Enhancing Cyber Threat Detection using Machine Learning and Behavioral Analytics

Ankur Sharma
Bourns, Inc. / IT & Security
Western Governors University, Millcreek, USA

ABSTRACT
Conventional security systems that depend on static rules and signature-based detection are increasingly challenged by the sophistication of modern cyber threats. This study examines how machine learning and behavioral analytics can be combined to strengthen cyber threat detection in contemporary digital environments. A two-phase design is adopted: an analytical review of current machine learning algorithms, behavioral monitoring methods, and adaptive cybersecurity frameworks, followed by an empirical evaluation on the NSL-KDD benchmark using its official training and testing split. Five detection approaches are evaluated: Decision Tree, Random Forest, support vector machine (SVM), a multilayer perceptron neural network, and an Isolation Forest trained exclusively on normal traffic to emulate behavioral baseline profiling. The Decision Tree attains the highest test accuracy (81.4%, F1 = 0.811), while the Random Forest achieves the strongest ranking ability (AUC = 0.960); the behavioral Isolation Forest reaches 79.6% accuracy and an AUC of 0.946 without observing a single labeled attack, together with the lowest false-positive rate (2.35%). A feature-group ablation shows that behavioral traffic features raise the unsupervised detector's F1-score from 0.311 to 0.793 and that behavioral features alone support an AUC of approximately 0.90 across all model families. Detection of rare attack classes (R2L, U2R) remains weak across every model, quantifying the impact of class imbalance. Based on the combined analytical and experimental evidence, an integrated multi-layered framework is proposed that couples behavioral profiling with adaptive machine learning classification to support proactive, scalable cyber defense in enterprise and cloud environments.

General Terms
Security, Algorithms, Experimentation, Performance, Measurement.

Keywords
Cyber threat detection, machine learning, behavioral analytics, anomaly detection, intrusion detection systems, NSL-KDD, adaptive security, user and entity behavior analytics.

1. INTRODUCTION
Cybersecurity threats have grown more complex as enterprise networks have become larger and more interconnected, driven by the proliferation of digital infrastructure, cloud services, and the Internet of Things (IoT). Attacks are now more adaptive and persistent, extending well beyond traditional malware and phishing, and organizations face elevated risk as adversaries exploit weaknesses in both security architectures and human users [1]. Rule-based and signature-based intrusion detection systems remain poorly suited to zero-day attacks, insider threats, polymorphic malware, and advanced persistent threats (APTs) because they depend on fixed rules and previously observed attack patterns.

Recent advances in artificial intelligence (AI) and machine learning have opened new possibilities for intelligent threat detection. Machine learning methods can detect anomalies, classify malicious traffic, and analyze large volumes of network data with greater efficiency and scalability than conventional systems [2]. As AI-powered security systems become embedded in cloud and enterprise settings, they have shown potential to improve adaptive defense, automated monitoring, and real-time threat mitigation [3], and AI-enabled big data analytics has emerged as a crucial element of cyber resilience and proactive detection [4]. The same data-driven pattern recognition has demonstrated value across domains as diverse as agricultural yield prediction [5], industrial predictive maintenance [6], educational outcome modeling [7], and gamified learning systems [8], underscoring the generality of the analytical machinery that cybersecurity applications now leverage.

Behavioral analytics has likewise become increasingly significant in cybersecurity research for its ability to track user and system behavior and to flag deviations from established operational baselines. In contrast to static detection, behavioral analytics improves the identification of insider threats, unauthorized access attempts, and stealthy attacks, and - when paired with intelligent machine learning systems - has demonstrated strong potential to raise detection accuracy while reducing the false-positive alarms commonly reported for traditional systems [4, 9]. This capability is especially applicable to cloud-based systems and zero-trust networks, where continuous monitoring and intelligent authentication protect distributed digital environments [10].

Nevertheless, operational and technical issues hinder real-world deployment. A central challenge is the imbalance of cybersecurity datasets, in which malicious traffic constitutes a small fraction of observations; imbalance degrades model training, classification performance, and generalizability [11]. Feature scaling, preprocessing, and model optimization are similarly consequential for detection reliability [12], and the selection of appropriate evaluation metrics with sound statistical validation is critical for trustworthy machine learning in security contexts [13].

The growing complexity of AI-based security systems has also renewed interest in Machine Learning Operations (MLOps) for continuous deployment, monitoring, scaling, and lifecycle management of models in cybersecurity infrastructures [14]. Recent work further highlights AI-powered monitoring, intelligent security analytics, and automated forensic analysis as key enablers of modern cyber defense [15]. Collectively, these developments mark a movement away from reactive security toward intelligent, adaptive, behavior-led architectures supporting real-time detection and autonomous mitigation.

Although prior research has examined machine learning for cybersecurity, comparatively little work integrates behavioral analytics with adaptive machine learning in a unified detection

architecture, and much of the existing literature evaluates classification accuracy alone without quantifying how behavioral features contribute to generalization against previously unseen attacks. This study addresses that gap through three contributions. First, it synthesizes the current literature on machine learning, behavioral analytics, and AI-driven defense to identify recurring strengths, limitations, and research gaps. Second, it empirically evaluates five representative detection approaches - including an unsupervised behavioral baseline trained only on normal traffic - on the NSL-KDD benchmark's official cross-distribution split, and quantifies the contribution of behavioral feature groups through a controlled ablation. Third, it consolidates the analytical and experimental findings into an integrated, multi-layered framework for adaptive, scalable, real-time cyber threat detection.

2. LITERATURE REVIEW

2.1 Cyber Threat Detection and Response

With growing reliance on digital infrastructure, cloud-based systems, and interconnected enterprise environments, the frequency and sophistication of cyber threats against governments, organizations, and individuals continue to increase. Cybersecurity research has consequently concentrated on protecting information systems from unauthorized access, data breaches, malware, ransomware, insider threats, and APTs. Traditional security systems rely on signature-based detection, which identifies malicious activity by matching it to previously defined signatures; such tactics remain effective against known patterns but increasingly struggle with novel attack types and variants such as polymorphic malware and zero-day exploits [1].

As intrusions grow in sophistication, anomaly-based detection systems have emerged that alert on outlier activity rather than signature matches, reflecting a broader shift from reactive to proactive architectures with adaptive, autonomous defenses [16]. Anomaly-based systems, however, are prone to high false-positive rates because genuine variation in user behavior is difficult to distinguish from malicious activity. In parallel, machine learning has enabled security systems to classify threats, predict attacks, and adapt through intelligent pattern recognition over large volumes of security data in real time [4], which is increasingly necessary as enterprise traffic outgrows the capacity of conventional monitoring infrastructures.

