random forest models: Learning from imbalanced dataset. Agronomy, 12(11): 2613. https://doi.org/10.3390/agronomy12112613
[31] Zheng, L., Han, Q., Junhu, Z. (2022). A combination method of resampling and random forest for imbalanced data classification. In 2022 4th International Conference on Advances in Computer Technology, Information Science and Communications (CTISC), Suzhou, China, pp. 1-5. https://doi.org/10.1109/CTISC54888.2022.9849803
[32] Boyd, K., Eng, K.H., Page, C.D. (2013). Area under the precision-recall curve: Point estimates and confidence intervals. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2013), Prague, Czech Republic, pp. 451-466. https://doi.org/10.1007/978-3-642-40994-3_29
[33] Pivar, J., Vugec, D.S. (2025). Cyber incident landscape and profiling: Exploring patterns, motives, and impacts through EuRepoC data. In 2025 MIPRO 48th ICT and Electronics Convention, Opatija, Croatia, pp. 782-787. https://doi.org/10.1109/MIPRO65660.2025.11132012
[34] Pseftelis, T., Chondrokoukis, G. (2025). Understanding cyber incident dynamics in the European Union: A study of actor types and sector vulnerabilities. https://doi.org/10.20944/preprints202504.2169.v1
[35] Delgado, J.J., Fidalgo, E., Alegre, E., Carofilis, A., Martínez-Mendoza, A. (2024). CECILIA: Enhancing CSIRT effectiveness with transformer-based cyber incident classification. In Proceedings of the 1st International Conference on NLP & AI for Cyber Security, pp. 186-195.
[36] Bertholat, J., Merad, M., Barbier, J. (2023). Methodological insights for the prevention of cyber-attacks risks in the energy sector: An empirical study. In 33rd European Safety and Reliability Conference. Southampton, United Kingdom.