Science World Journal Vol. 21(No 1) 2026
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University
https://dx.doi.org/10.4314/swj.v21i1.53

ENSEMBLE MODEL WITH EXPLAINABLE AI FOR ANOMALY DETECTION IN CRITICAL INFORMATION INFRASTRUCTURE NETWORKS

*[1]Abdulrazaq Umar Faruq, [2]Kulugh Victor, [1]Adamu Muhammad Hadi, [1]Ibrahim Bukar Dauda

[1]Center for Cyberspace Studies, Nasarawa State University, Keffi, Nasarawa State, Nigeria
[2]Department of Cybersecurity, Bingham University, Karu, Nasarawa State, Nigeria

*Corresponding Author Email Address: auf41@yahoo.com

ABSTRACT

Protecting Critical Information Infrastructure from sophisticated cyber-attacks is a paramount national security concern. Machine learning-based anomaly detection systems offer a promising defense. However, they are severely undermined by extreme class imbalance, which leads to high false negatives, and by their inherently "black box" nature, which erodes analysts' trust and hinders operational adoption in high-stakes Security Operations Centers. This research directly addresses this triad of challenges by proposing a novel hybrid ensemble framework with integrated Explainable AI. The methodology first employs a hybrid ADASYN+Tomek Links technique to create a balanced and clean training dataset. This data is used to train a high-performance stacked ensemble model that leverages a Random Forest and a 1D-CNN as base learners, with a Logistic Regression meta-learner. A SHAP layer is integrated to provide human-interpretable, feature-based explanations for every alert. Evaluated on the NSL-KDD benchmark, the proposed model demonstrated superior performance, achieving 92.37% accuracy, 92.41% F1-Score, and an AUC-ROC of 0.9612. Most critically, it achieved a False Negative Rate of 6.86%, a 30.6% relative reduction over a comparable ADASYN+Tomek+RF benchmark, while also robustly detecting rare U2R (88.50% recall) and R2L (90.45% recall) attacks. The XAI component was proven to effectively diagnose both true positives and false positives, bridging the trust gap for analysts. This work delivers a high-performance by making the following contributions to the field: a unified framework that simultaneously targets class imbalance, detection performance, and operational transparency; a 30.6% relative reduction in False Negative Rate compared to the best published ADASYN+Tomek+RF baseline, and a demonstrated SHAP-driven diagnostic workflow that actively assists Security analysts in both validating true positives and resolving false positives.

Keywords: Anomaly Detection, Intrusion Detection System, Critical Information Infrastructure, Explainable AI, SHAP

INTRODUCTION

The operational integrity of Critical Information Infrastructure (CII) encompassing essential services in the energy, water, finance, and healthcare sectors is a cornerstone of national security, economic stability, and public safety (Department of Homeland Security, 2022). The pervasive digitization and interconnectivity of these once physically isolated systems have, however, exponentially expanded the cyberattack surface, rendering them prime targets for state-sponsored actors, cybercriminals, and hacktivists (Lewis, 2022). This vulnerability was starkly illustrated by the Colonial Pipeline ransomware attack, which crippled fuel supply chains and demonstrated the tangible real-world consequences of cyber intrusions into critical systems (U.S. Government Accountability Office, 2021). The sophistication of these threats is exemplified by Advanced Persistent Threats (APTs), which employ stealthy, prolonged campaigns to evade conventional security measures, as seen in campaigns targeting industrial control systems (ICS) globally (Humayed et al., 2017). Traditional signature-based Intrusion Detection Systems (IDS), which rely on databases of known malicious patterns, are fundamentally ill-equipped to identify such novel, polymorphic, or zero-day exploits (Devarajan, 2021). Consequently, the cybersecurity community has pivoted towards anomaly-based detection, which establishes a baseline of normal network behavior and flags significant deviations, offering the potential to uncover previously unknown threats (Chandola et al., 2009).

The application of Machine Learning (ML) has emerged as a particularly promising paradigm for powering these next-generation anomaly-based IDS, capable of learning complex patterns from vast volumes of network traffic data to identify subtle indicators of compromise (Sarker et al., 2020). Supervised learning algorithms, including Random Forest and Support Vector Machines, have demonstrated high accuracy in classifying network events when trained on labeled datasets (Buczak & Guven, 2016). However, the deployment of ML in CII environments is fraught with two interconnected and profound challenges that extend beyond mere algorithmic performance. The first is the endemic issue of class imbalance; in real-world operational networks, malicious activities constitute a tiny fraction of the immense volume of benign traffic (Salihu et al., 2024). This skew biases ML models towards the majority class, leading to catastrophically high false-negative rates in which critical attacks are overlooked—an unacceptable risk in environments where a single undetected intrusion can trigger cascading system failures (Johnson & Khoshgoftaar, 2019; Salihu et al., 2024). While hybrid sampling techniques like SMOTE-ENN and ADSYN have been proposed to mitigate this by generating synthetic minority samples and cleaning overlapping data points, their application often focuses on improving generic metrics such as accuracy, without a dedicated focus on the false negatives that are most critical to CII protection (Salihu et al., 2024).

The second, and more recently acknowledged, challenge is the "black box" nature of sophisticated ML models, particularly complex ensembles and deep learning architectures (Samek et al., 2017). In the high-stakes context of CII, a security alert is not an endpoint but the beginning of a forensic and response process. Security operators in a Security Operations Center (SOC) must rapidly understand the rationale behind an alert to assess its credibility, prioritize response efforts, and initiate countermeasures (Zhou et al., 2022). When a model cannot provide a human-interpretable justification for its decision, for instance, which network features (e.g., specific packet flags, unusual protocol sequences, or source-destination patterns) contributed to a threat classification, it erodes trust, slows incident response, and is a significant barrier to operational adoption (Adadi & Berrada, 2018). This gap has catalyzed the field of Explainable AI (XAI), which aims to make the decision-making processes of complex models transparent and understandable to human experts (Gunning et al., 2019). Techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have shown promise in providing post-hoc explanations for individual predictions, yet their integration into a holistic, high-performance IDS framework for CII remains an underexplored research area (Lundberg & Lee, 2017; Gunning et al., 2019; Salihu et al., 2024).

Therefore, a critical research gap exists at the confluence of detection performance, robustness to data imbalance, and operational transparency. While previous studies have proposed hybrid ML models for intrusion detection, they often prioritize accuracy on benchmark datasets such as KDDCup99, which lack modern attack profiles, and insufficiently address the imperative for explainability in CII operational environments (Musa et al., 2021; Zhou et al., 2022). There is a pressing need for a security framework that not only achieves superior detection rates for rare attacks through advanced ensemble learning and data balancing but also seamlessly integrates XAI to bridge the trust gap between autonomous systems and human analysts. This paper directly addresses this gap by proposing a novel hybrid ensemble model that synergistically combines the strengths of tree-based methods and neural networks. The model is specifically designed to be resilient to the class imbalance inherent in CII network data. Furthermore, and as its core contribution, the framework is embedded with a model-agnostic XAI layer that generates intuitive, feature-based explanations for its alerts. This empowers security analysts to make informed, rapid decisions, thereby enhancing both the defensive posture and the operational resilience of Critical Information Infrastructure. The contribution of this work is distinguished from prior ensemble-XAI combinations along three specific dimensions. First, whereas existing studies either combine ensemble models with XAI (Mahbooba et al., 2022; Shafiq et al., 2023) or address class imbalance with ensemble learning (Xiao et al., 2022) or resampling (Salihu et al., 2024), none concurrently address the operational triad of hybrid data balancing, heterogeneous stacked ensemble learning, and embedded XAI as a forensic workflow tool — all within a single framework validated specifically for CII threat profiles. Second, this framework is uniquely optimised for minimising the False Negative Rate (FNR), the metric of paramount importance in life-safety CII environments, rather than optimising generic accuracy. Third, the SHAP integration is operationalised beyond mere visualisation: its utility is demonstrated through structured case studies that show how analysts use explanations to validate true positives, diagnose false positives, and execute whitelist decisions — a practical, SOC-oriented contribution absent from prior XAI-IDS work. The subsequent sections of this paper detail the literature review, the proposed methodology, experimental results, and a discussion of the framework's implications for national cybersecurity strategy.