2.2 Machine Learning in Cybersecurity

Machine learning is among the most impactful technologies in contemporary security research, improving detection accuracy, automating analysis, and uncovering new attack patterns. Depending on the training process and data availability, techniques are broadly categorized as supervised, unsupervised, or reinforcement learning; supervised algorithms learn from labeled data to separate malicious from benign activity, while unsupervised models expose hidden structure and anomalies in unlabeled data [2]. Typical methods include Decision Trees, Random Forests, support vector machines (SVMs), artificial neural networks, and deep learning architectures, applied to intrusion detection, malware classification, phishing detection, fraud analysis, and network traffic monitoring [3]. Deep architectures additionally enable automated feature extraction and recognition of advanced attack patterns in high-dimensional data.

Several obstacles nevertheless persist. Cybersecurity datasets are typically imbalanced - malicious traffic often represents a very small fraction of observations - which encourages classifiers to overfit the majority class, weakens detection of rare attacks, and increases false negatives [11]. This concern is acute in enterprise networks where rare but harmful activity may be missed for lack of high-quality training data. Feature engineering and preprocessing likewise exert a strong influence: scaling, normalization, and dimensionality reduction measurably improve model efficiency and classification reliability across tasks [12]. Finally, evaluation must be statistically sound; in imbalanced security settings, raw accuracy can mislead, and metrics such as precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC) are recommended for reliable, reproducible assessment [13].

2.3 Behavioral Analytics for Threat Detection

By tracking and analyzing user, device, and network patterns, behavioral analytics has become a vital component of modern security strategies. Whereas signature-driven systems match known indicators, behavioral analytics detects departures from an established behavioral baseline, improving identification of insider threats, credential misuse, account compromise, and stealthy attacks [9]. In enterprise settings, User and Entity Behavior Analytics (UEBA) systems continuously monitor login frequency, device interactions, access behavior, and communication patterns to expose operational anomalies, and their combination with machine learning supports adaptive learning and continuously updated threat classification [4].

Behavioral analytics is particularly consequential in cloud security and zero-trust architectures, where continuous authentication and verification are essential; zero-trust strategies assume threats may originate inside or outside the perimeter, so user behavior and network activity are monitored continuously, and AI-based behavioral systems have shown strong potential for autonomous mitigation and adaptive defense in cloud infrastructures [10]. Remaining challenges include confidentiality and data governance, the computational cost of processing high-volume behavioral data, and baseline instability in highly dynamic environments where frequent user and workload changes erode model validity over time.

2.4 AI-Driven Cyber Defense Systems

AI has shifted cybersecurity from a reactive posture toward predictive, autonomous threat intelligence. AI-based architectures integrate machine learning, behavioral analysis, automation, and intelligent monitoring to strengthen organizational resilience, with recent studies emphasizing automated response, adaptive identification, and proactive defense [15]. Deployments increasingly combine edge computing, cloud-based analytics, and threat intelligence to improve real-time monitoring across distributed environments, reducing latency and improving scalability in large enterprise infrastructures [17]. Adaptive monitoring services built on cloud-native tooling further demonstrate how continuous telemetry can feed intelligent detection pipelines in production environments [18], and AI-driven fraud detection illustrates the same adaptive-learning principles applied to financial cloud platforms [19].

Operationally, the rise of MLOps has strengthened the deployment and management of security-focused machine learning, supporting continuous monitoring, automated deployment, scaling, and lifecycle management in complex enterprise scenarios [14]. Related MLHOps frameworks emphasize operational dependability, governance, and sustainability of deployed models [20] - considerations that apply directly to the durability of AI-based cyber defense.

2.5 Challenges and Research Gaps

Despite this progress, notable gaps remain. Much of the literature optimizes classification accuracy on static datasets while paying limited attention to scalability, adaptability, explainability, and real-time production deployment, and static benchmarks do not fully reflect evolving real-world threats. The integration of behavioral analytics with machine learning - balancing detection sensitivity against false-positive rates within a single adaptive architecture - remains underdeveloped, and few studies quantify the marginal contribution of behavioral features to detection performance. Finally, the growth of AI-powered defense raises questions of ethical governance, adversarial machine learning, model transparency, and privacy-preserving analytics. These gaps motivate the present study's combination of literature synthesis, controlled experimentation on a public benchmark, and framework development for scalable, adaptive, real-time detection in heterogeneous environments.

3. METHODOLOGY

3.1 Research Design

A two-phase, mixed analytical-empirical design was adopted. The first phase is a comparative analytical review of recent peer-reviewed research on machine learning, behavioral analytics, and adaptive cyber defense, used to identify recurring strengths, limitations, and design principles. The second phase is a controlled experimental evaluation on a public intrusion-detection benchmark, designed to test three questions that the literature raises but seldom quantifies: (1) how well representative supervised models generalize to attack types absent from training data; (2) whether an unsupervised detector trained only on normal behavior - an operational analogue of behavioral baseline profiling - can approach supervised performance; and (3) how much behavioral feature groups contribute to detection performance relative to basic connection features. The two phases jointly inform the integrated framework proposed in Section 3.3.

3.2 Literature Selection

The analytical phase drew exclusively on peer-reviewed journal and conference publications, prioritizing work published between 2023 and 2025 to reflect the current threat and technology landscape. Studies were included if they (1) addressed cybersecurity applications of machine learning directly; (2) contributed analytical, conceptual, or empirical findings on threat detection; (3) examined operational challenges, intelligent security architectures, or adaptive defense; and (4) appeared in peer-reviewed venues with verifiable digital object identifiers. Non-peer-reviewed publications, duplicates, and methodologically unrelated articles were excluded. Thematic areas covered intrusion detection, behavioral anomaly detection, AI-augmented defense, cloud security architectures, security analytics, and adaptive threat mitigation.

3.3 Proposed Integrated Framework

Synthesis of the reviewed literature informed the design of an integrated detection architecture consisting of five operational layers: (1) continuous data acquisition and monitoring; (2) behavioral profiling and anomaly detection; (3) machine learning classification; (4) threat intelligence correlation and decision support; and (5) automated response and mitigation, closed by an adaptive feedback loop that retrains behavioral baselines and classification models as the environment evolves. Figure 1 presents the layered architecture, and Figure 2 shows the corresponding operational workflow from raw telemetry to automated response. The experimental design in Sections 3.4–3.6 instantiates the framework's two analytical layers - behavioral profiling and machine learning classification - so that their individual and combined contributions can be measured.

