UDC 37:004.056:004.8

**Artughrul Gayibov**
Baku Engineering University, Khirdalan, Azerbaijan
ORCID ID 0009-0009-7349-0286
*agayibov@beu.edu.az*

**Vagif Gasimov**
Baku Engineering University, Khirdalan, Azerbaijan
ORCID 0000-0003-3192-4225
*vaqasimov@beu.edu.az*

**Islam Islamov**
Baku Engineering University, Khirdalan, Azerbaijan
ORCID 0000-0001-8645-0640
*isislamov@beu.edu.az*

**Jabir Mammadov**
Baku Engineering University, Khirdalan, Azerbaijan
ORCID 0000-0003-3939-4708
*camammadov@beu.edu.az*

**Seyran Gayibov**
Baku Engineering University, Khirdalan, Azerbaijan
ORCID ID 0000-0001-7434-6248
*sqayibov@beu.edu.az*

# MACHINE LEARNING APPLICATION FOR INTELLIGENT THREAT CLASSIFICATION IN CLOUDFLARE BOT MANAGEMENT FIREWALL LOGS OF UNIVERSITY DIGITAL EDUCATION INFRASTRUCTURE

**Abstract.** The object of this study is the automated classification of Cloudflare Bot Management (CBM) firewall log records collected from a university web application. The study addresses the problem of distinguishing potentially malicious requests from legitimate academic traffic in the digital educational infrastructure of a higher education institution. Static threshold-based firewall rules are often insufficient for analyzing multidimensional security indicators, including WAF machine-learning attack scores, bot-related signals, TLS fingerprint information, and request-level metadata. A comparative machine-learning framework was developed using six supervised classifiers: Random Forest (RF), Gradient Boosting (GB), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gaussian Naive Bayes (GNB). The models were trained and evaluated on 24 engineered features extracted from 500 CBM log entries using a stratified 80/20 train-test split and 5-fold cross-validation. RF, GB, and DT achieved 100% test accuracy on the constructed dataset, while SVM, KNN, and GNB achieved 96.00%, 93.00%, and 44.00%, respectively. These results should be interpreted in the context of the proposed labelling procedure, because the target variable was derived from WAF score aggregation and the same group of WAF indicators was also used among the input features. Feature importance analysis showed that WAF machine-learning attack scores formed the dominant group of classification indicators. The Decision Tree model produced an interpretable threshold rule based on the normalized "wafMlAttackScore". This rule is not presented as a fully deployed production firewall policy, but as a candidate rule that can support further expert validation and configuration of university WAF security policies. The proposed approach therefore contributes to the interpretation and secondary classification of CBM firewall logs in educational ICT security management.

**Keywords:** university network security; ensemble machine learning; WAF threat detection; firewall log analysis; educational ICT infrastructure.

## 1. INTRODUCTION

© Artughrul Gayibov, Vagif Gasimov, Islam Islamov, Jabir Mammadov, Seyran Gayibov, 2026. 

The digital transformation in higher education has made university networks the cornerstone of academic activities. Learning management systems, student portals, platforms for remote exams, and repositories for research collaboration all rely on the constant availability and integrity of the institution's web services. Current industry data confirms that the education sector is among the industries most frequently targeted by web-based attacks, as universities process sensitive personal, financial, and research data while maintaining open-access environments designed for geographically dispersed user groups [1], [2]. Protecting this infrastructure, particularly the web-facing services exposed through university portals, learning platforms, and administrative systems, is a direct prerequisite for reliable ICT-supported education.

Web Application Firewalls (WAFs) represent the primary perimeter security layer for university web services. Modern cloud-based WAF platforms such as Cloudflare Bot Management (CBM) generate detailed security logs for every incoming request, enriched with machine learning-generated attack probability scores, bot confidence metrics, TLS fingerprint signals, and geographic metadata. These logs capture the full threat context of network activity targeting web services in the education sector. Translating this rich, high-dimensional data into reliable real-time threat classifications remains an unresolved issue directly related to educational ICT management [3].

Automated threat detection in firewall logs refers to the classification of each logged network event as either harmless traffic or a security threat based on a combination of security indicators. In the CBM context, these include WAF ML attack assessments in the categories of SQL injection, cross-site scripting (XSS), and remote code execution (RCE), as well as bot confidence signals and network metadata. Unlike signature-based detection, ML-based classification continuously learns decision boundaries from labeled data and can identify new attack variants [3].