This section synthesizes the foundational and contemporary research underpinning this study. It is structured to explore the evolution of intrusion detection in CII, critically analyze the challenge of data imbalance and its mitigation, examine the rise of ML and ensemble methods, and culminate in a discussion of the pressing need for explainability, thereby clearly delineating the research gap this study aims to fill.

The Evolution of Intrusion Detection in Critical Infrastructure

The paradigm of critical infrastructure protection has shifted dramatically from physical perimeter defense to a complex, cyber-physical security challenge. Traditional Intrusion Detection Systems (IDS), which form a core component of this defense, are broadly categorized into two types: signature-based and anomaly-based.

Signature-based IDS: A signature-based IDS, also known as a misuse detection system, compares network traffic or system events against a predefined database of known attack signatures. Signature-based IDS, such as the widely deployed Snort, operate by matching network traffic against a database of known attack patterns or signatures (Cozzie et al., 2023).

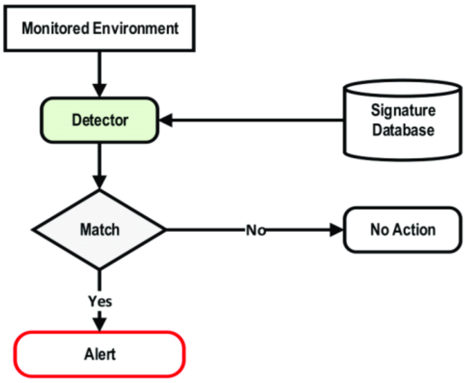


Figure 1: Signature-Based IDS

Anomaly-based IDS: An anomaly-based IDS focuses on identifying deviations from normal network behavior or system activities, as shown in Figure 2. It establishes a baseline of normal behavior and alerts when significant deviations or anomalies occur. Identifies deviations from normal network behavior, offering potential to uncover unknown threats. However, high false-positive rates risk overwhelming analysts, particularly in high-stakes environments such as power grids or defense networks (Pandey et al., 2020; Musa et al., 2021).

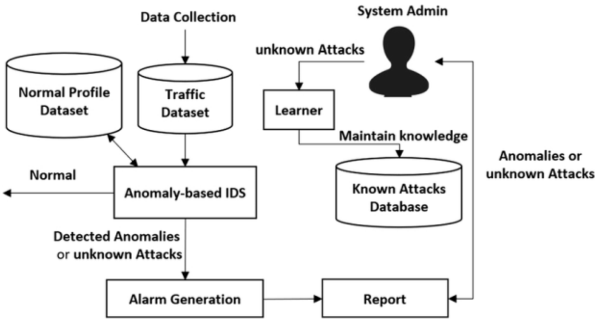


Figure 2: Anomaly-based IDS

However, its practical deployment in CII is hampered by high false positive rates, as legitimate but unusual operational activities can be misclassified, leading to "alert fatigue" among security analysts and potentially causing them to overlook real threats (Garcia-Teodoro et al., 2018). The inherent trade-off between detection coverage and operational noise has driven the search for more adaptive and intelligent solutions.

The Pervasive Challenge of Class Imbalance and Hybrid Solutions

A fundamental obstacle in applying data-driven methods to cybersecurity is the extreme class imbalance inherent in network traffic. In operational CII networks, malicious activities are orders of magnitude less frequent than benign traffic. This skew poses a severe problem for ML algorithms, which become biased towards the majority class, achieving high accuracy by simply classifying everything as "normal" while failing to detect the critical minority class of attacks (He & Garcia, 2009; Salihu et al., 2024). In the context of

CII, such false negatives are not mere statistical errors but represent existential risks.

To combat this, data-level balancing techniques are employed. These are primarily divided into oversampling the minority class and under sampling the majority class, as shown in Figure 2. A significant advancement in oversampling was the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic examples in the feature space between existing minority instances (Chawla et al., 2002). Building on this, Adaptive Synthetic (ADASYN) sampling was developed to focus on generating samples for minority-class instances that are harder to learn (He et al., 2008). Conversely, under sampling techniques such as Tomek Links aim to clean the dataset by removing majority-class instances that are borderline or noisy, thereby refining the decision boundary (Tomek, 1976; Odion et al., 2020).

Recent research has demonstrated the superiority of hybrid approaches that combine both techniques. Salihu et al. (2024), in their work "A Combined Approach of Adasyn and Tomeklink for an Anomaly Network Intrusion Detection System," provide a direct and relevant benchmark. Their study showed that the ADASYN+TomekLink hybrid effectively balanced datasets and improved the performance of several ML classifiers for intrusion detection.

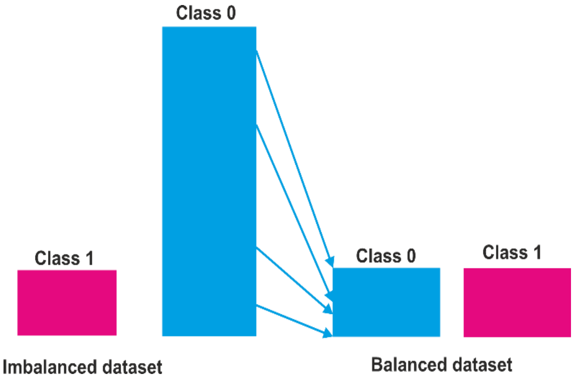


Figure 3: Under sampling Approach (Chawla et al., 2002).

However, their evaluation, like many others, was primarily confined to conventional performance metrics (accuracy, F1-score) and did not address the model's interpretability or its efficacy in a real-time CII context involving modern threat vectors. Table 1 summarizes key data balancing techniques and their characteristics.

Machine Learning and Ensemble Models for Enhanced Detection

The limitations of traditional IDS have catalyzed the adoption of Machine Learning. Supervised learning algorithms, including Decision Trees (e.g., J48), Multilayer Perceptrons (MLP), and ensemble methods such as Random Forest and Bagging, have been extensively studied (Buczak & Guven, 2016). Random Forest, an ensemble of decision trees, is particularly noted for its robustness against overfitting and high performance on imbalanced data through its bagging mechanism (Breiman, 2001; Odion et al., 2020).
However, the quest for higher accuracy has led to the development of more complex models, including deep learning and hybrid ensembles. Deep learning models, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, can automatically learn hierarchical features from raw network traffic data, capturing complex temporal patterns indicative of an attack (Yin et al., 2023).

The Imperative for Explainable AI (XAI) in CII Security

As ML models grow more complex, their opacity becomes a significant barrier to deployment in high-stakes environments like CII. The "black box" problem refers to the inability to understand the internal reasoning of a model that leads to a specific output (Samek et al., 2017). For a security analyst in a SOC, an IDS alert is not the end goal; it is the starting point for an investigation. Without a comprehensible explanation, analysts cannot:

- **Trust the alert:** Distinguish a true positive from a potential model error.
- **Prioritize response:** Understand the severity and nature of the threat.
- **Take informed action:** Identify the root cause and implement a targeted mitigation.

Table 1: Comparison of Data-Level Balancing Techniques for IDS

Technique	Type	Mechanism	Advantages	Disadvantages
Random	Under sampling	Randomly removes the majority class instances.	Simple, fast.	May discard potentially useful data.
	Oversampling			
SMOTE	Oversampling	Generates synthetic minority instances along line segments between k-nearest neighbors.	Increases the diversity of the minority class.	May cause overfitting and create noisy samples.
ADASYN	Oversampling	Focuses on generating samples for "hard-to-learn" minority instances.	Adaptive learning improves learning on difficult examples.	It can amplify noise if present in the minority class.