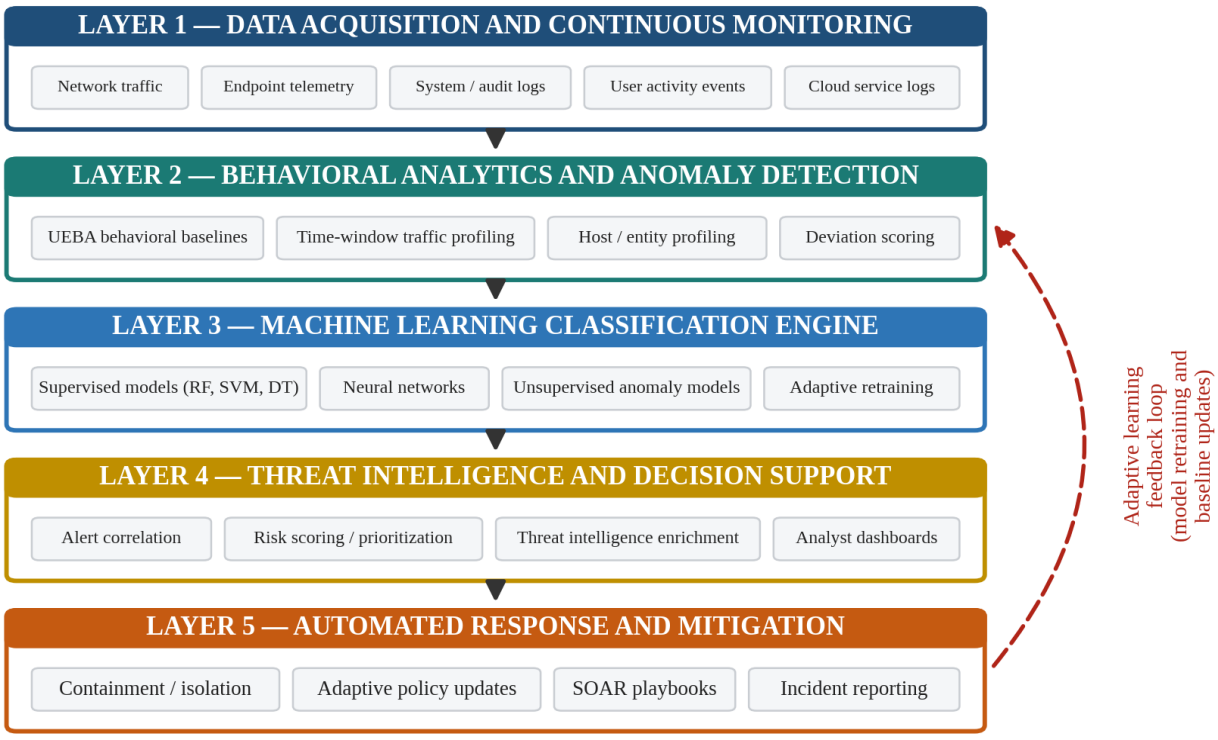

Figure 1: Integrated machine learning and behavioral analytics cybersecurity framework

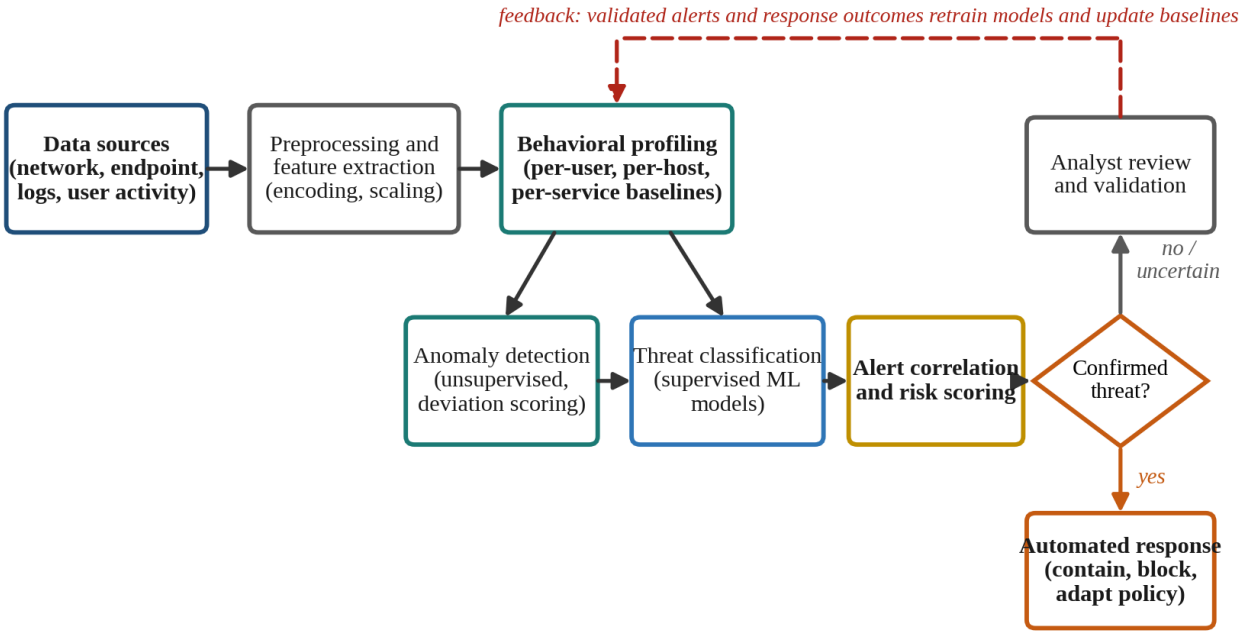

Figure 2: Operational workflow of AI-driven cyber threat detection and adaptive response

3.4 Experimental Dataset and Preprocessing

Experiments used NSL-KDD [21], a refined version of the KDD Cup 1999 corpus that removes duplicate records and remains a standard public benchmark for network intrusion detection. The official split was preserved: 125,973 training records (KDDTrain+) and 22,544 testing records (KDDTest+). Each record contains 41 features spanning four groups: basic connection features (e.g., duration, protocol, service, bytes transferred), content features describing in-session behavior (e.g., failed logins, root shell access, file creations), time-based traffic features computed over a two-second window (e.g., connection counts, same-service rates, error rates), and host-based traffic features computed over a 100-connection window to the same destination. The content, time-based, and host-based groups collectively encode behavioral characteristics of users, sessions, and hosts, and are treated as the behavioral feature set in the ablation of Section 4.4.

Labels were binarized (normal vs. attack), and attack labels were additionally mapped to the four standard categories - denial of service (DoS), probing (Probe), remote-to-local (R2L), and user-to-root (U2R) - for per-category analysis. Table 1 summarizes the composition. Two properties make this benchmark demanding and realistic for the study's questions. First, the class prior shifts between splits (46.5% attacks in training vs. 56.9% in testing). Second, and more importantly, the test set contains 17 attack labels never seen during training (e.g., mscan, snmpguess, httptunnel, apache2), accounting for 3,750 records or 29.2% of all test attacks; performance on KDDTest+ therefore directly measures generalization to previously unseen threats. Categorical features (protocol type, service, flag) were one-hot encoded and numeric features standardized to zero mean and unit variance, with all encoders and scalers fitted on training data only, yielding 122 model inputs. Consistent preprocessing of this kind is a documented prerequisite for reliable classification performance [12].