***The problem statement.*** Traditional, static, rule-based WAF configurations apply fixed thresholds to individual security features. While these configurations are easy to manage, they are ill-suited for the heterogeneous and constantly changing traffic patterns in university networks, as they cannot capture the complex, nonlinear interactions between high-dimensional security indicators [4]. This leads to increased false-positive rates, which block legitimate academic users, as well as increased false-negative rates, which allow real threats to go undetected. The consequences for ICT systems in the education sector are significant: service disruptions impair the continuity of learning, while undetected intrusion attempts jeopardize obligations to protect student data. Intelligent, adaptive threat classification is therefore an operational prerequisite for responsible ICT management in higher education.

***Analysis of recent studies and publications.*** Research in the field of automated network threat detection has evolved in two main directions. Rule-based WAF configurations and threshold-based filtering [4] offer transparency but are known to be inflexible in the face of evolving attack patterns. Machine learning approaches, which include supervised classification, anomaly detection, and ensemble methods, have demonstrated significantly improved accuracy by learning decision boundaries from labeled security event data [3], [5], [6].

Resende and Drummond [5] found that random forest classifiers achieve an accuracy of approximately 95% on the NSL-KDD intrusion detection benchmark dataset, with tree-based ensembles proving particularly well-suited for effectively processing high-dimensional feature sets of mixed types. Ahmad et al. [6] systematically evaluated SVM, KNN, and Naive Bayes classifiers using standard intrusion detection datasets and confirmed the competitive generalization ability of SVM across various attack categories. Gao et al. [7] demonstrated that adaptive ensemble models, which combine multiple classifiers, outperform individual approaches through the complementary aggregation of learning biases. Khraisat et al. [8] summarized the landscape of intrusion detection and identified ensemble diversity as a primary

determinant of detection robustness, noting that most previous evaluations relied on older network flow datasets rather than the ML-enriched log features generated by modern cloud WAF platforms.

Sommer and Paxson [9] identified a fundamental challenge in applying machine learning to network security: the semantic gap between the outputs of classifiers and the operational requirements of security analysts. Their analysis shows that high classification accuracy alone is insufficient without interpretable decision justifications, a finding that directly justifies the component for extracting interpretable rules in the present framework. Shone et al. [10] proposed a deep autoencoder architecture for network intruder detection that achieves competitive accuracy while reducing reliance on labeled training data. In doing so, they demonstrated the potential of deep learning as a complement to shallow ensemble methods, albeit with limited interpretability in production environments.

Chen and Guestrin [11] demonstrated the superiority of gradient boosting in various benchmark tests for classification in the field of cybersecurity through sequential residual correction. Quinlan [12] established the extraction of decision tree rules as a practical mechanism for translating trained models into human-readable classification guidelines applicable to IDS and firewall configurations. Neither study addresses the specific feature space of modern CBM protocols, which includes attack probability scores generated by Cloudflare Upstream ML, bot confidence signals, and JA3/JA4 TLS fingerprints, features that differ qualitatively from the raw network flow statistics used in traditional intrusion detection research.

In the context of Cloudflare-protected web services, WAF and Bot Management signals represent platform-generated security indicators rather than raw packet- or flow-level measurements. Cloudflare documentation describes WAF Attack Score as a machine-learning-based detection signal used to assess request-level attack likelihood, while Bot Management produces bot-related scores and detection signals that can be used in rules and security analytics [13], [14]. JA3 and JA4 fingerprinting further support the profiling of TLS clients and automated traffic patterns in Cloudflare environments [15]. Therefore, CBM firewall logs constitute a specific type of modern security log data that differs from classical intrusion detection datasets such as NSL-KDD. This distinction is important because the present study does not claim to benchmark general network intrusion detection; rather, it evaluates secondary classification and rule extraction from Cloudflare-enriched firewall records.

Gasimov and Ismailov [16] analyzed information-security systems in the corporate networks of higher education institutions and proposed simulation-based modelling for controlling unauthorized access requests through protection mechanisms. Buinytska et al. [17] examined artificial intelligence in the open university ecosystem and identified infrastructure and security as one of the relevant domains for AI implementation in higher education. Shchavinsky et al. [18] showed that AI-supported methods can improve the development of practical and managerial competencies of cybersecurity specialists in higher education institutions.