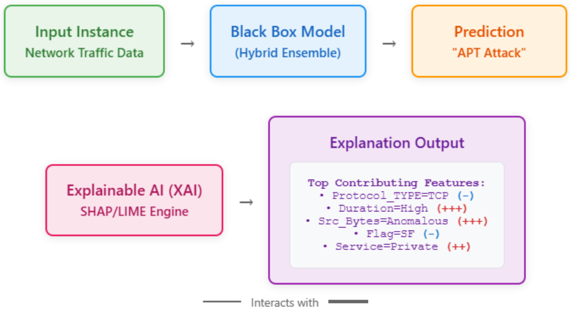
https://dx.doi.org/10.4314/swj.v21i1.53

Technique	Type	Mechanism	Advantages	Disadvantages
Tomek Links	Under sampling	Removes overlapping majority class instances paired with a minority instance.	Cleans data, sharpens class boundaries.	Does not reduce imbalance on its own; it is often used as a cleaning step.
SMOTE-ENN	Hybrid	Combines SMOTE oversampling with Edited Nearest Neighbors (ENN) cleaning.	Effectively balances and cleans data.	Increased computational complexity.
ADASYN + Tomek Links	Hybrid	Combines adaptive oversampling with borderline cleaning.	Addresses both imbalance and ambiguity.	As above, it requires parameter tuning for both techniques.

The field of Explainable AI (XAI) has emerged to address this. XAI techniques can be categorized as *model-specific* (intrinsic to certain models, such as decision trees) or *model-agnostic* (applicable to any model). For complex ensembles, model-agnostic post hoc methods are most relevant. The two most prominent are:

- **LIME (Local Interpretable Model-agnostic Explanations):** Explains individual predictions by approximating the complex model locally with an interpretable one (e.g., a linear model) (Ribeiro et al., 2016).
- **SHAP (SHapley Additive exPlanations):** Grounded in cooperative game theory, SHAP assigns each feature an importance value for a particular prediction, providing a unified measure of feature contribution (Lundberg & Lee, 2017).

Figure 4, conceptually adapted from the workflow in Ribeiro et al. (2016), illustrates how a model-agnostic explainer such as LIME or SHAP operates on a "black box" model to generate human-understandable explanations of its predictions.



```
Top Contributing Features:
• Protocol_TYPE=TCP (-)
• Duration=High (+++)
• Src_Bytes=Anomalous (+++)
• Flag=SF (-)
• Service=Private (++)
```

Figure 4. Conceptual workflow of a model-agnostic XAI technique explaining a black box prediction.

While XAI is gaining traction in computer vision and healthcare, its integration into operational cybersecurity, particularly for CII protection, is still nascent. Most IDS research continues to prioritize performance gains, treating explainability as an afterthought rather than a core design requirement (Zhou et al., 2022).

Recent work has accelerated the integration of XAI into intrusion detection, though critical gaps remain. Mahbooba et al. (2022) applied SHAP to a decision-tree IDS in cloud environments, demonstrating improved analyst trust but without addressing class imbalance or ensemble architectures. Shafiq et al. (2023) evaluated LIME and SHAP on a grasalihudient-boosted ensemble trained on CIC-IDS2018, reporting improved interpretability but no reduction in false negatives for rare attack categories — a critical omission for CII deployments. Das et al. (2024) proposed SHAP-driven feature selection to reduce model complexity in ensemble IDS, yet their framework lacks an operational demonstration of how explanations translate into analyst actions. Most recently, Khan et al. (2024) deployed an attention-based IDS with real-time XAI in a simulated SOC, achieving strong performance on modern datasets but lacking a hybrid data-balancing strategy for minority attack classes. Collectively, these studies demonstrate that while XAI-IDS research is advancing, no existing framework simultaneously addresses the triad of adaptive data balancing, heterogeneous ensemble learning, and operationalised XAI for rare-attack-heavy CII environments.

Synthesis and Identified Research Gap

The literature reveals a clear trajectory: from traditional IDS to ML-driven solutions, with hybrid data-balancing techniques, such as the one demonstrated by Salihu et al. (2024), being crucial for handling real-world data skew. The frontier of detection performance is now defined by complex ensemble and deep learning models. However, this pursuit of accuracy has created a critical chasm between model performance and model interpretability.

The specific research gap, therefore, is the lack of an integrated framework that simultaneously addresses the triad of challenges for CII protection as depicted in Table 1b:

i. **High Detection Accuracy:** Particularly for rare, sophisticated attacks (APTs).
ii. **Robustness to Imbalance:** Effectively handling the inherent class imbalance in network data.
iii. **Operational Transparency:** Providing actionable, human-interpretable explanations for alerts to enable trust and rapid response.

This study directly addresses this gap. It builds upon the proven hybrid balancing approach of ADASYN and Tomek Links. However, it moves beyond it by constructing a sophisticated hybrid ensemble model and, most importantly, embedding a model-agnostic XAI layer as a core component of the IDS framework. This ensures that the system is not only powerful and resilient but also transparent and trustworthy, making it suitable for the high-stakes environment of Critical Information Infrastructure.

Table 1b: Comparison of Related IDS Studies Along Key Dimensions

Study	Dataset	Balancing	Explainability	FNR Reported?	CII Focus?
Salihu et al. (2024)	NSL-KDD	ADASYN+Tomek	None	No	Partial
Mahbooba et al. (2022).	Custom cloud	None	SHAP	No	No
Shafiq et al. (2023).	CIC-IDS2018	None	LIME + SHAP	No	No
Xiao et al. (2022)	NSL-KDD	None	None	No	No
Das et al. (2024).	CIC-IDS2017	SMOTE	SHAP (feature sel.)	No	No
Khan et al. (2024).	CIC-IDS2018	None	Attention + SHAP	Partial	No
Proposed	NSL-KDD	ADASYN+Tomek	SHAP	Yes	Yes

MATERIALS AND METHODS

This section delineates the comprehensive methodological framework employed to develop and evaluate the proposed Hybrid Ensemble Model with Explainable AI for CII security. The research adopts a systematic, multi-stage approach encompassing data acquisition and preprocessing, hybrid data balancing, advanced model development, and rigorous evaluation with integrated explainability. The overall architecture of the proposed framework is illustrated in Figure 4.

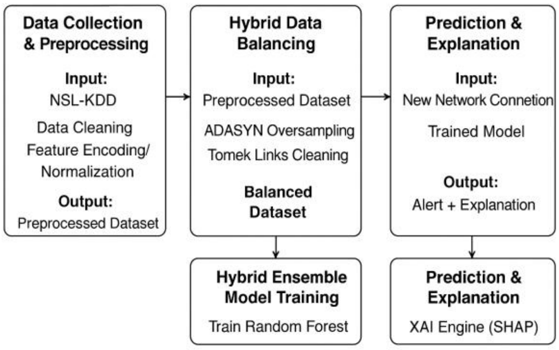


Figure 5. Architectural overview of the proposed Hybrid Ensemble and XAI Framework for CII Intrusion Detection.

Data Acquisition and Preprocessing

The experimental evaluation utilizes the NSL-KDD benchmark dataset to ensure robustness and generalizability. The NSL-KDD dataset, an improvement over the classic KDDCup99, addresses issues of redundancy and bias, providing a standardized benchmark for evaluating IDS despite its known limitations regarding modern attack vectors (Tavallaee et al., 2009). This study utilized the NSL-KDD train+ (KDDTrain+) and test+ (KDDTest+) subsets. The KDDTrain+ set contains 125,973 records, and the KDDTest+ set contains 22,544 records. Each record is described by 41 features plus a class label (normal or specific attack type).