Table 1. Composition of the NSL-KDD official training and testing split

Category	Train records	Train %	Test records	Test %
Normal	67,343	53.46	9,711	43.08
DoS	45,927	36.46	7,460	33.09
Probe	11,656	9.25	2,421	10.74
R2L	995	0.79	2,885	12.80
U2R	52	0.04	67	0.30
Total	125,973	100.0	22,544	100.0

3.5 Models and Experimental Setup

Five detection approaches were evaluated, chosen to represent the model families most frequently reported in the reviewed literature. Four are supervised classifiers trained on the full labeled training set: a Decision Tree (unlimited depth), a Random Forest of 100 trees, an SVM with radial basis function kernel trained on a stratified subsample of 20,000 records to keep kernel computation tractable (a standard practice reflecting the scalability limits of kernel methods on large corpora), and a multilayer perceptron (MLP) neural network with hidden layers of 128 and 64 units, trained with early stopping. The fifth approach operationalizes behavioral baseline profiling: an Isolation Forest of 100 trees trained exclusively on the 67,343 normal training records, with no exposure to labeled attacks, flagging test records as anomalous when they deviate from the learned normal profile. All models were implemented in scikit-learn 1.8 [22] with a fixed random seed (42) for reproducibility, and all training and inference ran on a single commodity workstation; wall-clock training times are reported alongside detection metrics to reflect operational cost.

To quantify the contribution of behavioral information (Section 4.4), the Random Forest analysis was complemented by a feature-group ablation in which every model family was retrained under three conditions: connection features only

(basic plus content, 22 features), behavioral traffic features only (time-based plus host-based, 19 features), and all 41 features. Feature importances from the full-feature Random Forest were aggregated across one-hot categories to attribute importance to the original 41 features.

3.6 Evaluation Metrics

Given the class imbalance and distribution shift documented above, evaluation followed the recommendations of [13]: accuracy is reported for comparability but interpretation emphasizes precision, recall, F1-score, and threshold-independent AUC, together with the false-positive rate (FPR), which governs analyst alert load in operational settings, and the false-negative rate (FNR), which measures missed attacks. Confusion matrices and per-attack-category detection rates are reported to expose error structure that aggregate metrics conceal.

3.7 Ethical Considerations

The study used only openly available scholarly literature and a public benchmark dataset; no human subjects, confidential organizational data, or live attack traffic were involved, so formal institutional ethical approval was not required. Ethical research practice was maintained throughout, including accurate attribution and restraint from unwarranted security claims. The broader ethical challenges of AI-driven defense - privacy, algorithmic bias, behavioral surveillance, and explainability - are examined in Section 5.3 and Section 5.4.

4. RESULTS AND ANALYSIS

4.1 Comparative Insights from the Literature

The analytical synthesis indicates clear and complementary strengths across model families. Supervised techniques - particularly Random Forest, SVM, and Decision Tree - are repeatedly reported to perform strongly on intrusion detection, malware classification, and phishing identification; Random Forests are valued for robustness on high-dimensional, noisy security data, while SVMs perform well on structured datasets but raise scalability concerns in high-volume, real-time analysis. Neural and deep architectures excel at complex, nonlinear attack patterns such as APTs and polymorphic malware, at the cost of computational demand, large training sets, and limited explainability. Unsupervised methods are consistently identified as valuable for exposing previously unknown attacks without signatures, tempered by higher false-positive rates in dynamic environments. Across the reviewed studies, hybrid systems that combine supervised classification with behavioral analytics are reported to outperform standalone models in adaptability, contextual interpretation, and anomaly identification. Table 2 consolidates these findings, which motivate the model selection and the behavioral ablation evaluated next.

Table 2. Comparative analysis of machine learning techniques for cyber threat detection (literature synthesis)

Technique	Major strength	Limitation	Typical application
Random Forest	High classification accuracy; robust to noise	Computational complexity	Intrusion detection
SVM	Effective on structured datasets	Limited scalability	Malware detection
Decision Tree	Interpretability and simplicity	Overfitting risk	Threat classification
Neural networks	Strong pattern recognition	High training requirements	**ehavioral anomaly detection**
Deep learning	Detection of complex threats	Explainability challenges	APT detection
Unsupervised learning	Detection of unknown threats	Higher false-positive rates	Anomaly detection

4.2 Overall Detection Performance on NSL-KDD

Table 3 reports the performance of all five models on the official KDDTest+ split, and Figure 3 visualizes the four headline classification metrics. Three results stand out. First, every model attains high precision (0.960–0.974) at low false-positive rates (2.35%–3.89%): alerts raised by any of the models are overwhelmingly genuine, an operationally important property because false alarms drive analyst alert fatigue. Second, recall is markedly lower (0.629–0.703) for all models, so aggregate accuracy settles in the 77.4%–81.4% range. This asymmetry is not incidental: 29.2% of test attacks belong to the 17 attack types absent from training, so a substantial share of malicious records resembles nothing the supervised models have seen. The precision–recall gap therefore quantifies, on a controlled benchmark, the generalization limitation that the literature attributes to signature-style supervised learning. Third, the simplest model, the Decision Tree, achieves the best test accuracy (81.4%) and F1-score (0.811), outperforming the Random Forest (77.9%, 0.765) and SVM (77.4%, 0.760) under distribution shift, while the MLP is second (80.7%, 0.801).

Table 3. Detection performance on the official NSL-KDD test set (KDDTest+); best value per column in bold

Model	Accuracy	Precision	Recall	F1	AUC	FPR	FNR	Train (s)
Decision Tree	**0.8141**	0.9598	**0.7029**	**0.8115**	0.8322	0.0389	**0.2971**	2.7
Random Forest	0.7787	0.9689	0.6314	0.7646	**0.9598**	0.0268	0.3686	14.4
SVM (RBF, 20k)	0.7740	0.9598	0.6294	0.7603	0.8882	0.0348	0.3706	3.4
MLP neural network	0.8068	0.9678	0.6833	0.8010	0.9026	0.0301	0.3167	37.2
Isolation Forest (behavioral)	0.7962	**0.9738**	0.6597	0.7865	0.9463	**0.0235**	0.3403	0.2