Burov et al. [19] considered cybersecurity as an integral part of an innovative digital educational environment and argued that traditional cybersecurity approaches are insufficient against new sophisticated attacks, thereby requiring threat detection, threat discovery, and resilience-oriented strategies. These recent studies strengthen the educational and institutional relevance of the present work, which focuses on ML-based threat detection in university firewall logs.

The reviewed studies also demonstrate that cybersecurity in higher education should be addressed at several interconnected levels. Gasimov and Ismailov [16] focus on the simulation and control of unauthorized access in institutional networks, while Buinytska et al. [17] consider artificial intelligence as a component of the broader open university ecosystem.

Shchavinsky et al. [18] examine AI-supported cybersecurity training, whereas Burov et al. [19] emphasize the resilience of the digital educational environment against emerging threats. However, these studies do not experimentally compare machine-learning classifiers using request-level Cloudflare WAF and Bot Management signals. They also do not investigate how model decisions can be transformed into interpretable candidate firewall rules for university web services. Therefore, the present study complements existing literature by connecting educational ICT security management with the secondary classification and interpretation of modern cloud-based firewall logs.

The literature reviewed [3]-[19] reveals three consistent limitations: (i) To date, no study has conducted a comparative evaluation of six complementary ML classifiers specifically based on Cloudflare CBM WAF log features characteristic of university network environments; (ii) no framework for ICT security in the education sector has yet been proposed that simultaneously enables highly accurate classification and extracts actionable, human-readable firewall rules; and (iii) the direct contribution of such a framework to university network security policies and ICT management practices has not yet been empirically demonstrated.

***The research goal.*** To address these gaps, this study aims to design and empirically evaluate an interpretable secondary machine-learning classifier for Cloudflare Bot Management firewall logs in university digital infrastructure. To this end, six supervised classifiers are integrated with feature importance analysis and the extraction of decision tree rules to enable both highly accurate classification and interpretable, policy-ready firewall rules that are applicable to the university's ICT security management. The following objectives are pursued: (1) building a feature engineering pipeline to extract 24 meaningful security indicators from raw CBM JSON log entries; (2) training, comparing, and cross-validating six supervised classifiers using a stratified 80/20 train-test split and 5-fold cross-validation; and (3) deriving interpretable candidate firewall rules from the trained Decision Tree model for further expert validation in the university's network security management.

The main contributions are: (i) a systematic comparison of six supervised ML classifiers based on CBM-WAF log features reflecting a university web application's traffic; (ii) an analysis of feature importance that identifies WAF ML attack scores as the dominant classification signal within the constructed dataset; and (iii) the derivation of interpretable candidate decision rules that can support, but not replace, expert validation and configuration of educational institutions' firewall policies.

## 2. THE THEORETICAL BACKGROUNDS

University networks represent a unique security environment that must balance the open accessibility essential for academic collaboration, distance learning, and the public dissemination of research results with the protection of sensitive personal data, intellectual property of research, and administrative systems. This tension is further intensified by the volume and heterogeneity of data traffic: a single institutional web application can be used simultaneously by instructors, students, researchers, administrative staff, external collaborators, and automated systems, each generating legitimate traffic profiles that partially overlap with attack signatures [1].

Ensemble machine learning addresses the limitations of individual classifiers by aggregating predictions from multiple models with complementary inductive biases. Random Forest reduces variance through bootstrap aggregation over decision trees [22], while gradient boosting [11] reduces biases through sequential residual correction. Decision trees offer the additional benefit of structural interpretability: trained models can be read directly as conditional classification rules, enabling security administrators without expert knowledge to understand, review, and implement model decisions within existing WAF policies [12].

Recent studies published in Information Technologies and Learning Tools confirm that information and cybersecurity in higher education institutions should be considered as an essential component of the digital educational environment rather than as a purely technical perimeter-protection problem. Gasimov and Ismailov [13] emphasize that the insufficient level of information security in service and corporate networks can lead to significant operational losses and that higher education institutions require systematic mechanisms for controlling unauthorized access requests.