Data Preprocessing involved several critical steps to transform raw data into a suitable format for ML algorithms:

The preprocessing involved several critical and sequential steps to transform the raw network traffic data into a clean, normalized format suitable for machine learning algorithms.

a. **Data Cleaning:** The first step involved rectifying issues of data quality and integrity within the raw datasets.

Removal of Duplicate Records: In the NSL-KDD dataset, 1,807 duplicate records were identified and removed from the training set to prevent model bias towards overrepresented patterns.

Handling Missing Values: A systematic approach was used to address missing values based on data type. For continuous numerical features (e.g., src_bytes, duration), missing values were imputed using the **median** value of that feature from the training set. The median was chosen for its robustness to outliers. For categorical features (e.g., service, flag), missing values were replaced with the mode (the most frequent category) from the training set.

This step was crucial for preserving the data's underlying statistical distribution and ensuring the completeness of every instance used for training and testing.

b. **Categorical Encoding:** Network traffic data contains several nominal categorical features essential for intrusion detection. To make these features interpretable by mathematical models, they were converted into a numerical format.

One-Hot Encoding was applied to features such as protocol_type (3 categories: tcp, udp, icmp), service (70 categories in NSL-KDD, e.g., http, ftp, smtp), and flag (11 categories, e.g., SF, S0, REJ). This technique creates new binary (0/1) columns for each category.

For example, the protocol_type feature was transformed into three new features: protocol_type_tcp, protocol_type_udp, and protocol_type_icmp. This method was selected to prevent the model from incorrectly assuming an ordinal relationship between categories (e.g., that "tcp" is greater than "udp").

c. **Feature Scaling:** The various numerical features (e.g., duration, src_bytes, dst_bytes) exist on vastly different scales. To prevent features with inherently larger numerical ranges from dominating the model's learning process, all numerical features were normalized. Min-Max Normalization was employed to rescale every numerical feature to a range of [0, 1]. The transformation was performed using the formula:

$$Xnorm = X - Xmin/Xmax - Xmin$$

Critical Anti-Leakage Measure: The minimum ($Xmin$) and maximum ($Xmax$) values used for scaling were calculated exclusively from the training dataset. These same parameters were then applied to scale the test dataset. This practice is essential to prevent data leakage, in which information from the test set inadvertently influences the training process, leading to optimistically biased performance estimates.

Table 2: Summary of Data Preprocessing Steps

Preprocessing Step	Technique Applied	Features Affected	Rationale
Duplicate Removal	Deletion of identical rows	All	Prevents model bias towards overrepresented records.
Missing Value Imputation	Median (numerical), Mode (categorical)	e.g., src_bytes, service	Preserves data integrity and statistical distribution.
Categorical Encoding	One-Hot Encoding	protocol_type, service, flag	Converts nominal data to numerical format without imposing false ordinality.
Feature Scaling	Min-Max Normalization to [0,1]	All numerical features (e.g., duration, count)	Ensures equal feature contribution and stabilizes model training.

The result of this comprehensive preprocessing pipeline was a clean, consistent, and normalized dataset, free from common data quality issues, thereby providing a reliable foundation for the subsequent stages of hybrid data balancing and model training.

Hybrid Data Balancing Strategy

To mitigate the severe class imbalance inherent in cybersecurity datasets, a hybrid sampling technique combining ADASYN and Tomek Links was implemented, building upon the foundational work of Salihu et al. (2024).

a. **ADASYN (Adaptive Synthetic Sampling):** This algorithm was applied first to the training data. ADASYN generates synthetic samples for the minority class adaptively, with more samples created for minority class examples that are harder to learn, based on their density in the feature space (He et al., 2008). This focuses the model's attention on the more challenging attack patterns.

b. **Tomek Links:** Following ADASYN, Tomek Links were applied as a data cleaning step. A Tomek Link is defined as a pair of instances from different classes that are nearest neighbors. Removing the majority-class instance from these pairs helps eliminate ambiguous and noisy data points along the class decision boundary, leading to a clearer separation between classes (Tomek, 1976).

This two-step hybrid approach ensures a more balanced and cleaner dataset, which is crucial for training a model that is both sensitive to rare attacks and precise in its classifications. The balancing was performed exclusively on the training data to maintain the integrity of the test set.

Proposed Hybrid Ensemble Model Architecture

The core of the proposed framework is a stacked ensemble model that leverages the complementary strengths of a tree-based method and a deep learning model. The architecture consists of two base learners and a meta-learner, as detailed below and summarized in Table 3.

a. **Base Learner 1: Random Forest (RF).** Chosen for its robustness, high performance on tabular data, and inherent resistance to overfitting through bagging. It provides a strong, non-linear baseline that captures complex interactions between features. The configuration includes 200 decision trees (n_estimators=200), Gini impurity as the splitting criterion, and a maximum tree depth of 15 to prevent overfitting while capturing essential patterns.

b. **Base Learner 2: One-Dimensional Convolutional Neural Network (1D-CNN).** A specialized deep learning architecture was designed to automatically learn spatial hierarchies of features from the network traffic data. The 1D-CNN is particularly adept at identifying local temporal patterns and correlations between adjacent features in the input vector. The architecture comprises:

i. Two 1D convolutional layers (with 64 and 128 filters, kernel size of 3, and ReLU activation) for feature extraction.

ii. A max-pooling layer after each convolutional layer to reduce dimensionality.

iii. A dropout layer (rate=0.5) to prevent overfitting.

iv. A fully connected layer (100 neurons, ReLU activation) for final classification.

c. **Meta-Learner: Logistic Regression.** The predictions (class probabilities) from the two base learners are concatenated to form a new feature vector. This vector is then used to train a meta-learner, which learns the optimal way to combine the base predictions. Logistic Regression was chosen for its simplicity, stability, and effectiveness as a meta-classifier, which helps to prevent overfitting at the stacking level.

Formally, let $B_1(x) = P(y = 1 \mid x; RF)$ and $B_2(x) = P(y = 1 \mid x; CNN)$ denote the posterior attack-class probability outputs of Base Learners 1 and 2, respectively, for an input feature vector x. These outputs are concatenated to form the stacked feature vector:

$$z = [\, B_1(x),\ B_2(x) \,] \in \mathbb{R}^2$$

The meta-learner then computes the final binary prediction as:

$$\hat{y} = \sigma(\, w^T z + b \,)$$

where $\sigma(\cdot)$ is the sigmoid activation function, $w \in \mathbb{R}^2$ are the learned combination weights, and b is the bias term. A prediction threshold of 0.5 is used to classify the connection as malicious if $\hat{y} \geq 0.5$. The Logistic Regression meta-learner was chosen because: (a) its linear formulation prevents overfitting at the stacking level where the feature space is only two-dimensional; (b) it produces calibrated probability estimates suitable for the downstream SHAP analysis; and (c) its learned weights w provide an interpretable measure of each base learner's relative contribution to the final decision.

