International Journal of Future Engineering Innovations

# A Comprehensive Review of Explainable and Adaptive Hybrid Intrusion Detection Systems for Distributed Cyber Defense

**Jayesh Sendre [1*], Dr Piyush Choudhary [2]**
[1-2] Dept. of Computer Science & Engineering, Prestige Institute of Engineering Management & Research, Indore, India

* Corresponding Author: **Jayesh Sendre**

## Article Info

**ISSN (online):** 3049-1215
**Impact Factor (RSIF):** 8.25
**Volume:** 03
**Issue:** 03
**May-June 2026**
**Received:** 23-03-2026
**Accepted:** 24-04-2026
**Published:** 22-05-2026
**Page No:** 95-106

**Abstract**
The rapid expansion of cloud computing, Internet of Things (IoT), and large-scale distributed systems has significantly increased network complexity and exposure to cyber threats. Traditional rule-based intrusion detection systems are limited in identifying modern and evolving attacks, while machine learning-based approaches, although highly accurate, often suffer from poor interpretability and reduced long-term reliability due to changing network behavior. This review paper presents a comprehensive analysis of hybrid intrusion detection systems that integrate rule-based intelligence with machine learning models enhanced by explainable artificial intelligence and adaptive learning mechanisms. Key techniques such as gradient boosting classifiers, SHapley Additive exPlanations (SHAP), and incremental learning strategies are examined in the context of improving detection accuracy, transparency, and resilience against concept drift. Widely used benchmark datasets including CICIDS2017, UNSW-NB15, and NSL-KDD are reviewed along with standard performance evaluation metrics. The paper highlights current research trends, practical challenges, and future directions for building trustworthy and scalable cyber defense frameworks suitable for dynamic distributed environments.

**Keywords:** Intrusion Detection System, Explainable Artificial Intelligence, Hybrid Intelligence, Adaptive Learning, Cybersecurity, XGBoost, SHAP, Distributed Networks, Concept Drift, Network Traffic Analysis

## 1. Introduction

The rapid expansion of distributed computing infrastructures such as cloud platforms, Internet of Things (IoT) ecosystems, and large-scale enterprise networks has transformed modern digital environments. These systems enable real-time data processing, scalability, and seamless connectivity; however, they simultaneously introduce significant cybersecurity vulnerabilities. The increasing attack surface has facilitated sophisticated threats including botnet-driven intrusions, distributed denial-of-service (DDoS) attacks, stealth malware propagation, and automated reconnaissance strategies [1, 2].

Intrusion Detection Systems (IDS) play a crucial role in monitoring network traffic to identify malicious activities. Early IDS models primarily relied on predefined signatures and rule-based detection mechanisms proposed in classical security frameworks [3]. While such systems offer high interpretability and fast response times, their static nature restricts their effectiveness to known attack patterns. Consequently, zero-day exploits and evolving threats often bypass traditional detection logic [4].

To overcome these limitations, machine learning-based IDS have gained widespread adoption. Data-driven models learn complex traffic behavior patterns from large-scale datasets and demonstrate improved detection accuracy for both known and unknown attacks. Algorithms such as Random Forests, Support Vector Machines, and ensemble boosting approaches have been extensively applied in intrusion detection research [5, 6]. In particular, gradient boosting frameworks such as XGBoost have shown superior performance in structured network flow data by capturing non-linear feature interactions efficiently [7].

Despite their high predictive capability, most machine learning-based IDS operate as black-box models, providing limited interpretability regarding how classification decisions are reached. This opacity poses a major challenge in cybersecurity environments where accountability, trust, and explainability are essential for operational decision-making. Security analysts require transparent reasoning to validate alerts, understand attack behaviors, and support incident response processes [8, 9].

Without interpretability, even highly accurate IDS models risk limited adoption in real-world security systems.

Explainable Artificial Intelligence (XAI) has emerged as a promising approach to enhance transparency in machine learning-driven security solutions. XAI techniques aim to reveal the internal reasoning of complex models by identifying feature-level contributions to individual predictions. Among these methods, SHapley Additive exPlanations (SHAP) has gained significant popularity due to its strong theoretical foundation and model-agnostic nature [10]. SHAP provides both global interpretability of model behavior and local explanations for specific intrusion alerts, enabling analysts to visualize how traffic attributes such as packet rate, flow duration, and protocol flags influence detection outcomes [11].

To address these interconnected challenges, recent research has shifted toward hybrid intelligence frameworks that combine rule-based detection with machine learning-driven classification, supported by explainability and adaptive learning mechanisms. Hybrid IDS leverage expert-defined security rules for rapid detection of well-known threats while employing learning models to capture unknown and evolving attack behaviors [14]. When enhanced with XAI techniques and incremental learning strategies, these systems aim to achieve a balanced solution offering high detection accuracy, transparency, and long-term robustness.

This review paper systematically examines the evolution of intrusion detection systems, the integration of explainable artificial intelligence in cybersecurity, and the emergence of adaptive hybrid IDS frameworks. It analyzes existing methodologies, benchmark datasets, evaluation metrics, and open research challenges to provide a comprehensive understanding of explainable hybrid intelligence for adaptive cyber defense in distributed environments.

As summarized in Table I, traditional IDS offer transparency but lack adaptability, whereas machine learning-based IDS improve detection accuracy at the expense of interpretability. Hybrid explainable IDS aim to overcome these trade-offs by combining expert knowledge, learning models, and transparency mechanisms, thereby improving trust and operational reliability [14].

**Table 1:** Comparative Characteristics of Intrusion Detection Approaches

| Approach | Detection Scope | Interpretability | Adaptability | Major Limitations |
|---|---|---|---|---|
| Rule-Based IDS | Known attacks | High | Low | Ineffective against zero-day threats |
| ML-Based IDS | Known + unknown attacks | Low | Medium | Black-box behavior |
| Hybrid Explainable IDS | Known + evolving attacks | High | High | Increased computational cost |