In addition to perimeter-level threat detection, the protection of university digital educational infrastructure also requires complementary mechanisms for securing stored and transmitted information. Gasimov et al. [20] proposed a maze-based image encryption method constructed by random number generation, demonstrating how algorithmic transformation techniques can be used to strengthen data confidentiality. Although encryption methods and WAF-based threat classification address different layers of cybersecurity, both contribute to a multi-layered protection strategy for educational ICT systems. In this context, the present study focuses on the perimeter and log-analysis layer, where machine learning is applied to CBM WAF records for the intelligent detection of malicious network activity.

The integration of machine learning-based threat detection into IT management in education directly addresses the operational requirements for intrusion detection and prevention systems in institutional environments, as formulated by Scarfone and Mell [4]: Automated classification reduces reliance on manual log review, transparent decision rules meet audit and data protection requirements, and adaptive training plans allow security posture to be adjusted to the threat landscape. These features are particularly valuable in the university environment, where resources for security operations are typically limited compared to commercial enterprises [1].

Cloudflare Bot Management generates machine learning-based log features, including wafMlAttackScore, wafRceAttackScore, wafMlXssAttackScore, and wafMlSqliAttackScore, which encode attack-related assessments produced by the upstream Cloudflare security model for common web attack categories. In this article, the term “upstream model” refers to the machine-learning-based scoring mechanism already provided by the Cloudflare security platform before the proposed secondary classifier is trained on exported CBM log records. This concentrates relevant information about web application security threats and reduces the effort required for feature creation in downstream classifiers. This feature makes it particularly suitable for resource-constrained IT environments in education, where ease of use and traceability are paramount.

## 3. RESEARCH METHODS

### 3.1. Dataset and Feature Engineering

The dataset comprises 500 JSON-formatted CBM firewall log entries collected from a Cloudflare CBM-protected web application serving a university. Each raw log entry contains nested security metadata including WAF analysis results, bot detection outputs, and request-level network information representative of academic institution web traffic.

A feature engineering pipeline extracts 24 numerical features per log entry. Table 1 summarizes the four feature categories and their constituent variables.

*Table 1*

**Feature categories and descriptions extracted from CBM firewall logs**

| Category | Feature | Description |
|---|---|---|
| WAF Scores | *wafMlAttackScore* | General ML attack probability |
| | *wafRceAttackScore* | Remote code execution score |
| | *wafMlXssAttackScore* | Cross-site scripting score |
| | *wafMlSqliAttackScore* | SQL injection score |
| Bot Detection | *botScore* | Bot confidence (0 to 99) |
| | *botDetectionCount* | Number of detection signals |
| | *botScoreSrcName* | Detection engine source |
| | *isBot* | UA-based bot flag |
| | *isCrawler* | UA-based crawler flag |
| Network Info | *clientCountryName* | Geographic origin |
| | *clientAsn* | Autonomous system number |
| | *clientASNDescription* | ASN provider description |
| | *ja3HashPresent* | TLS fingerprint presence (JA3) |
| | *ja4SignalCount* | JA4 TLS signal count |
| Request Info | *action* | Firewall action taken |
| | *userAgentLength* | User-agent string length |
| | *clientRequestHTTPProtocol* | HTTP protocol version |
| | *clientRequestHTTPMethodName* | HTTP request method |

The binary target variable threat is derived through a two-step rule-based labelling procedure. This procedure is used because independently verified expert labels were not available for the pilot dataset. Therefore, the target variable should be interpreted as a constructed WAF-score-based threat indicator rather than as an externally validated ground-truth attack label. First, the four WAF attack score dimensions are normalized and combined into a composite threat level score using an arithmetic mean:

$$\text{threat_level}_i = (1/4) \times \Sigma_k \, s_{i,k} \quad (1)$$

where $s_{i,k}$ is the *k*-th WAF attack score for log entry *i*. Second, the 70th percentile threshold $\tau_{0.7}$ is applied to assign the binary label:

$$y_i = 1 \text{ if threat_level}_i > \tau_{0.7}; \text{ otherwise } y_i = 0 \quad (2)$$