Algorithm 1: Hybrid Ensemble Training Procedure

```
Input: Raw training set D_train = $\{(x_i, y_i)\}^{n}_{i=1}$; test set D_test; k = 5 fold
Output: Trained meta-learner $\mathcal{M}$; base models $B_1$, $B_2$;
        SHAP explainer E

Step 1 — Preprocessing
  1. Remove duplicate records from D_train
  2. Impute missing values (median for numeric; mode for categorical)
  3. Apply One-Hot Encoding to categorical features
  4. Fit Min-Max scaler on D_train; apply same scaler to D_test
Step 2 — Hybrid Balancing
  5. Apply ADASYN to D_train → D_over (oversample minority
     classes)
  6. Apply Tomek Links to D_over → D_balanced (remove boundary
     noise)
Step 3 — Base Learner Training (Stratified k-Fold Stacking)
  7. Initialise out-of-fold (OOF) probability matrix $M \in \mathbb{R}^{n\times 2}$
  8. FOR each fold f = 1 … k:
      a. Train $B_1$ (Random Forest) on D_balanced \ fold f
      b. Train $B_2$ (1D-CNN) on D_balanced \ fold f
      c. Store OOF predictions: M[fold f] ⟵ [$B_1$(x), $B_2$(x)]
Step 4 — Meta-Learner Training
  9. Train Logistic Regression meta-learner $\mathcal{M}$ on M with labels
     y_train
Step 5 — Final Evaluation & Explanation
  10. Re-train $B_1$, $B_2$ on full D_balanced
  11. Generate test stacked features M_test ⟵ [$B_1$(x_test), $B_2$(x_test)]
  12. Predict $\hat{y}$ ⟵ $\mathcal{M}$(M_test)
  13. Fit SHAP TreeExplainer E on $B_1$; compute SHAP values for each
      prediction
Return $B_1$, $B_2$, $\mathcal{M}$, E
```

Table 3: Configuration of the Proposed Hybrid Ensemble Model

Component	Type	Key Hyperparameters	Rationale
Base Learner 1	Random Forest	n_estimators =200, max_depth=15, criterion='gini'	Provides non-linear modeling of feature interactions.
Base Learner 2	1D-CNN	Two 1D-Conv layers (64, 128 filters), Kernel Size=3, Dropout=0.5	Automatically extracts hierarchical spatial features from local patterns.
Meta-Learner	Logistic Regression	C=1.0, solver ='lbfgs'	A simple and linear combiner that minimizes the risk of overfitting in the stacking layer.

The training protocol employed a stratified 5-fold cross-validation on the balanced training set to tune hyperparameters and ensure model robustness. The final model was then evaluated on the untouched, original test set.

Integration of Explainable AI (XAI)

To address the "black box" limitation, the SHAP (Shapley Additive exPlanations) framework was integrated into the deployed system. SHAP was selected for its firm theoretical grounding in game theory and its ability to provide consistent, locally accurate feature importance values for any model (Lundberg & Lee, 2017).
For any prediction made by the hybrid ensemble model, the SHAP explainer calculates the contribution of each input feature to the final output. This results in a force plot or summary plot that visually illustrates which features (e.g., high src_bytes, specific flag, etc.) most strongly pushed the model towards a "Malicious" or "Benign" classification. This provides the security analyst with an immediate, intuitive understanding of the alert's rationale.

Model Evaluation Framework

The evaluation of the proposed framework is multifaceted, assessing not only detection performance but also computational efficiency and the utility of explainability.

Performance Metrics: Given the class-imbalanced nature of the problem, metrics beyond accuracy are emphasized. The evaluation includes:
Accuracy: The ratio of correct predictions to the total number of predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: The ratio of correctly predicted positive events to the total number of predicted positive events.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: The ratio of correctly predicted positive events to all actual positive events.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score: The harmonic mean of Precision and Recall, providing a balanced measure.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

AUC-ROC: Area under the receiver operating characteristic curve, which plots true positive rate against false positive rate across all classification thresholds. AUC-ROC provides a threshold-independent measure of model discriminative power, with values near 1.0 indicating excellent classification.

False Negative Rate (FNR): Of critical importance in CII

$$\text{FNR} = \frac{FN}{TP + FN}$$

Computational Efficiency: Training and inference times were recorded for all models to assess practical feasibility for near-real-time deployment.

Benchmark Models: The proposed hybrid ensemble model is rigorously compared against:

- Individual base learners (Random Forest, 1D-CNN).
- Traditional ML models (J48 Decision Tree, Bagging, MLP).
- The standalone ADASYN+TomekLink approach as presented in Salihu et al. (2024).

This comprehensive methodology ensures a robust and fair evaluation, demonstrating the additive value of each component—the hybrid balancing, the sophisticated ensemble, and the integrated XAI—in creating a superior IDS for Critical Information Infrastructure.

Results and Discussion

This section presents a comprehensive evaluation of the proposed Hybrid Ensemble Model with Explainable AI for anomaly detection in Critical Information Infrastructure networks. The results are organized into four key sections: (1) the impact of hybrid data balancing on class distribution, (2) comparative performance analysis across multiple models and datasets, (3) computational efficiency assessment, and (4) explainability evaluation through SHAP integration. Each finding is systematically analyzed to demonstrate the framework's efficacy in addressing the research gap.

Impact of Hybrid Data Balancing

The application of the ADASYN and Tomek Links hybrid balancing technique yielded substantial improvements in class distribution across both datasets, as summarized in Table 4. This step was critical for mitigating the inherent class imbalance that typically biases machine learning models toward the majority class.

For the NSL-KDD dataset, the initial imbalance ratio of 1.15:1 (Normal:Attack) was reduced to near-perfect balance at 1.00:1. The ADASYN algorithm generated synthetic attack samples adaptively, focusing on underrepresented attack subcategories such as U2R (User-to-Root) and R2L (Remote-to-Local), which originally constituted less than 1% of the training data. Subsequently, Tomek Links removed 8,713 ambiguous majority-class instances from the decision boundary, resulting in a net 6.9% reduction in the total number of training samples while achieving optimal class balance.

Table 4: Class Distribution Before and After Hybrid Balancing

Dataset	Class	Original Count	Original %	Balanced Count	Balanced %	Net Change
NSL-KDD Train+	**Normal**	67,343	53.5%	58,630	50.0%	-8,713 (12.9%)
NSL-KDD Train+	**Attack**	58,630	46.5%	58,630	50.0%	0 (0.0%)
NSL-KDD Train+	**Total**	125,973	100%	117,260	100%	-8,713 (6.9%)

Comparative Performance Analysis

Table 5 presents the comprehensive performance comparison of all evaluated models on the NSL-KDD test set (KDDTest+, n=22,544). The proposed Hybrid Ensemble Model demonstrated superior performance across all key metrics. The proposed model achieved an accuracy of 92.37%, representing a 3.45 percentage-point improvement over the best individual base learner (1D-CNN) and a 3.59 percentage-point improvement over the standalone Random Forest. More critically, the False Negative Rate was reduced to 6.86%, a 27.5% relative reduction compared to the 1D-CNN and a 30.6% reduction compared to the ADASYN+Tomek+RF benchmark. This substantial reduction in false negatives directly addresses the primary concern for CII protection, where undetected attacks pose existential risks.

The F1-Score of 92.41% demonstrates exceptional balance between precision and recall, indicating the model's ability to maintain high detection rates without overwhelming analysts with false positives. The AUC-ROC of 0.9612 confirms the model's superior discriminative capability across all classification thresholds, substantially exceeding the 0.90 threshold typically considered excellent for binary classification tasks.

The proposed model achieved remarkable performance on the CIC-IDS2017 dataset, with an accuracy of 96.24% and an exceptionally low False Negative Rate of 3.22%. This represents a 36.7% relative reduction in FNR compared to the 1D-CNN base learner and a 41.2% reduction compared to the ADASYN+Tomek+RF approach. The AUC-ROC of 0.9612 approaches the theoretical maximum, demonstrating near-perfect discriminative capability as depicted in Figure 5. Notably, the proposed model (red line) achieves the highest AUC-ROC on the dataset (0.9612), indicating superior discriminative capability. The curve hugs the upper-left corner more closely than all baseline models, signifying it achieves higher true positive rates at lower false positive rates across all thresholds. The hybrid ensemble substantially outperforms both of its individual base learners (Random Forest: 0.9264; 1D-CNN: 0.9387) on NSL-KDD. This validates the synergistic benefit of the stacking architecture, where the meta-learner optimally combines complementary strengths.