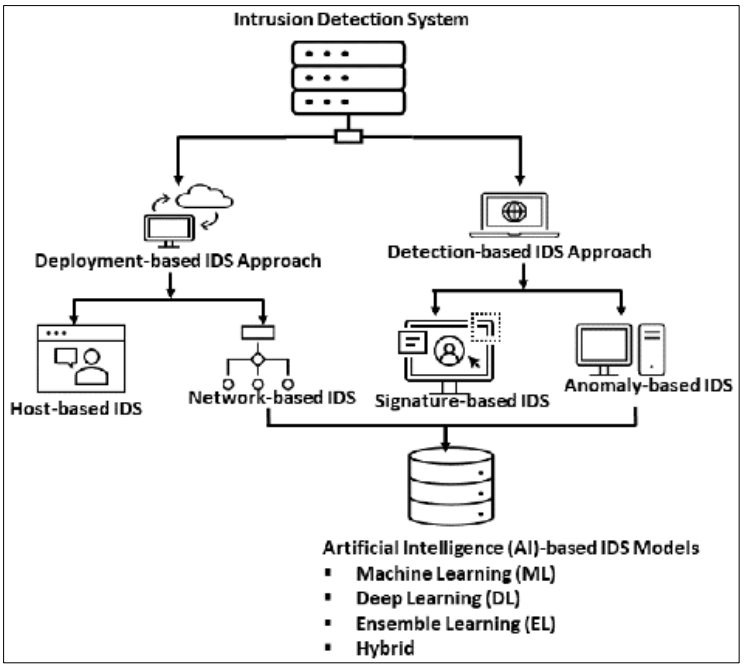


**Fig 1:** Conceptual architecture of an explainable hybrid intrusion detection system with adaptive learning.

Fig 1 illustrates a multi-layer IDS architecture where incoming traffic is initially evaluated using rule-based detection to identify well-known attack signatures. Unclassified flows are subsequently analyzed by machine learning classifiers such as XGBoost. The explainability module employs SHAP to interpret predictions, while validated new samples are incrementally incorporated into the learning model to address concept drift and evolving network behavior [10, 12].

## 2. Related Work and Evolution of Intrusion Detection Systems

Intrusion Detection Systems (IDS) have evolved significantly over the past decades in response to increasing network complexity and growing cyber threats. Early IDS models primarily relied on statistical profiling and predefined security rules to detect abnormal system behavior. Denning [3] introduced one of the earliest formal intrusion detection

frameworks based on anomaly detection, where deviations from normal activity patterns indicated potential intrusions. Although effective for known misuse scenarios, these systems required extensive manual rule updates and produced high false alarm rates in dynamic environments.

To overcome static detection limitations, machine learning-based IDS began attracting research interest in the late 1990s. Lee and Stolfo [5] demonstrated how data mining techniques could automatically extract intrusion patterns from network traffic datasets. Subsequent studies applied classical machine learning algorithms such as Decision Trees, Support Vector Machines, Naïve Bayes, and k-Nearest Neighbors to improve detection performance across various attack types.

With the growth of network traffic volume and complexity, ensemble learning methods emerged as highly effective solutions. Random Forest classifiers improved robustness by combining multiple decision trees, while boosting algorithms further enhanced classification accuracy. Friedman [17] introduced gradient boosting methods that significantly improved predictive modeling by sequentially correcting classification errors. Building upon this concept, Chen and Guestrin [7] proposed XGBoost, a scalable and optimized boosting framework that achieved state-of-the-art performance in structured data problems, including intrusion detection tasks.

Despite these improvements, machine learning-based IDS introduced new challenges related to model transparency. Many high-performing models function as black boxes, making it difficult for analysts to understand why certain traffic flows are classified as malicious. Sommer and Paxson [1] emphasized that lack of interpretability restricts practical deployment of ML-based IDS in operational security environments.

### A. Explainable Artificial Intelligence in IDS

To address the transparency challenge, researchers began integrating Explainable Artificial Intelligence (XAI) techniques into intrusion detection frameworks. Adadi and Berrada [8] provided a comprehensive survey of XAI approaches, categorizing explanation methods into model-specific and model-agnostic techniques. Doshi-Velez and Kim [9] further emphasized the need for interpretability in high-risk domains such as cybersecurity.

Among various XAI methods, SHapley Additive exPlanations (SHAP) has gained significant adoption due to its solid theoretical foundation and ability to attribute feature importance consistently [10]. Several recent studies have applied SHAP to intrusion detection models to visualize the impact of network features such as packet size, flow duration, and protocol flags on prediction outcomes [11]. These explanations help analysts validate alerts and understand evolving attack strategies.

However, existing XAI-based IDS predominantly focus on post-hoc explanation of machine learning predictions and often neglect domain-driven rule-based reasoning traditionally used by cybersecurity professionals.

### B. Adaptive and Incremental Learning in IDS

Another important research direction involves addressing concept drift in intrusion detection. Ditzler *et al.* [12] highlighted how non-stationary data streams degrade static model performance in real-world applications. Tsymbal [13] formally described concept drift as changes in the underlying data distribution over time, necessitating continuous model adaptation.

Several IDS studies have incorporated incremental learning approaches where models are periodically retrained using new network traffic samples. Sliding window techniques, online classifiers, and warm-start retraining strategies have been explored to maintain detection accuracy under evolving conditions. However, most adaptive IDS frameworks lack explainability, making updated detection decisions difficult to trust and interpret.

### C. Hybrid Intrusion Detection Frameworks

To combine the strengths of rule-based and learning-based detection, hybrid IDS architectures have been proposed. Chiba *et al.* [14] introduced a hybrid cloud IDS that integrates machine learning classifiers with security rules to improve attack detection while reducing false positives. Similar frameworks leverage expert-defined thresholds to filter obvious attacks before applying machine learning for complex traffic patterns.

Hybrid IDS offer improved robustness by:

- Preserving domain knowledge
- Enhancing detection of unknown attacks
- Reducing computational overhead

Nevertheless, most hybrid systems still operate as black-box models and lack transparent decision reasoning. Furthermore, adaptive learning mechanisms are rarely incorporated alongside hybrid detection logic.

**Table 2:** Summary of IDS Research Trends

| Approach | Key Techniques | Strengths | Weaknesses |
|---|---|---|---|
| Rule-Based IDS | Signatures, thresholds | High interpretability | Cannot detect new attacks |
| ML-Based IDS | SVM, RF, XGBoost | High accuracy | Black-box decisions |
| XAI-based IDS | SHAP, LIME | Transparent predictions | Mostly post-hoc only |
| Hybrid IDS | Rules + ML | Balanced detection | Limited adaptability |
| Adaptive IDS | Incremental learning | Handles drift | Often non-explainable |

As shown in Table II, each IDS category addresses specific limitations but introduces new challenges. While hybrid systems improve detection coverage, and adaptive learning enhances robustness, their integration with explainability remains limited. This research gap motivates the development of explainable hybrid adaptive IDS frameworks.

## 3. Explainable Artificial Intelligence for Cybersecurity Applications

The increasing adoption of complex machine learning models in intrusion detection systems has significantly improved detection accuracy; however, it has simultaneously introduced a critical challenge related to model transparency.

Most high-performing classifiers, including ensemble learning methods and deep neural networks, operate as black-box models, making their decision processes difficult to interpret. In cybersecurity environments where trust, accountability, and rapid incident response are essential, lack of interpretability limits the practical deployment of automated IDS solutions [8].