In the present dataset, higher normalized values of the four WAF score variables are treated as indicators of higher attack risk according to the internal preprocessing convention used in the experiment. The threshold $\tau_{0.7}$ denotes the 70th percentile of the composite threat level distribution in the available sample. The 70th percentile was selected as an exploratory threshold to obtain a sufficiently populated threat class for model comparison, resulting in 140 threat records and 360 benign records. Equal weights of 1/4 were assigned to the four WAF score dimensions because no expert-defined weighting scheme was available for prioritizing one attack category over another. This design choice avoids introducing an arbitrary preference for SQL injection, XSS, RCE, or general WAF attack scores at the labelling stage. However, because the label is derived from WAF score variables that are also included among the input features, the obtained classification accuracy should be interpreted as evidence of internal consistency and rule extractability, not as proof of independent real-world threat detection performance.

This procedure yields 360 benign entries (72.0%) and 140 threat entries (28.0%). Seven categorical features (action, botScoreSrcName, clientASNDescription, clientCountryName, clientRequestHTTPProtocol, clientRequestHTTPMethodName, source) are encoded using scikit-learn's "LabelEncoder" [21]. Missing values are imputed with zero. All features are standardized with "StandardScaler" to zero mean and unit variance. The dataset is partitioned into training (80%, n = 400) and test (20%, n = 100) subsets using stratified sampling to preserve class proportions.

### 3.2. Classifier Configuration

Six supervised classifiers are trained on identical preprocessed data to compare their suitability for secondary classification of CBM firewall log records. In this study, the term "framework" refers to the full feature engineering, model comparison, evaluation, feature-importance analysis, and rule-extraction pipeline. RF and GB are ensemble classifiers in the strict machine-learning sense, whereas DT, SVM, KNN, and GNB are included as comparative baseline models rather than as components of a single prediction-fusion ensemble. Random Forest (RF): 150 decision trees with maximum depth 20, using bootstrap aggregation to reduce variance [22]. Gradient Boosting (GB): 100 sequential estimators with iterative residual correction [11]. Decision Tree (DT): a single tree with maximum depth is 10, selected for classification and interpretable rule extraction applicable to firewall policy [12]. Support Vector Machine (SVM): RBF kernel with probability calibration [6]. K-Nearest Neighbors (KNN): k = 5, Euclidean distance. Gaussian Naive Bayes (GNB): probabilistic baseline classifier assuming feature-conditional Gaussian distributions.

Model performance is evaluated using accuracy, precision, recall, and F1-score on the held-out test set, supplemented by 5-fold stratified cross-validation on the training set. ROC curves with AUC values and precision-recall curves with average precision scores are generated for all six classifiers. Feature importance scores are derived from the trained Random Forest model. Interpretable decision rules are extracted from the trained Decision Tree using scikit-learn's "export_text" function [21].

## 4.FINDINGS

### 4.1. Classification Performance

Table 2 presents accuracy, precision, recall, F1-score, and 5-fold cross-validation statistics for all six classifiers on the held-out test set (*n* = 100).

*Table 2*

**Classification performance of all six models on CBM firewall log test data**

| Model | Train | Test | Prec. | Rec. | F1 | CV Mean | CV Std |
|---|---|---|---|---|---|---|---|
| Random Forest | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.0000 |
| Gradient Boosting | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.0000 |
| Decision Tree | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.0000 |
| SVM (RBF) | 0.9875 | 0.9600 | 0.9286 | 0.9286 | 0.9286 | 0.9575 | 0.0269 |
| KNN (k=5) | 0.9525 | 0.9300 | 0.8621 | 0.8929 | 0.8772 | 0.9150 | 0.0421 |
| Gaussian NB | 0.5250 | 0.4400 | 0.3293 | 0.9643 | 0.4909 | 0.5200 | 0.0504 |

RF, GB, and DT achieved perfect scores on both training and test sets in the constructed dataset (accuracy = 1.0000, F1 = 1.0000) with zero cross-validation variance. This result should be interpreted cautiously. The target variable was derived from WAF score aggregation, and WAF score variables were also included among the model inputs. Therefore, the perfect performance of the tree-based models primarily indicates that they successfully recovered the rule structure embedded in the constructed labels. SVM attained 96.00% test accuracy with balanced precision and recall (0.9286 each). KNN reached 93.00% test accuracy with marginally reduced precision (0.8621). GNB exhibited the lowest performance (test accuracy = 44.00%), with high recall (0.9643) but very low precision (0.3293), indicating systematic over-prediction of the threat class and rendering it unsuitable for direct deployment in university WAF environments without substantial feature preprocessing.