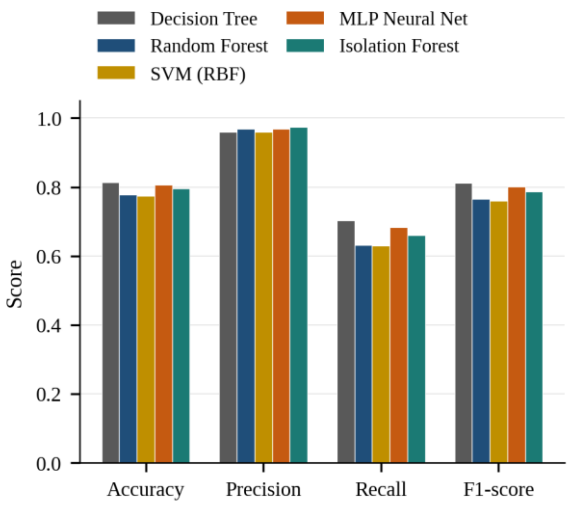

Figure 3: Classification performance of the five models on KDDTest+

The behavioral baseline result deserves emphasis. The Isolation Forest never observed a labeled attack, yet it reaches 79.6% accuracy and an F1-score of 0.787 - higher than the fully supervised Random Forest and SVM - with the lowest false-positive rate of any model (2.35%) and the highest precision (0.974). It is also the cheapest model to train by two orders of magnitude (0.2 s versus 2.7–37.2 s), because it fits only normal traffic. This directly supports the framework's premise: a behavioral profile of normality, maintained continuously and inexpensively, is a competitive first line of detection precisely in the regime where labeled attack data are scarce or stale. The training-time column also confirms the literature's scalability observations - the kernel SVM required subsampling to remain tractable, whereas the tree-based methods scaled to the full corpus without difficulty.

4.3 Discrimination Ability and Error Analysis

Figure 4 presents ROC curves. Ranked by threshold-independent discrimination, the Random Forest is strongest (AUC = 0.960), followed by the behavioral Isolation Forest (0.946), the MLP (0.903), the SVM (0.888), and the Decision Tree (0.832). The contrast between Table 3 and this ranking is analytically informative: the Random Forest orders attack and normal records almost correctly (96% probability that a random attack scores above a random normal record) yet loses accuracy at the default 0.5 decision threshold, where its ensemble-averaged probabilities for unfamiliar attacks fall short of the cutoff - its 4,730 false negatives (Figure 5) are concentrated exactly there. The Decision Tree shows the opposite profile: hard, axis-aligned splits generalize surprisingly well at the fixed threshold (3,813 false negatives, the fewest) but provide poor ranking (AUC = 0.832). Operationally, this implies the Random Forest and Isolation Forest are the preferable alarm-scoring engines provided the alert threshold is calibrated to the deployment environment - for example against a validation stream or an FPR budget - rather than left at the default, a tuning task that the proposed framework's adaptive feedback layer is designed to perform continuously.

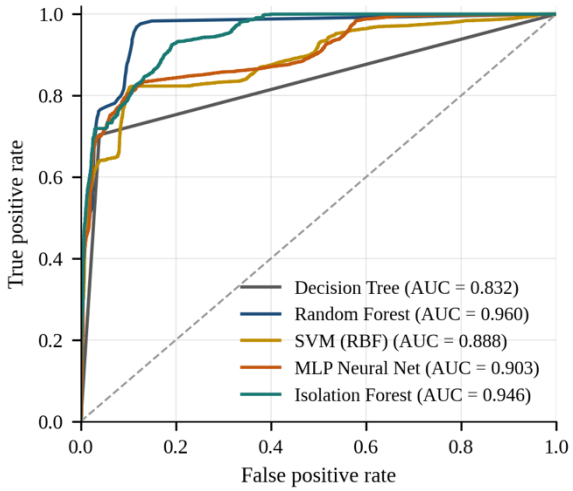

Figure 4: ROC curves on KDDTest+ (higher area indicates better ranking of attacks above normal traffic)

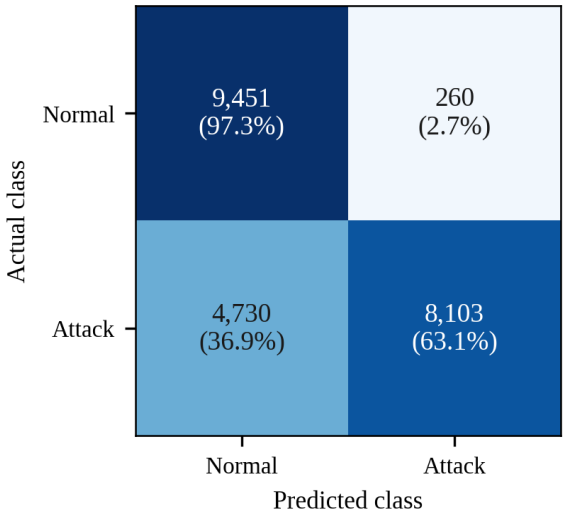

Figure 5: Confusion matrices on KDDTest+ (counts, with row-normalized percentages)

Table 4 decomposes recall by attack category and exposes structure that aggregate metrics hide. High-volume categories are detected reliably: DoS detection ranges from 79.3% to 86.0% and Probe from 72.2% to 93.7%. The behavioral Isolation Forest is the single best Probe detector (93.7%), consistent with the behavioral signature of scanning - many connections, high error rates, and unusual service diversity - standing far outside the normal baseline. Rare categories collapse for every model: R2L detection spans only 5.9%–24.8% and U2R 13.4%–44.8%. The cause is visible in Table 1: R2L and U2R constitute 0.79% and 0.04% of training data respectively, while R2L additionally grows to 12.8% of the test set, so supervised learners have almost no signal to fit and the anomaly detector must separate low-and-slow application-layer misuse from ordinary user variation. These figures give concrete, per-category form to the imbalance problem described analytically in [11], and identify where targeted resampling, cost-sensitive learning, or category-specific behavioral features would yield the greatest marginal benefit.

Table 4. Per-category detection rate on KDDTest+ (for Normal, the value is the correct-classification rate); best value per row in bold

Category	Decision Tree	Random Forest	SVM (RBF)	MLP	Isolation Forest
Normal	0.961	0.973	0.965	0.970	**0.977**
DoS	0.846	0.828	0.793	**0.860**	0.795
Probe	0.815	0.722	0.781	0.825	**0.937**
R2L	**0.248**	0.059	0.091	0.113	0.086
U2R	0.313	0.134	0.134	**0.448**	0.313

4.4 Contribution of Behavioral Features

The ablation isolates what behavioral information adds. Figure 6 and Table 5 report F1-scores when each model family is trained on connection features only (basic + content, 22 features), behavioral traffic features only (time-based + host-based, 19 features), and all 41 features. For the supervised models the connection group already carries most of the signal at the default threshold (F1 0.787–0.809), and behavioral features alone still support respectable supervised performance (F1 0.703–0.739). The decisive effect appears in the unsupervised detector: with connection features only, the Isolation Forest is barely usable (F1 = 0.311), but with behavioral traffic features alone it reaches F1 = 0.793 - its best condition, slightly above the all-feature configuration (0.787). Behavioral aggregates are thus not merely helpful but constitutive for baseline-deviation detection: windows of connection counts, error rates, and service diversity define what "normal" looks like in a way that per-connection byte counts do not.