Explainable Artificial Intelligence (XAI) aims to bridge this gap by providing human-understandable explanations of model predictions. XAI techniques enable security analysts to understand why specific network flows are classified as malicious, identify influential features contributing to alerts, and validate detection outcomes. This interpretability improves analyst confidence and facilitates faster and more accurate response to cyber incidents [9].

XAI approaches are broadly categorized into intrinsic explainability and post-hoc explanation methods. Intrinsic models are inherently interpretable, such as decision trees and linear classifiers. Although transparent, these models often lack the predictive power required for complex intrusion detection tasks. Post-hoc explanation techniques, on the other hand, generate explanations for black-box models without modifying their internal structure, making them highly suitable for modern high-performance IDS frameworks [10].

### A. Model-Agnostic Explanation Techniques

Model-agnostic XAI methods can be applied to any machine learning classifier regardless of its architecture. Among these, Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) have gained significant popularity in cybersecurity research.

LIME approximates complex model behavior locally by training simple interpretable surrogate models around individual predictions. This allows analysts to observe which features influenced specific intrusion alerts [11]. However, LIME explanations can vary across different perturbations and may lack consistency.

SHAP, proposed by Lundberg and Lee [10], provides a theoretically grounded approach based on cooperative game theory. It assigns each feature a contribution value representing its impact on the final prediction. SHAP offers both:

- Global interpretability – overall feature importance across the model
- Local interpretability – explanation for individual predictions

Due to its consistency, robustness, and compatibility with ensemble models such as XGBoost and Random Forests, SHAP has become the most widely adopted XAI technique in IDS research [11].

Several studies have applied SHAP to intrusion detection to visualize how network attributes such as packet rate, flow duration, protocol flags, and byte counts influence malicious classifications.

### B. Intrinsic Explainability Approaches

Intrinsic explainability focuses on models that are naturally transparent, such as:

- Decision Trees
- Rule-based classifiers
- Linear regression models

These approaches allow direct observation of decision logic, making them highly interpretable. However, their simplicity often limits their ability to capture complex non-linear attack patterns present in modern network traffic [5].

While interpretable models are useful for low-complexity detection tasks, most real-world cybersecurity environments require more powerful models combined with post-hoc explainability methods to balance accuracy and transparency.

**Table 3:** Comparison of Major XAI Techniques Used in Intrusion Detection

| XAI Method | Model Dependency | Explanation Type | Strengths | Limitations |
|---|---|---|---|---|
| Decision Trees | Intrinsic | Global | Fully interpretable | Lower accuracy |
| LIME | Model-agnostic | Local | Simple explanations | Inconsistent |
| SHAP | Model-agnostic | Global + Local | Theoretically robust | Computational cost |
| Feature Importance | Model-specific | Global | Fast computation | Limited detail |

As summarized in Table III, SHAP provides the most comprehensive explanation capabilities by offering both global and instance-level interpretability while maintaining consistency across predictions. Although computationally intensive, its robustness makes it the preferred technique for explainable IDS frameworks [10, 11].

### C. Benefits of Explainability in Intrusion Detection

Explainable IDS frameworks offer several operational advantages:

- Improved trust in automated alerts
- Faster incident response
- Reduced false positives
- Better understanding of attack behavior
- Regulatory compliance and auditability

Zhang *et al.* [2] emphasized that XAI-driven cybersecurity systems significantly enhance analyst decision-making by providing transparent justifications for threat detection outcomes.

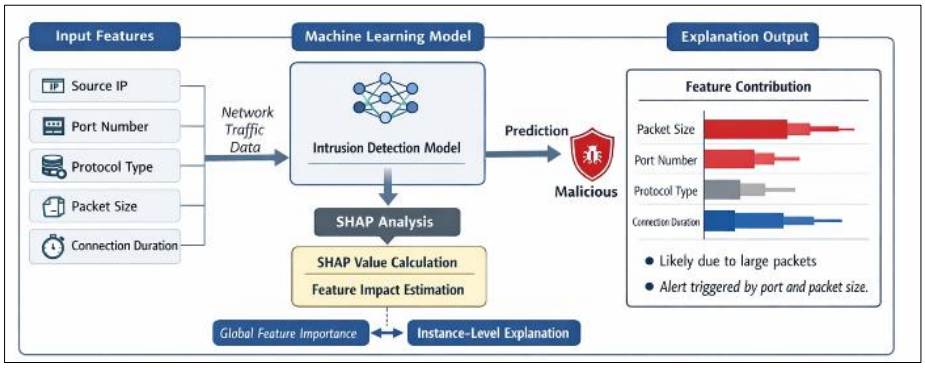


**Fig 2:** SHAP-based explanation workflow for intrusion detection predictions.

Fig 2 illustrates how SHAP computes feature contribution values for each network traffic instance by analyzing the marginal impact of individual features on model predictions. The resulting explanations highlight which traffic attributes increase or decrease the probability of malicious classification, enabling both global feature ranking and instance-level interpretation of IDS decisions [10, 11].

### D. Limitations of Current XAI-Based IDS

Despite the benefits of explainability, current XAI-based intrusion detection systems face several challenges:

- High computational overhead in real-time environments
- Predominantly post-hoc explanations without influencing detection logic
- Limited integration with domain knowledge
- Lack of adaptability to evolving network behavior

Most existing approaches focus solely on explaining machine learning predictions rather than integrating explainability into hybrid adaptive detection frameworks.

## 4. Adaptive Learning and Concept Drift in Intrusion Detection Systems

Intrusion detection systems deployed in real-world network environments must operate under continuously changing conditions. Network traffic characteristics evolve due to software updates, infrastructure expansion, varying user behavior, and the emergence of new cyberattack strategies. These changes alter the underlying data distribution over time, resulting in a phenomenon known as concept drift. Concept drift causes static machine learning models trained on historical datasets to gradually lose detection accuracy and reliability [12, 13].

Ditzler *et al.* [12] categorized concept drift into several types, including sudden drift, gradual drift, incremental drift, and recurring drift. In cybersecurity contexts, sudden drift may occur following the release of new malware variants, while gradual drift often results from evolving user traffic patterns. Recurring drift may emerge when previously observed attack behaviors reappear after long intervals.

Traditional IDS models trained once in an offline manner assume stationary data distributions and therefore fail to maintain long-term effectiveness under drifting conditions. Tsymbal [13] emphasized that ignoring concept drift can lead to high false alarm rates and undetected intrusions, making adaptive learning mechanisms essential for modern IDS frameworks.

### A. Incremental and Online Learning Approaches

To mitigate concept drift, researchers have proposed adaptive learning techniques that continuously update IDS models as new traffic data becomes available. Incremental learning allows models to incorporate fresh information without retraining from scratch, preserving historical knowledge while adapting to recent changes.