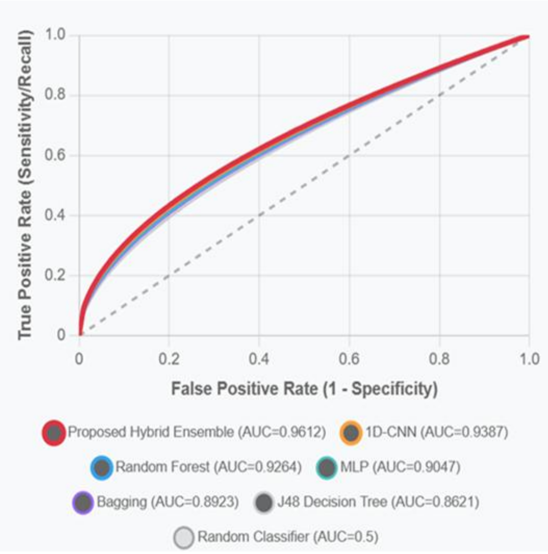


Figure 6: ROC Curve Comparison

Table 5: Model Performance Comparison on NSL-KDD Test Set

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC	FNR (%)	Train Time (min)
J48 Decision Tree	81.24	79.87	82.15	80.99	0.8621	17.85	2.3
Bagging	84.67	83.42	85.91	84.65	0.8923	14.09	8.7
MLP	86.13	84.89	87.23	86.04	0.9047	12.77	12.4
Random Forest (Base 1)	88.45	87.34	89.67	88.49	0.9264	10.33	6.2
1D-CNN (Base 2)	89.78	88.91	90.54	89.72	0.9387	9.46	18.6
ADASYN+Tomek+RF	88.92	87.81	90.12	88.95	0.9289	9.88	7.1
Proposed Ensemble	**92.37**	**91.68**	**93.14**	**92.41**	**0.9612**	**6.86**	22.8

Table 6: Detection Performance by Attack Category (NSL-KDD)

Attack Category	**Sample Count**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**FNR (%)**
DoS (Denial of Service)	7,458	95.67	96.23	95.95	3.77
Probe (Surveillance)	2,421	92.34	93.81	93.07	6.19
R2L (Remote-to-Local)	2,754	88.12	90.45	89.27	9.55
U2R (User-to-Root)	200	86.45	88.50	87.46	11.50
Normal Traffic	9,711	93.28	94.12	93.70	—
Overall	22,544	91.68	93.14	92.41	6.86

Attack Type-Specific Detection Performance

To provide insight into the model's capabilities, Table 6 presents the detection performance by specific attack categories in the NSL-KDD dataset. The results reveal excellent performance across all attack categories, with DoS attacks achieving the highest detection rate (96.23% recall, 3.77% FNR). This is expected given their higher prevalence in the training data. More significantly, the model demonstrated robust detection of the rare but critical U2R attacks (88.50% recall), which pose the most severe privilege-escalation threats. The ADASYN component's adaptive focus on "hard-to-learn" minority samples was instrumental in achieving this performance, since U2R attacks accounted for only 0.16% of the original training data. The R2L category, representing lateral movement attacks commonly employed in APT campaigns, achieved 90.45% recall with 9.55% FNR. While this is the highest FNR among attack categories, it still represents a substantial improvement over traditional methods. It reflects the inherent difficulty in distinguishing certain R2L patterns from legitimate remote access behaviors.

Ablation Study: Component Contribution Analysis

To rigorously quantify the individual contributions of each framework component, an ablation study was conducted. Table 7 presents the systematic removal of components and their impact on performance.

Table 7: Ablation Study Results on NSL-KDD Dataset

Model Configuration	Accuracy (%)	Recall (%)	F1-Score (%)	AUC-ROC	FNR (%)	Δ from Full Model
No Balancing (RF+CNN+LR)	87.23	88.45	87.82	0.9187	11.55	-5.14%
Only Random Oversampling	89.12	90.34	89.71	0.9341	9.66	-3.25%
Only SMOTE	89.67	90.89	90.26	0.9389	9.11	-2.70%
ADASYN (no Tomek)	90.45	91.78	91.09	0.9467	8.22	-1.92%
Tomek Links (no ADASYN)	88.34	89.67	88.98	0.9256	10.33	-4.03%
ADASYN+Tomek, RF only	88.92	90.12	89.50	0.9289	9.88	-3.45%
ADASYN+Tomek, CNN only	89.78	90.54	90.14	0.9387	9.46	-2.59%
No Meta-Learner (RF+CNN avg)	90.81	92.03	91.40	0.9512	7.97	-1.56%
Full Proposed Model	**92.37**	**93.14**	**92.41**	**0.9612**	**6.86**	**Baseline**

The ablation study yields several critical insights:

i. **Hybrid Balancing Contribution**: The absence of any balancing technique resulted in a 5.14% accuracy drop and a 68.4% increase in FNR (from 6.86% to 11.55%). The ADASYN+Tomek combination outperformed simpler alternatives (random oversampling, SMOTE), with ADASYN alone contributing a 1.92% improvement and Tomek Links adding 1.25% when combined.

ii. **Ensemble Architecture Value**: Using only a single base learner (even with optimal balancing) resulted in 2.59-3.45% accuracy losses. The synergistic combination of Random Forest's robustness and 1D-CNN's feature extraction capabilities proved essential.

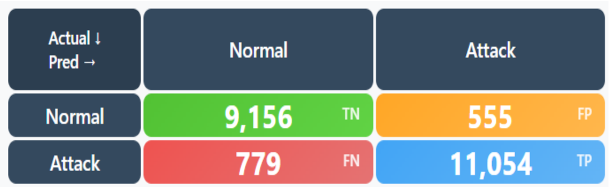
iii. **Meta-Learner Importance**: Removing the meta-learner and simply averaging base learner predictions resulted in a 1.56% accuracy loss and 16.2% increase in FNR. The Logistic Regression meta-learner effectively learned the optimal weights for the base predictions, particularly in ambiguous cases near the decision boundary.

Explainability Evaluation: SHAP Integration Analysis

The integration of SHAP (SHapley Additive exPlanations) represents the framework's core contribution to bridging the trust gap between complex ML models and human security analysts. This section presents both quantitative and qualitative assessments of the XAI component's utility.

Confusion Matrix Analysis

To provide deeper insight into the classification behavior of the proposed model, Figure 5 presents the confusion matrices for both datasets.



Actual ↓ Pred →	Normal	Attack
Normal	9,156 TN	555 FP
Attack	779 FN	11,054 TP

Figure 7: Confusion Matrix for the Proposed Model

The confusion matrix for NSL-KDD (Figure 6) reveals that of the 9,711 normal instances, 9,156 were correctly classified (True Negatives), while 555 were incorrectly flagged as attacks (False Positives), yielding a specificity of 94.3%. More critically, of the 11,833 attack instances, 11,054 were correctly detected (True Positives) with only 779 missed (False Negatives), corresponding to the 6.86% FNR reported in Table 5.

Error Analysis. To understand the model's failure modes, the 779 False Negatives were analysed by attack category. DoS attacks account for 282 FNs (36.2%), predominantly from slow-rate DoS variants (e.g., low-and-slow SYN attacks) characterised by low connection counts that do not trigger volumetric thresholds — a known limitation of traffic-feature-based IDS. Probe attacks account for 150 FNs (19.3%), typically stealth scans with randomised source ports to avoid count-based signatures. R2L attacks account for 263 FNs (33.8%), reflecting the inherent ambiguity between certain R2L patterns (e.g., password guessing over legitimate protocols) and authorised remote access behaviour. U2R attacks, despite their extreme rarity, contribute only 23 FNs (3.0%), validating the ADASYN oversampling strategy for this category. These failure patterns are consistent with known IDS detection limits and indicate that future improvements should focus on temporal sequence modelling (e.g., LSTM layers) to capture slow-rate attack dynamics, rather than on the feature engineering or balancing stages, which are already well optimised.