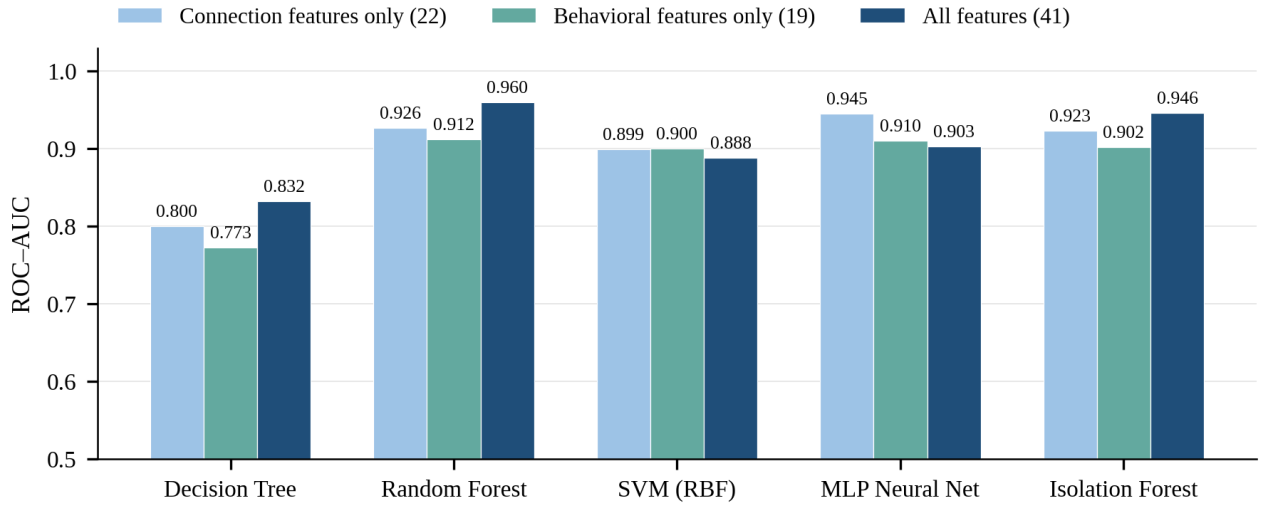

Figure 6: Feature-group ablation - F1-score by feature condition and model family

Table 5. Feature-group ablation results (F1-score / AUC on KDDTest+)

Model	Connection only	ehavioral only	All features
Decision Tree	0.787 / 0.800	0.728 / 0.773	0.811 / 0.832
Random Forest	0.800 / 0.927	0.703 / 0.912	0.765 / 0.960
SVM (RBF)	0.787 / 0.899	0.728 / 0.900	0.760 / 0.888
MLP neural network	0.809 / 0.945	0.739 / 0.910	0.801 / 0.903
Isolation Forest	0.311 / 0.923	0.793 / 0.902	0.787 / 0.946

Two further observations refine the picture. First, behavioral-only AUC is approximately 0.90 for every model family (0.900–0.912, and 0.773 only for the threshold-bound Decision Tree), meaning nineteen windowed behavioral aggregates alone rank attacks above normal traffic nearly as well as the full 122-dimensional encoding - strong quantitative support for behavioral analytics as a detection substrate rather than a supplement. Second, for the Random Forest, adding all features to the connection set slightly reduces F1 at the default threshold (0.800 to 0.765) even as AUC improves (0.927 to 0.960): richer inputs let the ensemble bind more tightly to training-distribution signatures, sharpening ranking while shifting scores for unseen attacks below the fixed cutoff. This is a concrete instance of why static training pipelines degrade under threat drift and why the proposed framework treats threshold calibration and periodic retraining as first-class operations rather than maintenance afterthoughts.

Figure 7 completes the analysis with aggregated Random Forest feature importances. The four most informative features are basic connection attributes (source bytes, TCP flag, destination bytes, service), but six of the top twelve are behavioral traffic aggregates - including the two-second connection count, same-service rate, and destination-host service statistics - confirming at the feature level that behavioral and connection information are complementary: connection features identify what a flow is, while behavioral features establish whether its context is normal.

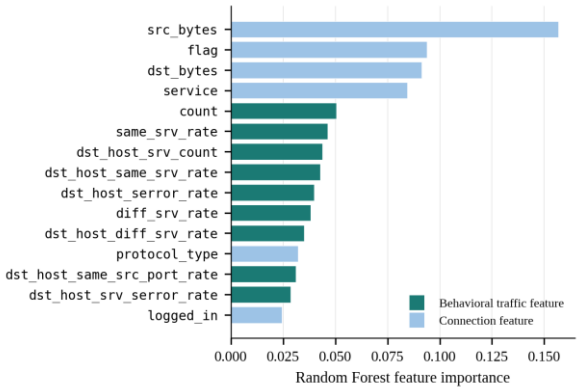

Figure 7: Top aggregated Random Forest feature importances, colored by feature group

4.5 Behavioral Analytics and Anomaly Detection Findings

Read together, the literature synthesis and the experiments converge on a consistent account of behavioral analytics. The reviewed studies report that behavioral monitoring exposes insider misuse, credential abuse, and stealthy attacks that signature matching misses, and that UEBA-style profiling of logins, access frequency, and communication patterns deepens incident visibility [4, 9]. The experiments make this concrete: a detector that models only normal behavior detects two-thirds of attacks (including 93.7% of probing activity) at the lowest false-positive rate of any evaluated model, and behavioral feature windows alone sustain an AUC of about 0.90 across model families. The literature's caveats are equally visible in the data: behavioral baselines drift as environments change, and the R2L results (Table 4) show that low-and-slow application-layer misuse can hide inside ordinary behavioral variation - the false-positive/sensitivity tension that continuous retraining, contextual enrichment, and governance-aware tuning must manage in production [10].

4.6 Operational Challenges Identified

The combined evidence identifies four operational challenges for intelligent detection systems. First, class imbalance: rare categories are precisely the ones models miss (Table 4), and the training prior (0.04%–0.79% for U2R/R2L) leaves little for supervised learners to fit [11]. Second, distribution shift: 29.2% of test attacks are types never seen in training, producing the uniform precision–recall asymmetry in Table 3; static models quantifiably decay against novel threats. Third, threshold and calibration management: the divergence between accuracy and AUC rankings (Sections 4.2–4.3) shows that deployment quality depends as much on decision-threshold governance as on model choice. Fourth, scalability and cost: kernel methods required subsampling while tree ensembles and the behavioral detector scaled easily, and training cost varied by two orders of magnitude - differences that compound in high-throughput enterprise telemetry and motivate the MLOps-style lifecycle management emphasized in the literature [14, 20].