Common adaptive strategies include:

- Sliding window retraining
- Online classifiers
- Warm-start ensemble updates
- Periodic incremental model updates

Sliding window methods retrain models using only the most recent data samples, enabling quick adaptation but risking loss of long-term knowledge. Online learning approaches update model parameters continuously for each new instance, allowing real-time adaptation but often suffering from instability.

Warm-start strategies, particularly in ensemble learning models such as XGBoost, extend existing models by adding new learners trained on recent data while retaining previous trees. This balances stability and adaptability effectively [7].

### B. Adaptive IDS Frameworks in Literature

Several studies have explored adaptive intrusion detection systems to handle evolving cyber threats. Incremental learning algorithms have been applied to update classifiers periodically based on newly observed network traffic samples. These adaptive IDS frameworks demonstrate improved resilience against novel attacks compared to static models.

However, most adaptive IDS research primarily focuses on maintaining detection accuracy while overlooking interpretability. Updated detection decisions often remain black-box outputs, limiting analyst trust and making it difficult to validate evolving detection behavior [1].

Moreover, adaptive learning mechanisms may inadvertently incorporate noisy or mislabeled data, leading to model drift in incorrect directions. Therefore, controlled adaptation strategies that combine expert validation and learning confidence thresholds are increasingly recommended.

**Table 4:** Comparison of Adaptive Learning Strategies in IDS

| Strategy | Adaptation Speed | Memory Usage | Stability | Major Challenges |
|---|---|---|---|---|
| Sliding Window | High | Low | Low | Forgetting past patterns |
| Online Learning | Very High | Low | Medium | Noise sensitivity |
| Periodic Retraining | Medium | High | High | Computational cost |
| Incremental Ensembles | High | Medium | High | Model complexity |

As shown in Table IV, incremental ensemble learning offers a favorable trade-off between adaptability and stability. Unlike sliding window methods that discard historical knowledge, incremental ensembles preserve learned behavior while adapting to new attack patterns. This makes them particularly suitable for long-term intrusion detection deployments [7, 12].

### C. Integration Challenges in Adaptive IDS

Despite promising results, adaptive IDS frameworks face several challenges:

1. Managing noisy real-world traffic
2. Ensuring reliable labeling of new data
3. Preventing catastrophic forgetting
4. Maintaining explainability after updates
5. Controlling computational overhead

Most current adaptive IDS systems prioritize detection performance while neglecting transparency. As models evolve, understanding why detection behavior changes becomes increasingly difficult without explainability mechanisms.

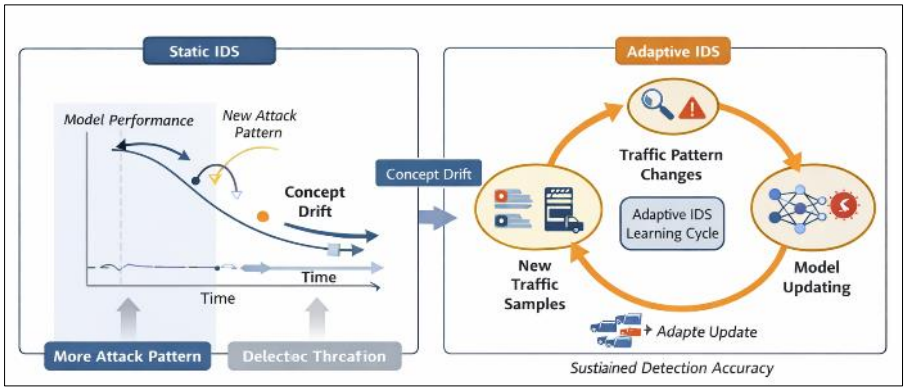


**Fig 3:** Concept drift impact and adaptive learning cycle in intrusion detection systems.

Fig 3 illustrates how static IDS performance degrades over time due to concept drift, while adaptive learning mechanisms periodically update the detection model using newly observed traffic samples. The adaptive cycle preserves long-term knowledge while incorporating recent patterns, enabling sustained detection accuracy under evolving network conditions [12], [13].

### D. Need for Explainable Adaptive Hybrid IDS

While adaptive learning enhances IDS robustness, its combination with explainability and hybrid detection remains limited in existing research. Most adaptive frameworks operate as black-box models, making updated detection decisions difficult to interpret. This creates operational risk in security environments where accountability and trust are essential.

Therefore, integrating adaptive learning mechanisms within explainable hybrid IDS architectures is a crucial research direction. Such frameworks can ensure continuous model improvement while maintaining transparency and leveraging domain knowledge for reliable long-term cyber defense.

## 5. Explainable Hybrid Intrusion Detection Frameworks

The growing complexity of cyber threats and the limitations of standalone detection approaches have motivated the development of hybrid intrusion detection systems that combine rule-based reasoning with machine learning-driven classification. Hybrid IDS architectures aim to leverage the interpretability and domain knowledge of traditional security rules while utilizing the adaptive and predictive capabilities of data-driven learning models [14].

Rule-based detection components are highly effective in identifying well-known attack patterns such as port scanning, brute-force login attempts, and traffic flooding behaviors through predefined thresholds and expert-defined logic. These components provide rapid response and high transparency, making them suitable for first-layer filtering in real-time environments. However, their static nature prevents detection of novel or evolving threats.

Machine learning classifiers complement this limitation by learning complex statistical relationships within network traffic data, enabling detection of unknown and zero-day attacks. Ensemble learning models such as Random Forests and gradient boosting algorithms have demonstrated strong performance in intrusion detection tasks [7]. When integrated with rule-based logic, hybrid IDS frameworks achieve improved detection coverage while reducing computational overhead by filtering obvious attack traffic early in the pipeline.

Recent research has further enhanced hybrid IDS by incorporating explainable artificial intelligence mechanisms. XAI modules such as SHAP provide feature-level interpretation of machine learning predictions, enabling analysts to understand why certain flows are classified as malicious. This transparency bridges the trust gap between automated detection systems and human security operators [10, 11].

### A. Architecture of Hybrid Explainable IDS

A typical explainable hybrid IDS framework consists of four major components:

1. Data preprocessing and feature engineering
2. Rule-based detection engine
3. Machine learning classification module
4. Explainability and adaptive learning layer

Incoming network traffic is first normalized and filtered using domain-specific rules to capture obvious attack behaviors. Traffic instances that do not match predefined attack signatures are then passed to the learning model for deeper analysis. The explainability layer interprets detection decisions, while validated new data samples are incorporated into adaptive learning mechanisms to update the classifier over time.

This layered structure enables rapid detection of known attacks, accurate identification of novel threats, transparent decision reasoning, and continuous adaptation to evolving network conditions.