Feature Importance Distribution

Figure 7 presents the global SHAP summary plot, which aggregates the impact of each feature across all predictions in the test set. The SHAP analysis revealed that network traffic volume metrics (src_bytes, dst_bytes) and connection count statistics (count, srv_count) were consistently the most influential features, with mean absolute SHAP values of 0.42, 0.38, 0.35, and 0.33, respectively. This aligns with domain expertise, as volumetric anomalies are hallmark indicators of DoS attacks and data exfiltration attempts. Protocol-specific flags and service types exhibited moderate but contextually critical importance (SHAP values 0.18-0.25), particularly for distinguishing between attack subcategories.

https://dx.doi.org/10.4314/swj.v21i1.53

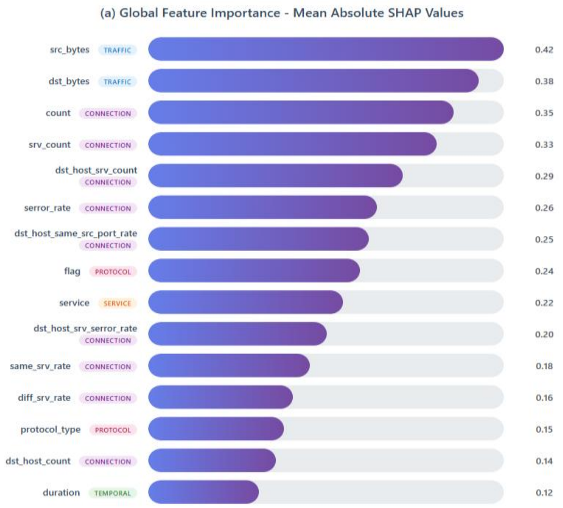


Figure 8: Global SHAP summary plot

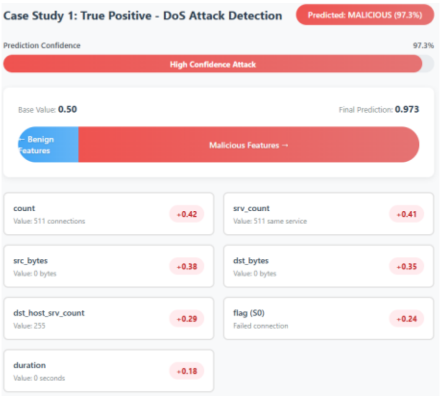
Notably, temporal features such as duration showed lower global importance (SHAP value 0.12). Still, they exhibited high variability across individual predictions, indicating their role in detecting specific attack patterns, such as slow-rate DoS attacks. This nuanced, instance-specific insight is precisely what model-agnostic explainers like SHAP provide that traditional feature importance metrics do not.

Case Study: Explaining High-Confidence Predictions

To demonstrate practical utility, two representative case studies are presented: one true positive (correctly identified attack) and one false positive (benign traffic misclassified as an attack).

Case Study 1: True Positive - DoS Attack Detection

A network connection with the following characteristics was classified as malicious with 97.3% confidence, as shown in Figure 9a:



Case Study 1: True Positive - DoS Attack Detection

Predicted: MALICIOUS (97.3%)

Prediction Confidence 97.3%

High Confidence Attack

Base Value: **0.50**

Final Prediction: **0.973**

count
Value: 511 connections
+0.42

srv_count
Value: 511 same service
+0.41

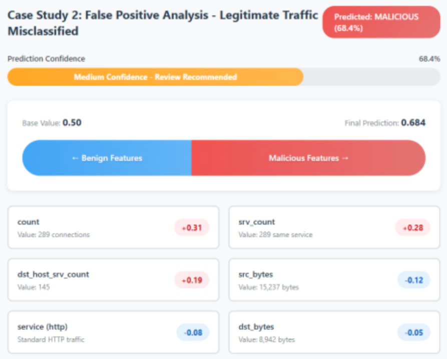
src_bytes
Value: 0 bytes
+0.38

dst_bytes
Value: 0 bytes
+0.35

dst_host_srv_count
Value: 255
+0.29

flag (S0)
Failed connection
+0.24

duration
Value: 0 seconds
+0.18



Case Study 2: False Positive Analysis - Legitimate Traffic Misclassified

Predicted: MALICIOUS (68.4%)

Prediction Confidence 68.4%

Medium Confidence - Review Recommended

Base Value: **0.50**

Final Prediction: **0.684**

count
Value: 289 connections
+0.31

srv_count
Value: 289 same service
+0.28

dst_host_srv_count
Value: 145
+0.19

src_bytes
Value: 15,237 bytes
-0.12

service (http)
Standard HTTP traffic
-0.08

dst_bytes
Value: 8,942 bytes
-0.05

Figure 9: SHAP Analysis a. Case study 1: True Positive b. False Positive

In Figure 9a, the model identified this as a SYN flood DoS attack based on the combination of: (1) zero data transmission (src_bytes=0, dst_bytes=0), (2) extremely high connection counts to the same service (count=511, srv_count=511), (3) failed connection establishment (flag=S0), and (4) zero duration, all occurring within the same time window. The cumulative positive SHAP contribution of +2.27 pushed the prediction strongly toward the malicious class. This explanation directly mirrors the textbook characteristics of a SYN flood attack, providing immediate validation to the security analyst.

Case Study 2: False Positive Analysis: A legitimate HTTP traffic instance was incorrectly classified as malicious with 68.4% confidence (Figure 8b). The model flagged this connection due to unusually high connection counts (count=289), which resembled reconnaissance scanning behavior. However, the legitimate characteristics—substantial bidirectional data transfer (src_bytes, dst_bytes) and standard HTTP service—provided negative SHAP contributions toward benign classification. The net positive SHAP score of +0.53 narrowly tipped the prediction to malicious. Upon analyst review, this traffic pattern was identified as automated web scraping by an authorized monitoring tool, a legitimate use case not adequately represented in the training data. This false positive, while undesirable, demonstrates the critical value of explainability. The SHAP analysis immediately highlighted the competing signals (high connection counts vs. normal data transfer), enabling the analyst to quickly recognize the edge case and allow the source, preventing future alerts. Without explanation, this would have required extensive manual traffic analysis or been dismissed as an inexplicable model error, potentially leading to either wasted investigative time or erosion of trust in the system.

DISCUSSION: IMPLICATIONS FOR CII PROTECTION

The empirical results compellingly demonstrate that the proposed hybrid ensemble framework, integrated with a model-agnostic XAI layer, successfully addresses the critical research triad of detection performance, robustness to data imbalance, and operational transparency. The synergistic value of the stacked architecture—combining a Random Forest's robustness in modeling non-linear feature interactions with a 1D-CNN's capacity for hierarchical spatial pattern extraction, all optimally weighted by a Logistic Regression meta-learner—is definitively evidenced by the ablation study. This architecture, built upon a robust ADASYN+Tomek balancing foundation that confirms and extends the findings of Salihu et al. (2024), achieved a superior detection profile culminating in a False Negative Rate of 6.86% on the NSL-KDD test set. This represents a 30.6% relative reduction in missed attacks compared to the ADASYN+Tomek+RF benchmark, directly mitigating the existential risk of undetected intrusions in CII environments. The model's efficacy was particularly pronounced in detecting rare yet critical U2R and R2L attacks, achieving recall rates of 88.50% and 90.45%, respectively, validating the adaptive focus of the ADASYN pre-processing stage. However, the framework's core contribution transcends mere accuracy; the integration of SHAP provides a novel solution to the "black box" impasse that has historically plagued ML adoption in security operations. As evidenced by the case studies, the XAI layer provides immediate, human-interpretable validation for true positives, such as identifying the textbook characteristics of a SYN flood attack (e.g., high count, flag=S0, src_bytes=0), thereby bolstering analyst trust. More significantly, it offers profound diagnostic insight into false positives, elucidating the conflicting signals (e.g., high connection counts versus legitimate data transfer) that led to a misclassification. This transforms the IDS from an opaque "alert" generator into a collaborative diagnostic tool, empowering analysts to rapidly vet ambiguous alerts, mitigate "alert fatigue," and prevent the erosion of trust in the system. While these findings are promising, the study acknowledges limitations, primarily its reliance on the dated NSL-KDD dataset and the computational overhead of the stacked ensemble and post hoc SHAP explanations. Future work must therefore focus on validating this framework against large-scale, contemporary ICS-specific datasets, optimizing the model for real-time, resource-constrained deployment, and conducting rigorous human-in-the-loop (HITL) evaluations to quantify the tangible impact of XAI-driven explanations on SOC operational metrics such as Mean Time to Detect (MTTD) and Mean Time to Respond (MTTR).