5. DISCUSSION

5.1 Interpretation of Findings

The results substantiate, with controlled measurements, the central claims that motivated this study. Signature-style supervised learning is genuinely strong where its assumptions hold - precision approached 0.97 and false-positive rates stayed below 4% for every model - but its recall degrades predictably when nearly a third of attacks fall outside the training distribution. Behavioral modeling changes the failure mode rather than merely the score: the Isolation Forest, informed only by what normal traffic looks like, matched or exceeded two fully supervised models and dominated the detection of scanning behavior, while the ablation showed behavioral feature windows to be the substrate that makes unsupervised detection viable at all (F1 0.311 versus 0.793). Neither paradigm suffices alone: supervised models miss the unknown, and behavioral models struggle with misuse that imitates normality (R2L). The empirical case for the integrated framework is therefore not that hybrid systems are conventional wisdom in the literature, but that the two approaches fail in measurably complementary ways.

5.2 Machine Learning and the Evolution of Intelligent Cybersecurity

The findings also refine how machine learning should be positioned within security operations. Model selection proved less decisive than lifecycle management: the accuracy ranking and the AUC ranking disagreed, the best threshold-fixed model (Decision Tree) was the worst ranker, and the best ranker (Random Forest) lost accuracy to an uncalibrated cutoff under drift. In production terms, detection quality is a property of the pipeline - data preparation [12], metric selection [13], threshold calibration, monitoring, and retraining [14] - rather than of an algorithm in isolation. Deep architectures extend detection to complex nonlinear patterns [3], but the interpretability constraints emphasized in the literature remain: in regulated environments, organizations increasingly require security systems that explain classifications and automated responses, so future intelligent defense must pair predictive performance with explainable AI to preserve accountability and trust.

5.3 Behavioral Analytics and Adaptive Threat Intelligence

The behavioral results carry direct architectural implications. Because a normality profile is cheap to train (0.2 s here), label-free, and effective, it can be refreshed continuously at the network edge and per entity - exactly the continuous-verification posture that zero-trust architectures prescribe [10] and that edge-integrated defense systems operationalize [17]. Behavioral scoring also contextualizes supervised alerts: an event flagged by both a classifier and a baseline deviation warrants immediate priority, whereas classifier-only alerts on behaviorally normal activity are candidates for automated triage - a mechanism that addresses the alert-fatigue problem identified across the reviewed studies. The governance obligations are equally clear: continuous behavioral monitoring processes sensitive user activity, so privacy protection, data governance, and proportionate retention must be designed into the profiling layer rather than appended afterward, consistent with the operational-governance emphasis of MLHOps frameworks [20].

5.4 Operational Challenges and Deployment Limitations

Table 6 consolidates the principal operational challenges, their observed or reported impact, and mitigation strategies supported by the literature and the present measurements. Beyond the four challenges quantified in Section 4.6, two deployment risks from the literature warrant standing attention. Explainability: opaque, high-capacity models complicate incident investigation and regulatory compliance, arguing for interpretable models or post-hoc explanation in the classification layer. Adversarial machine learning: input perturbation, data poisoning, and model-evasion techniques

target the detection models themselves; robust training, input validation, and monitored model integrity are necessary countermeasures, and fraud-detection deployments in financial clouds illustrate both the value and the exposure of adaptive models in adversarial settings [19]. Continuous adaptive monitoring pipelines of the kind demonstrated on cloud-native tooling [18] provide the operational substrate on which these mitigations run.

Table 6. Operational challenges and mitigation strategies for AI-driven cybersecurity systems

Challenge	Impact (observed / reported)	Mitigation strategy
Imbalanced datasets	R2L/U2R detection collapses to 5.9–44.8% (Table 4)	Resampling, cost-sensitive learning, targeted features
Distribution shift / novel attacks	Recall 0.63–0.70 vs. precision 0.96–0.97 (Table 3)	**ehavioral baselining; periodic retraining**
Threshold miscalibration	Accuracy and AUC rankings diverge (Sections 4.2–4.3)	Continuous calibration against FPR budgets
False-positive alert load	FPR 2.35–3.89% at enterprise volume	**ehavioral context scoring; alert triage**
Scalability limits	Kernel SVM required subsampling; MLP costliest to train	Distributed/cloud processing; efficient ensembles
Explainability gaps	Opaque decisions impede investigation and compliance	Explainable AI in the classification layer
Adversarial manipulation	Poisoning and evasion target detection models	Robust training; input validation; model monitoring
Privacy in behavioral monitoring	Continuous profiling of user activity	Privacy-preserving analytics; governance policies

5.5 Theoretical and Practical Contributions

Theoretically, the study contributes a quantified account of the complementarity between supervised classification and behavioral anomaly detection on a controlled cross-distribution benchmark, including the first-order effects of behavioral feature groups on both supervised and unsupervised performance. Practically, the results translate into deployable guidance: maintain an inexpensive, continuously retrained behavioral baseline as a first detection line; use a high-AUC classifier (ensemble methods performed best) as the scoring engine with explicitly governed thresholds; prioritize alerts confirmed by both layers; and direct imbalance remediation at the specific categories where detection measurably fails. The proposed five-layer framework packages these practices with threat-intelligence correlation and automated response, providing a reference architecture for enterprise and cloud deployments seeking real-time, adaptive defense [15, 17].

5.6 Limitations and Future Research Directions

Four limitations bound the findings. First, the empirical phase used a single public benchmark; NSL-KDD's provenance in simulated 1999 traffic [21], however refined, means absolute figures will differ on modern telemetry, even though the structural findings - precision–recall asymmetry under novel attacks, behavioral-feature primacy for anomaly detection, and rare-class collapse - are properties of the learning problem rather than of the corpus. Second, evaluation was binary at the top level; full multi-class detection and attack attribution were examined only through per-category recall. Third, models were evaluated offline on a fixed split, whereas production systems face streaming data, concept drift, and adversarial adaptation. Fourth, the framework's response and orchestration layers were designed but not operationally exercised. Future work should accordingly validate the framework on contemporary datasets (e.g., CICIDS-family corpora) and live enterprise settings, incorporate explainable AI and adversarially robust training in the classification layer, examine federated and privacy-preserving formulations of behavioral profiling, and evaluate continuous-learning MLOps pipelines under real drift [14, 20].