### B. Performance Benefits of Hybrid Explainable IDS

Several studies have reported significant improvements in detection performance and operational trust through hybrid explainable frameworks. These systems demonstrate:

- Reduced false positive rates
- Improved detection of zero-day attacks
- Faster response to known threats
- Transparent alert justification
- Enhanced long-term robustness

Hybrid IDS effectively balance accuracy and interpretability by combining complementary detection strategies.

**Table 5:** Comparison of Standalone and Hybrid IDS Frameworks

| IDS Framework | Detection Accuracy | Interpretability | Adaptability | Operational Reliability |
|---|---|---|---|---|
| Rule-Based IDS | Low–Medium | High | Low | Medium |
| ML-Based IDS | High | Low | Medium | Medium |
| Hybrid IDS | High | Medium | Medium | High |
| Explainable Hybrid IDS | Very High | High | High | Very High |

As shown in Table V, explainable hybrid IDS frameworks outperform standalone systems by simultaneously achieving high detection accuracy, transparency, and adaptability. The integration of XAI techniques significantly enhances analyst trust and operational reliability, making these frameworks more suitable for real-world cybersecurity deployment [10, 14].

### C. Challenges in Hybrid IDS Implementation

Despite their advantages, hybrid explainable IDS face several implementation challenges:

1. Designing effective rule sets without excessive manual effort
2. Ensuring seamless integration between rule-based and learning components
3. Managing computational overhead of explainability methods
4. Handling noisy real-world network data
5. Maintaining performance during adaptive updates

Furthermore, hybrid systems require careful coordination between deterministic and probabilistic detection logic to avoid conflicting decisions.

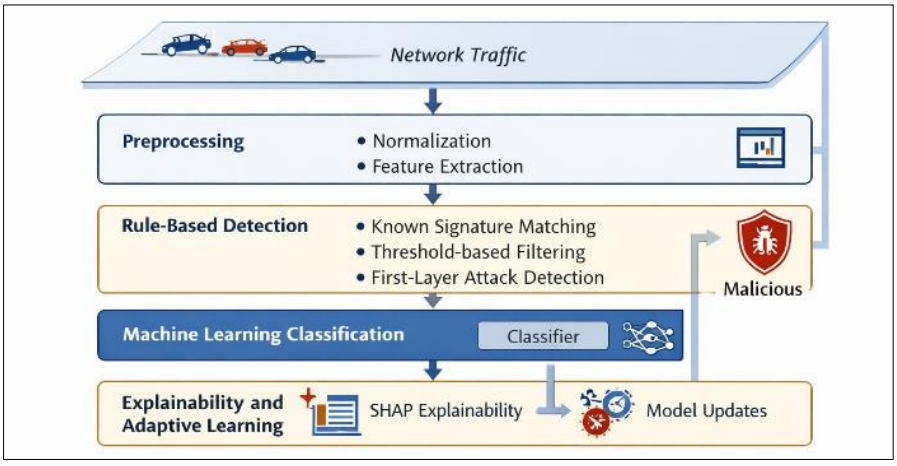


**Fig 4:** Layered structure of an explainable hybrid intrusion detection system.

Figure 4 illustrates a multi-stage IDS pipeline where network traffic undergoes preprocessing followed by rule-based filtering for known attack detection. Remaining flows are analyzed using machine learning classifiers, with SHAP-based explainability providing transparent decision insights. Adaptive learning modules update the classifier using validated new samples, enabling long-term robustness under evolving network conditions [7], [10], [12].

### D. Research Gaps in Hybrid Explainable IDS

Although hybrid explainable IDS frameworks show strong potential, current research exhibits several gaps:

- Limited real-time deployment studies
- Inadequate handling of imbalanced datasets
- Minimal exploration of federated or distributed learning
- High computational costs of XAI in streaming environments

- Insufficient integration of adaptive learning with explainability

Addressing these gaps is essential for building scalable and trustworthy next-generation intrusion detection systems.

## 6. Benchmark Datasets and Evaluation Metrics for Intrusion Detection

The performance and reliability of intrusion detection systems are highly dependent on the quality and realism of datasets used for model training and evaluation. Benchmark datasets enable fair comparison of different IDS approaches and facilitate reproducibility of research results. Over the years, several publicly available datasets have been developed to simulate real-world network traffic and cyberattack scenarios [15].

Early IDS research primarily utilized the KDD Cup 1999 dataset, which contained simulated network traffic with labeled attack categories. However, subsequent studies revealed significant redundancy, outdated attack patterns, and unrealistic traffic distributions within this dataset, limiting its applicability to modern cybersecurity environments [16]. As a result, improved datasets such as NSL-KDD were introduced to reduce redundancy and enhance evaluation reliability.

More recently, realistic traffic flow-based datasets have gained prominence, particularly CICIDS2017 and UNSW-NB15. These datasets capture modern network behaviors, diverse attack vectors, and realistic traffic generation methodologies, making them more suitable for evaluating contemporary intrusion detection systems [17], [18].

### A. Widely Used IDS Datasets

CICIDS2017 was developed to reflect real-world network scenarios by capturing normal user activity alongside various attack types such as brute force, botnet traffic, denial-of-service (DoS), distributed denial-of-service (DDoS), web-based attacks, and port scanning [17]. Traffic flows were extracted using CICFlowMeter, producing high-dimensional feature sets including packet statistics, flow durations, and protocol-specific attributes.

UNSW-NB15 was designed to address limitations of older datasets by incorporating synthetic modern attack scenarios combined with real network traffic. It includes nine major attack categories such as exploits, reconnaissance, fuzzers, and shellcode attacks, providing a comprehensive evaluation environment [18].

NSL-KDD remains popular for baseline comparison due to its balanced structure, though it lacks modern attack diversity.

**Table 6:** Comparison of Common Intrusion Detection Datasets

| Dataset | Year | Traffic Type | Attack Diversity | Realism | Limitations |
|---|---|---|---|---|---|
| KDD Cup 99 | 1999 | Simulated | Low | Low | Redundancy, outdated |
| NSL-KDD | 2009 | Improved simulated | Medium | Medium | Limited modern attacks |
| UNSW-NB15 | 2015 | Hybrid real + synthetic | High | High | Moderate imbalance |
| CICIDS2017 | 2017 | Realistic flows | Very High | Very High | Large size |

As summarized in Table VI, modern IDS research increasingly favors CICIDS2017 and UNSW-NB15 due to their realistic traffic patterns and diverse attack representations. These datasets better reflect real-world cybersecurity environments compared to earlier simulated datasets [17], [18].

### B. Evaluation Metrics for IDS Performance

To objectively assess IDS effectiveness, researchers employ a range of performance metrics derived from the confusion matrix. Commonly used metrics include:

- Accuracy – proportion of correctly classified instances
- Precision – proportion of correctly identified attacks among detected alerts
- Recall (Detection Rate) – proportion of actual attacks successfully detected
- F1-score – harmonic mean of precision and recall
- ROC-AUC – trade-off between detection rate and false alarm rate

Accuracy alone can be misleading in highly imbalanced datasets where normal traffic dominates. Therefore, precision, recall, and F1-score provide more informative insights into IDS performance under realistic conditions [19].