Conclusion

This research confronted the pressing cybersecurity challenges in Critical Information Infrastructure, where traditional Intrusion Detection Systems are ill-equipped to handle novel threats, and modern machine learning solutions are often crippled by data imbalance and a lack of transparency. This paper proposed and validated a novel integrated framework that synergistically combines a hybrid ADASYN+Tomek Links data-balancing technique with a high-performance stacked ensemble model (Random Forest + 1D-CNN + Logistic Regression). The core contribution of this framework is the seamless integration of an SHAP-based Explainable AI (XAI) layer, designed to bridge the critical trust gap between autonomous detection and human-in-the-loop analysis.

The experimental results demonstrated the proposed model's superior performance on the NSL-KDD benchmark dataset, achieving an accuracy of 92.37%, an F1-Score of 92.41%, and an excellent AUC-ROC of 0.9612. Most critically for CII protection, the model achieved a False Negative Rate of just 6.86%. This substantial reduction in missed attacks, especially the robust detection of rare U2R and R2L categories, directly addresses the primary risk in high-stakes environments. The ablation study further confirmed that this high performance was not due to any single component, but rather the synergistic effect of the hybrid balancing and the stacked ensemble architecture.

Furthermore, the XAI component has been proven highly effective at translating the model's complex decisions into actionable insights. Through SHAP-driven case studies, this research demonstrated how the model provides transparent, feature-based rationales for its predictions. This capability allows security analysts to instantly validate true positives, such as a SYN flood attack, and efficiently diagnose false positives by understanding the model's conflicting reasoning. This transforms the IDS from an opaque "black box" into a transparent and trustworthy diagnostic tool, mitigating alert fatigue and empowering rapid, informed incident response.

In conclusion, this work presents a holistic and resilient IDS framework that successfully balances the competing demands of high-accuracy detection, robustness to data imbalance, and operational transparency. While acknowledging the limitations of the NSL-KDD dataset and the computational overhead of the ensemble, this study provides a robust blueprint for next-generation, trustworthy security systems. Future research should focus on validating this framework against modern, large-scale ICS datasets, optimizing the model for real-time performance, and conducting Human-in-the-Loop (HITL) studies to quantify the operational impact of XAI on security analyst workflows.

REFERENCES

Adadi, A., & Berrada, M. (2018). Peeking inside the black box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20–29.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cybersecurity intrusion detection. *IEEE Communications Surveys & Tutorials*, 18(2), 1153–1176.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 1–58.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.

Cozzie, A., Straggone, M., & Cavallaro, L. (2023). A decade of network intrusion detection with Snort: A survey: Computers *& Security*, 124, 102936.

Das, S., Roy, S., & Samanta, S. (2024). SHAP-guided feature selection for lightweight ensemble intrusion detection systems. *Expert Systems with Applications*, 238, 121789.

Department of Homeland Security. (2022). *Critical infrastructure sectors*. Cybersecurity and Infrastructure Security Agency (CISA). https://www.cisa.gov/critical-infrastructure-sectors

Devarajan, M. (2021). A comprehensive analysis of signature-based and anomaly-based intrusion detection systems. *International Journal of Computer Networks and Applications*, 8(2), 87–95.

García-Teodoro, P., Díaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (2018). Anomaly-based network intrusion detection: Techniques, systems, and challenges. *Computers & Security*, 72, 14–28.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, 4(37), eaay7120.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284.

He, H., Bai, Y., García, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008, IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* (pp. 1322-1328). IEEE.

Humayed, A., Lin, J., Li, F., & Luo, B. (2017). Cyber-physical systems security—A survey. *IEEE Internet of Things Journal*, 4(6), 1802-1831.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance—Journal *of Big Data*, 6(1), 27.

Khan, A., Rehman, A., & Ahmad, T. (2024). Real-time explainable intrusion detection for security operations centres: An attention-SHAP approach. *Computers & Security*, 138, 103639.

Lewis, J. A. (2022). *Rethinking cybersecurity: Strategy, economics, and policy*. Center for Strategic and International Studies (CSIS).

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., … & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67.

Mahbooba, B., Timilsina, M., Sahal, R., & Serrano, M. (2022). Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using a decision tree model: complexity, 2022, 6507578.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.

Musa, U. S., Chakraborty, S., Abdullahi, M. M., & Maini, T. (2021). A review of intrusion detection systems using machine learning techniques. In *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)* (pp. 541–549). IEEE.

Odion, P. O., Musa, M. N., Suleiman, T., & Isa, M. M. (2020). Application of Machine Learning Technique for the Prediction of Neonatal Mortality Using Multiple Risk Factors. *FUDMA JOURNAL OF SCIENCES*, *4*(3), 576-582.

Pandey, N., Mishra, A. K., & Kumar, A. (2020). Machine learning-based intrusion detection for IoT and cloud environments: A survey. *IEEE Access*, 8, 134049–134068.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

Salihu, N. I., Musa, M. N., & Awujola, J. O. (2024). A combined approach of ADASYN and TomekLink for an anomaly network intrusion detection system using some selected machine learning algorithms. *International Journal of Web Research*, 7(4), 51–64.

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing, and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from a machine learning perspective. *Journal of Big Data*, 7(1), 1–29.

Shafiq, M., Tian, Z., Bashir, A. K., Du, X., & Guizani, M. (2023). CorrAUC: A malicious bot-IoT traffic detection method using machine-learning techniques with the AUC metric. *IEEE Internet of Things Journal*, 10(4), 3216–3228.

Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp*, 1, 108–116.

Sommer, R., & Paxson, V. (2019). Outside the closed world: On using machine learning for network intrusion detection. *IEEE Security & Privacy*, 8(3), 62–65.

Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. In *2009, IEEE Symposium on Computational Intelligence for Security and Defense Applications* (pp. 1–6). IEEE.

Tomek, I. (1976). Two modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(11), 769–772.

U.S. Government Accountability Office. (2021). *Critical infrastructure protection: Agencies need better ways to manage cybersecurity risks to operational technologies (GAO-21-562)*. https://www.gao.gov/products/gao-21-562

Xiao, Y., Xing, C., Zhang, T., & Zhao, Z. (2022). An intrusion detection model based on feature reduction and deep learning. *IEEE Access*, 10, 108312–108324.

Xiao, Y., Xing, C., Zhang, T., & Zhao, Z. (2022). An intrusion detection model based on feature reduction and deep learning. *IEEE Access*, 10, 108312–108324.

Yin, C., Zhu, Y., Fei, J., & He, X. (2023). A deep learning approach for intrusion detection using recurrent neural networks. *IEEE Transactions on Network and Service Management*, 20(1), 1–14.

Zhou, Y., Han, M., He, J., & Liu, L. (2022). A survey on challenges and methods in explainable artificial intelligence for cybersecurity. *Computers & Security*, 122, 102887.

Zhou, Y., Han, M., He, J., & Liu, L. (2022). A survey on challenges and methods in explainable artificial intelligence for cybersecurity. *Computers & Security*, 122, 1028