6. CONCLUSION

This study set out to determine how machine learning and behavioral analytics can jointly strengthen cyber threat detection in modern digital environments, combining a synthesis of recent literature with a controlled evaluation on the official NSL-KDD split. The empirical results give the integration argument quantitative substance. Supervised models detect known attack families with precision near 0.97 and false-positive rates under 4%, yet lose 30–37% of attacks when nearly a third of test threats fall outside the training distribution. A behavioral detector trained only on normal traffic recovers much of that ground - 79.6% accuracy, AUC 0.946, the lowest false-positive rate measured, and the strongest detection of probing activity - while the feature ablation shows behavioral traffic aggregates to be the foundation of anomaly detection, lifting unsupervised F1 from 0.311 to 0.793 and alone sustaining an AUC of roughly 0.90 across model families. Persistent weaknesses on rare R2L and U2R categories quantify the imbalance problem and mark the clearest target for future gains.

These findings support an integrated, multi-layered architecture in which continuously retrained behavioral baselines provide label-free first-line detection, calibrated high-AUC classifiers score and rank threats, threat-intelligence correlation adds context, and automated response closes the loop under explicit governance of thresholds, privacy, and model lifecycle. With empirical validation on contemporary telemetry, explainable and adversarially robust components, and privacy-preserving behavioral profiling, such architectures offer a practical path from reactive, signature-bound security toward adaptive, intelligent cyber defense at enterprise and cloud scale.

7. REFERENCES

[1] Kommera, H. K. R. 2024. Adaptive cybersecurity in the digital age: emerging threat vectors and next-generation defense strategies. International Journal for Research in

Applied Science and Engineering Technology 12, 9, 558–564. DOI: 10.22214/ijraset.2024.64226.

[2] Afifi, H., Pochaba, S., Boltres, A., Laniewski, D., Haberer, J., Paeleke, L., et al. 2024. Machine learning with computer networks: techniques, datasets, and models. IEEE Access 12, 54673–54720. DOI: 10.1109/ACCESS.2024.3384460.

[3] Vemulapalli, V. K. 2025. AI-driven cybersecurity: the future of adaptive threat defense. World Journal of Advanced Research and Reviews 26, 2, 3248–3255. DOI: 10.30574/wjarr.2025.26.2.1953.

[4] Sultana, S., Uddin, M., Chy, M. A. R., Hasan, S. N., Hossain, E., Kaur, H., et al. 2025. AI-augmented big data analytics for real-time cyber attack detection and proactive threat mitigation. International Journal of Computational and Experimental Science and Engineering 11, 3. DOI: 10.22399/ijcesen.3564.

[5] Elbasi, E., Zaki, C., Topcu, A. E., Abdelbaki, W., Zreikat, A. I., Cina, E., et al. 2023. Crop prediction model using machine learning algorithms. Applied Sciences 13, 16, 9288. DOI: 10.3390/app13169288.

[6] Aminzadeh, A., Sattarpanah Karganroudi, S., Majidi, S., Dabompre, C., Azaiez, K., Mitride, C., and Sénéchal, E. 2025. A machine learning implementation to predictive maintenance and monitoring of industrial compressors. Sensors 25, 4, 1006. DOI: 10.3390/s25041006.

[7] Pelima, L. R., Sukmana, Y., and Rosmansyah, Y. 2024. Predicting university student graduation using academic performance and machine learning: a systematic literature review. IEEE Access 12, 23451–23465. DOI: 10.1109/ACCESS.2024.3361479.

[8] Swacha, J. and Gracel, M. 2023. Machine learning in gamification and gamification in machine learning: a systematic literature mapping. Applied Sciences 13, 20, 11427. DOI: 10.3390/app132011427.

[9] Hasan, S. N., Kaur, H., Mohonta, S. C., Siddiqa, K. B., Kaur, J., Haldar, U., et al. 2025. The influence of artificial intelligence on data system security. International Journal of Computational and Experimental Science and Engineering 11, 3. DOI: 10.22399/ijcesen.3476.

[10] Kolawole, I. 2025. Leveraging cloud-based AI and zero trust architecture to enhance U.S. cybersecurity and counteract foreign threats. World Journal of Advanced Research and Reviews 25, 3, 006–025. DOI: 10.30574/wjarr.2025.25.3.0635.

[11] Altalhan, M., Algarni, A., and Turki-Hadj Alouane, M. 2025. Imbalanced data problem in machine learning: a review. IEEE Access 13, 13686–13699. DOI: 10.1109/ACCESS.2025.3531662.

[12] Pinheiro, J. M. H., Oliveira, S. V. B. de, Silva, T. H. S., Saraiva, P. A. R., Souza, E. F. de, Godoy, R. V., et al. 2025. The impact of feature scaling in machine learning: effects on regression and classification tasks. IEEE Access 13, 199903–199931. DOI: 10.1109/ACCESS.2025.3635541.

[13] Rainio, O., Teuho, J., and Klén, R. 2024. Evaluation metrics and statistical tests for machine learning. Scientific Reports 14, 1, 6086. DOI: 10.1038/s41598-024-56706-x.

[14] Kreuzberger, D., Kühl, N., and Hirschl, S. 2023. Machine learning operations (MLOps): overview, definition, and architecture. IEEE Access 11, 31866–31879. DOI: 10.1109/ACCESS.2023.3262138.

[15] Reddy, K. V. S. P., Mohith, B., Babu, P. M., and Navtej, K. 2025. CyberSleuth AI: intelligent network forensics analyzer. International Journal for Research in Applied Science and Engineering Technology 13, 4, 1643–1647. DOI: 10.22214/ijraset.2025.68420.

[16] Chawande, S. 2024. The role of artificial intelligence in cybersecurity. World Journal of Advanced Engineering Technology and Sciences 11, 2, 683–696. DOI: 10.30574/wjaets.2024.11.2.0014.

[17] Ajayi, R. and Masunda, M. 2025. Integrating edge computing, data science and advanced cyber defense for autonomous threat mitigation. International Journal of Science and Research Archive 15, 2, 063–080. DOI: 10.30574/ijsra.2025.15.2.1292.

[18] Bansal, I. 2025. Implementing adaptive security monitoring: AWS CloudWatch and advanced threat detection techniques. International Journal of Computational and Experimental Science and Engineering 11, 3. DOI: 10.22399/ijcesen.3482.

[19] Palla, S. R. 2025. AI-driven fraud detection and security: a transformative approach for financial services cloud environments. World Journal of Advanced Research and Reviews 26, 1, 2040–2050. DOI: 10.30574/wjarr.2025.26.1.1296.

[20] Khattak, F. K., Subasri, V., Krishnan, A., Pou-Prom, C., Akinli-Kocak, S., Dolatabadi, E., et al. 2025. MLHOps: machine learning health operations. IEEE Access 13, 20374–20412. DOI: 10.1109/ACCESS.2024.3521279.

[21] Tavallaee, M., Bagheri, E., Lu, W., and Ghorbani, A. A. 2009. A detailed analysis of the KDD CUP 99 data set. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), IEEE, 1–6. DOI: 10.1109/CISDA.2009.5356528.

[22] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. 2011. Scikit-learn: machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.