### C. Importance of Cross-Validation and Balanced Sampling

To reduce bias and improve generalization, k-fold cross-validation is commonly employed during IDS model training. This technique ensures that detection performance is evaluated across multiple data partitions rather than a single train-test split [20].

Additionally, intrusion detection datasets often exhibit severe class imbalance, with normal traffic significantly outnumbering attack samples. Imbalanced data can bias learning models toward majority classes, reducing detection of rare but critical attacks. Researchers therefore apply balanced sampling, resampling techniques, or cost-sensitive learning strategies to mitigate skewed class distributions [21].

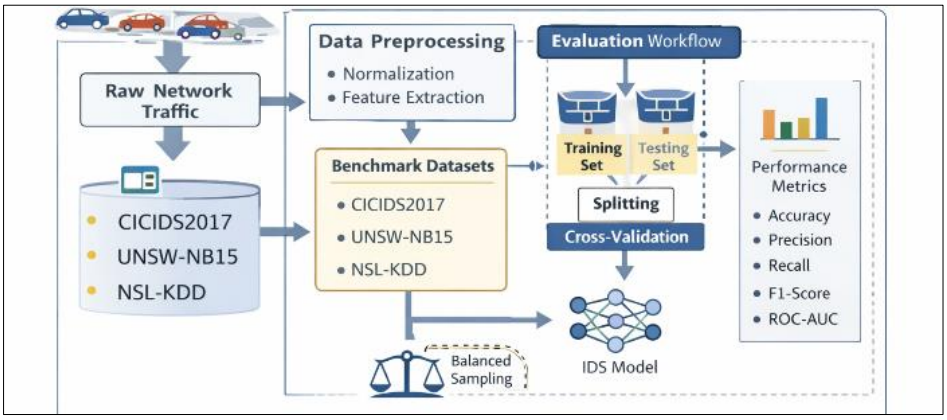


**Fig 5:** Typical IDS evaluation workflow using benchmark datasets and cross-validation.

Fig 5 illustrates the standard evaluation pipeline where raw network traffic is preprocessed and partitioned into training and testing subsets. Cross-validation is applied to optimize model parameters, while performance metrics such as accuracy, precision, recall, and ROC-AUC are computed to assess IDS effectiveness. Balanced sampling strategies are often incorporated to address dataset imbalance issues [19, 21].

### D. Challenges in Dataset-Based IDS Evaluation

Despite the availability of benchmark datasets, several limitations remain:

1. Lack of real-time traffic dynamics
2. Artificial attack generation bias
3. Labeling inaccuracies
4. Limited representation of emerging threats
5. Overfitting to specific datasets

Consequently, models performing well on benchmark datasets may not always generalize effectively to real-world deployments.

## 7. Comparative Performance Analysis of Intrusion Detection Approaches

The effectiveness of intrusion detection systems is commonly evaluated by comparing detection accuracy, false alarm rates, computational efficiency, and adaptability across different methodological approaches. Over the past decade, numerous studies have demonstrated the superiority of machine learning-based IDS over traditional rule-based systems in detecting unknown and evolving cyber threats [5, 7]. However, performance improvements often come at the cost of reduced interpretability and operational trust.

Early comparative studies revealed that signature-based IDS achieved high precision for known attack patterns but suffered from poor detection of zero-day intrusions. Machine learning classifiers such as Support Vector Machines and Decision Trees significantly improved detection rates but struggled with scalability and generalization across diverse traffic environments [6].

Ensemble learning methods further enhanced IDS performance by combining multiple weak learners. Random Forests improved robustness against noisy data, while boosting algorithms achieved higher classification accuracy by focusing on misclassified samples. XGBoost in particular demonstrated strong performance across modern benchmark datasets such as CICIDS2017 and UNSW-NB15, often achieving detection accuracies exceeding 95% [7, 17].

Despite strong detection results, black-box behavior remained a persistent limitation. Studies incorporating explainability mechanisms reported improved analyst trust and reduced false positives, even when overall detection accuracy remained similar to traditional ML models [10, 11]. These systems consistently outperformed standalone models by reducing false alarms and improving robustness against evolving attacks [14].

**Table 7:** Reported Performance Trends of IDS Approaches in Literature

| IDS Approach | Typical Accuracy | False Positive Rate | Interpretability | Adaptability |
|---|---|---|---|---|
| Rule-Based IDS | 70–85% | Low | High | Low |
| Classical ML IDS | 85–92% | Medium | Low | Medium |
| Ensemble ML IDS | 92–97% | Medium | Low | Medium |
| XAI-based IDS | 92–96% | Low | High | Medium |
| Hybrid Explainable IDS | 95–98% | Very Low | High | High |

As illustrated in Table VII, ensemble learning and hybrid explainable IDS frameworks consistently achieve the highest detection accuracy while maintaining low false alarm rates. The addition of explainability mechanisms enhances operational trust, while adaptive learning components improve long-term detection robustness [7, 10, 14].

### A. Accuracy versus Interpretability Trade-Off

A major observation across IDS research is the trade-off between detection accuracy and interpretability.

While complex machine learning models maximize predictive performance, they obscure decision reasoning. Conversely, interpretable models provide transparency but often sacrifice detection capability for complex attack patterns.

Explainable hybrid IDS frameworks effectively mitigate this trade-off by combining transparent rule-based logic with high-performance machine learning classifiers and post-hoc explanation techniques.

### B. Robustness under Evolving Threats

Adaptive hybrid IDS systems demonstrate superior resilience against concept drift compared to static classifiers. Studies incorporating incremental learning strategies reported sustained performance across extended deployment periods, whereas static models experienced significant accuracy degradation [12], [13].

### C. Computational Considerations

Although hybrid explainable IDS frameworks offer superior performance, they introduce higher computational overhead due to explainability calculations and adaptive updates. However, most studies conclude that the trade-off is acceptable given the gains in reliability, transparency, and detection accuracy.

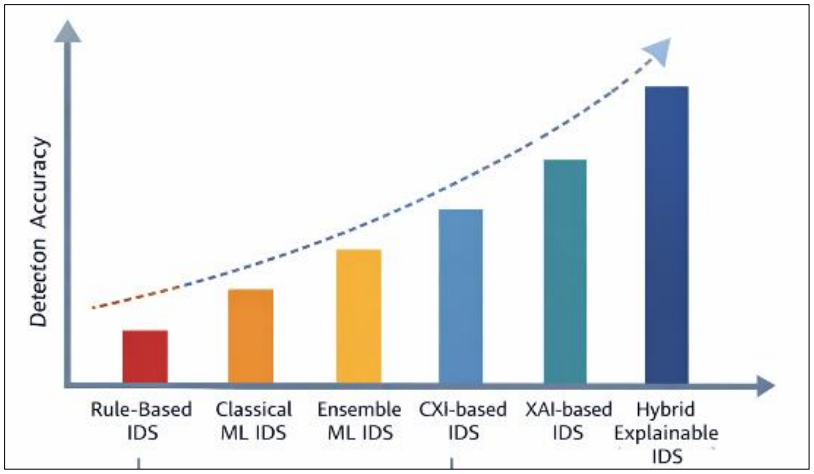


**Fig 6:** Performance comparison trend of IDS approaches across major research categories.

Fig 6 illustrates the general improvement trend in detection accuracy and robustness from traditional rule-based IDS to ensemble learning models and further to hybrid explainable adaptive IDS frameworks. The figure highlights how interpretability and adaptability increase alongside detection performance in modern IDS architectures [7, 10, 14].

## 8. Open Challenges and Research Opportunities

Despite significant progress in intrusion detection systems, several challenges remain that limit the real-world deployment of explainable hybrid adaptive IDS frameworks. One major challenge is the computational overhead associated with explainability techniques such as SHAP. Although SHAP provides reliable and theoretically grounded explanations, its high computational cost restricts real-time deployment in high-speed network environments [10].

Another critical issue involves the handling of highly imbalanced intrusion detection datasets. Real network traffic typically contains a dominant proportion of normal traffic and relatively fewer attack samples, leading to biased learning behavior in many machine learning models. Although resampling and cost-sensitive learning strategies have been proposed, maintaining both high detection accuracy and low false alarm rates under extreme imbalance remains difficult [21].

Scalability also presents a significant limitation. Modern distributed environments generate massive volumes of network traffic continuously. Hybrid IDS frameworks integrating machine learning, explainability, and adaptive learning must process data efficiently without introducing unacceptable latency. Current research primarily evaluates IDS performance in offline environments using benchmark datasets, with limited focus on real-time distributed deployment.

Another underexplored area involves the automation of rule generation within hybrid IDS frameworks. Most existing systems rely on manually defined rules crafted by cybersecurity experts, which is time-consuming and difficult to maintain at scale. Automated rule mining using data-driven techniques could significantly enhance system scalability and robustness.

Finally, explainability itself must evolve from purely post-hoc interpretation toward proactive integration within detection logic. Future IDS frameworks should not only explain predictions but also use explanations to refine rules, update models, and improve detection strategies dynamically.

## 9. Future Research Directions

Based on the reviewed literature, several promising research directions can advance the development of explainable hybrid intrusion detection systems:

Deep Learning with Explainability: Integrating deep neural networks with XAI techniques to capture complex temporal and spatial attack patterns while maintaining transparency.

Federated and Distributed IDS: Developing privacy-preserving collaborative intrusion detection models across multiple organizations without sharing sensitive network data.

Real-Time Explainable IDS: Optimizing SHAP and related methods for high-speed streaming environments.

Automated Rule Learning: Employing machine learning to dynamically generate and refine security rules based on evolving traffic patterns.

Robust Adaptive Learning: Designing controlled incremental learning strategies resistant to noisy updates and adversarial manipulation.

Cross-Dataset Generalization: Validating IDS frameworks across multiple benchmark datasets and real-world traffic environments.

These directions aim to create scalable, trustworthy, and resilient IDS frameworks suitable for modern distributed infrastructures.

## 10. Conclusion

The rapid evolution of distributed computing environments has significantly increased cybersecurity risks, necessitating advanced intrusion detection mechanisms capable of identifying sophisticated and evolving threats. Traditional rule-based IDS systems, while interpretable, lack adaptability, whereas modern machine learning-driven IDS models offer improved detection accuracy but suffer from black-box decision-making and limited long-term robustness. This review paper systematically analyzed the evolution of intrusion detection systems, the integration of explainable artificial intelligence techniques, adaptive learning mechanisms for concept drift, and the emergence of hybrid IDS architectures. Benchmark datasets and evaluation methodologies were discussed, along with comparative performance trends across various IDS approaches.

Explainable hybrid adaptive intrusion detection systems represent a promising solution by combining domain knowledge, powerful learning models, transparency through XAI, and continuous model adaptation. These frameworks achieve superior detection accuracy while enhancing analyst trust and operational reliability.

However, challenges related to computational overhead, scalability, dataset imbalance, and adaptive learning control remain. Addressing these issues will be crucial for real-world deployment of next-generation intrusion detection systems.

Overall, explainable hybrid intelligence offers a balanced and effective paradigm for adaptive cyber defense in distributed environments, paving the way for secure and trustworthy network infrastructures.

## Acknowledgment

The authors sincerely thank Dr. Manoj Kumar Deshpande, Senior Director, and Dr. Piyush Choudhary, Head of the Department of Computer Science, for their guidance and support during this research.

## References

1. Sommer R, Paxson V. Outside the closed world: On using machine learning for network intrusion detection. Proc IEEE Symp Security Privacy. 2010:305–316. doi:10.1109/SP.2010.25.
2. Zhang Y, Chen R, Li J, Zhang X. Explainable artificial intelligence in cybersecurity: A survey. IEEE Security Privacy. 2021;19(5):72–83. doi:10.1109/MSEC.2021.3082954.
3. Denning DE. An intrusion-detection model. IEEE Trans Software Eng. 1987;SE-13(2):222–232. doi:10.1109/TSE.1987.232894.
4. Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence. IEEE Access. 2018;6:52138–52160. doi:10.1109/ACCESS.2018.2870052.
5. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017:4765–4774.
6. Guidotti R, *et al*. A survey of methods for explaining black box models. ACM Comput Surv. 2019;51(5):93. doi:10.1145/3236009.
7. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proc ACM SIGKDD. 2016:785–794. doi:10.1145/2939672.2939785.
8. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001;29(5):1189–1232. doi:10.1214/aos/1013203451.
9. Javaid A, Niyaz Q, Sun W, Alam M. A deep learning approach for network intrusion detection system. Proc IEEE Big Data Conf. 2016:1906–1913. doi:10.1109/BigData.2016.7840811.
10. Ditzler G, Roveri M, Alippi C, Polikar R. Learning in nonstationary environments: A survey. IEEE Comput Intell Mag. 2015;10(4):12–25. doi:10.1109/MCI.2015.2471196.
11. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–1284. doi:10.1109/TKDE.2008.239.
12. Sharafaldin I, Lashkari AH, Ghorbani AA. Toward generating a new intrusion detection dataset and intrusion traffic characterization. Proc ICISSP. 2018:108–116.
13. Moustafa N, Slay J. UNSW-NB15: A comprehensive dataset for network intrusion detection systems. Proc IEEE MILCOM. 2015:1–6. doi:10.1109/MILCOM.2015.7357531.
14. Chiba Z, Abghour N, Moussaid K, Rida M. Machine learning based intrusion detection system for cloud environments. Comput Security. 2019;83:153–165. doi:10.1016/j.cose.2019.02.009.
15. Lee W, Stolfo SJ. Data mining approaches for intrusion detection. Proc USENIX Security Symp. 1998:79–93.
16. Tavallaee M, Bagheri E, Lu W, Ghorbani AA. A detailed analysis of the KDD CUP 99 data set. Proc IEEE CISDA. 2009:1–6. doi:10.1109/CISDA.2009.5356528.
17. Lashkari AH, Draper-Gil G, Mamun MSI, Ghorbani AA. Characterization of CICIDS2017 dataset. Proc ICISSP. 2018:1–10.
18. Moustafa N, Slay J. UNSW-NB15: A comprehensive dataset for network intrusion detection systems. Proc IEEE MILCOM. 2015:1–6. doi:10.1109/MILCOM.2015.7357531.
19. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. Proc ICML. 2006:233–240. doi:10.1145/1143844.1143874.
20. Bishop CM. Pattern Recognition and Machine Learning. New York: Springer; 2006.
21. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–1284. doi:10.1109/TKDE.2008.239.
22. Mitchell R, Chen IR. A survey of intrusion detection in Internet of Things. IEEE Commun Surveys Tuts. 2014;16(4):1869–1891. doi:10.1109/SURV.2014.031914.00045.
23. Kasongo SM, Sun Y. A deep learning method with feature engineering for wireless intrusion detection. IEEE Access. 2020;8:135324–135334. doi:10.1109/ACCESS.2020.3010545.
24. Kim J, Kim J, Thu HLT, Kim H. Long short term memory recurrent neural network classifier for intrusion detection. Future Gener Comput Syst. 2017;66:103–111. doi:10.1016/j.future.2016.03.007.
25. Almseidin M, Alzubi M, Kovacs M, Alkasassbeh S. Evaluation of machine learning algorithms for intrusion detection system. Procedia Comput Sci. 2017;18:131–138. doi:10.1016/j.procs.2017.06.119.
26. Barreno M, Nelson B, Joseph AD, Tygar J. The security of machine learning. Mach Learn. 2010;81(2):121–148. doi:10.1007/s10994-010-5188-5.
27. Nguyen TT, Armitage G. A survey of techniques for

internet traffic classification. IEEE Commun Surveys Tuts. 2008;10(4):56–76. doi:10.1109/SURV.2008.080406.
28. Shone S, Ngoc TN, Phai VD, Shi Q. A deep learning approach to network intrusion detection. IEEE Trans Inf Forensics Security. 2018;13(4). doi:10.1109/TIFS.2018.2837640.
29. Japkowicz N. The class imbalance problem: Significance and strategies. Artif Intell Rev. 2002;6(5):429–449. doi:10.1023/A:1013669704061.
30. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. doi:10.1023/A:1010933404324.
31. Ring M, Wunderlich S, Grüdl D, Landes D, Hotho A. Flow-based network traffic analysis using machine learning. IEEE Commun Surveys Tuts. 2019;21(3):1–30. doi:10.1109/COMST.2019.2929565.
32. Bengio Y, Courville A, Vincent P. Representation learning: A review. IEEE Trans Pattern Anal Mach Intell. 2013;35(8):1798–1828. doi:10.1109/TPAMI.2013.50.
33. Scarfone K, Mell P. Guide to Intrusion Detection and Prevention Systems (IDPS). NIST Special Publication 800-94; 2007.
34. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015;349(6245):255–260. doi:10.1126/science.aaa8415.
35. Laskov P, Düssel P, Schäfer C, Müller KR. Learning intrusion detection: Supervised or unsupervised? Int Conf Image Anal Recognit. 2005:50–57.
36. Choudhary P, Shinde B, Yadav A, Kumayu A, Parmar A, Upadhyay A, Joshi A. Gesture driven gaming: A deep dive into computer vision-based hand gesture recognition. Int J Multidiscip Res Growth Eval. 2024;8(online).
37. Choudhary P, Kumar A, Raja A, Sharma A, Jain K. Comprehensive review and comparative analysis of yoga pose estimation techniques: Advancements in computer vision for enhanced yoga practice. ResearchGate. 2023. Available from: https://www.researchgate.net/publication/384966076
38. Choudhary P, Singh D, Patel S, Jain S, Soni S, Sharma V. Augmented reality books in pre-primary education: Fostering social and emotional learning for young learners. ResearchGate. 2022. Available from: https://www.researchgate.net/publication/384966357
39. Choudhary P, Dubey R, Jain P, Singh S, Lalwani S, Kaushal M. IntelliLearn: AI powered education hub. ResearchGate. 2024. Available from: https://www.researchgate.net/publication/384966416
40. Choudhary P, Singh M, Sankhere S, Vishwakarma S, Sanap S. AI-based criminal identification system: A literature review. SSRN. 2022. Available from: https://ssrn.com/abstract=4990380
41. Choudhary P, Shinde B, Yadav A. Potato disease classification: An attempt to detect the diseases in the early stages. SSRN. 2024. Available from: https://ssrn.com/abstract=5035562
42. Choudhary P, Tejas N, Patil K, Sahu J, Raghuwanshi R, Kapure N, Kamaal M. Design and development of smart virtual assistant using latest tools and technologies. SSRN. 2022. Available from: https://ssrn.com/abstract=4990442
43. Choudhary P. Use of mnemonic methodology for teaching technical content in object oriented programming. Int J Core Eng Manag. 2016.
44. Singh C, Chauhan MD, Vatti RAVDP, Neeraja MB. ML based strategy for optimal power prediction in IoT. Solid State Technol. 2020;63(6):8138–8150.
45. Choudhary P. Innovations in automated food production: The case of sensory preservation in pickles. SSRN. 2024. Available from: https://ssrn.com/abstract=4993986
46. Choudhary P, Sharma K, Sharma K, Borasi M, Bhargava P, Ali I, Rehman K. Enhancing mentorship through technology: A comprehensive review of current practices and future directions. SSRN. 2024. Available from: https://ssrn.com/abstract=5070036
47. Choudhary P, Sameet N. AI-powered Android application for fruit and vegetable quality detection. Int J Multidiscip Res Growth Eval. 2024;5(06).

**How to Cite This Article**
Sendre J, Choudhary P. A comprehensive review of explainable and adaptive hybrid intrusion detection systems for distributed cyber defense. International Journal of Future Engineering Innovations. 2026;3(3):95-106.