Figure 1 presents confusion matrices for all six classifiers on the test set.

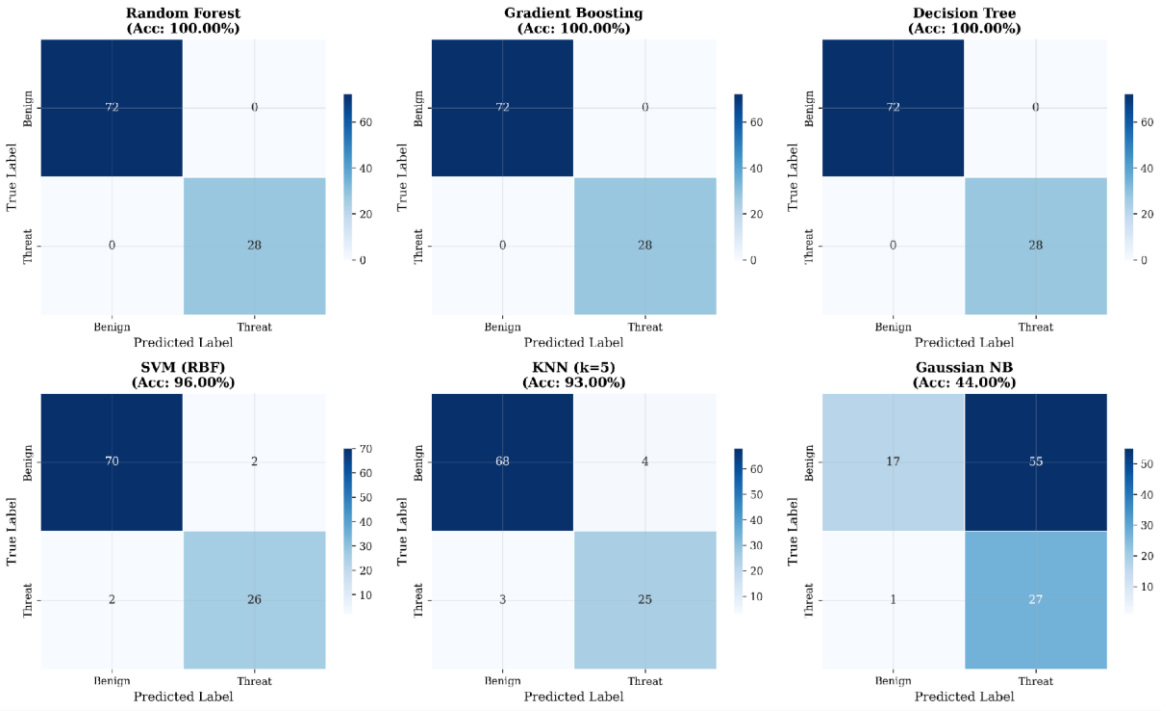



*Figure 1. Confusion matrices for all six classifiers on the test set (n = 100)*

In a university WAF environment, false positives correspond to legitimate academic requests that may be unnecessarily blocked or challenged, while false negatives correspond to potentially malicious requests that remain undetected.

*Table 3*

**False-positive and false-negative cases for each classifier on the test set**

| Model | TN | FP | FN | TP |
|---|---|---|---|---|
| Random Forest | 72 | 0 | 0 | 28 |
| Gradient Boosting | 72 | 0 | 0 | 28 |
| Decision Tree | 72 | 0 | 0 | 28 |
| SVM (RBF) | 70 | 2 | 2 | 26 |
| KNN (k = 5) | 68 | 4 | 3 | 25 |
| Gaussian NB | 17 | 55 | 1 | 27 |

From this perspective, RF, GB, and DT showed no error cases on the constructed test set. SVM and KNN produced a limited number of both FP and FN cases. GNB produced an excessive number of false positives, which would be unacceptable in practice because it could disrupt access to learning platforms, student portals, and administrative services.

Figure 2 shows a grouped bar chart comparing accuracy, precision, recall, and F1-score across all six classifiers.

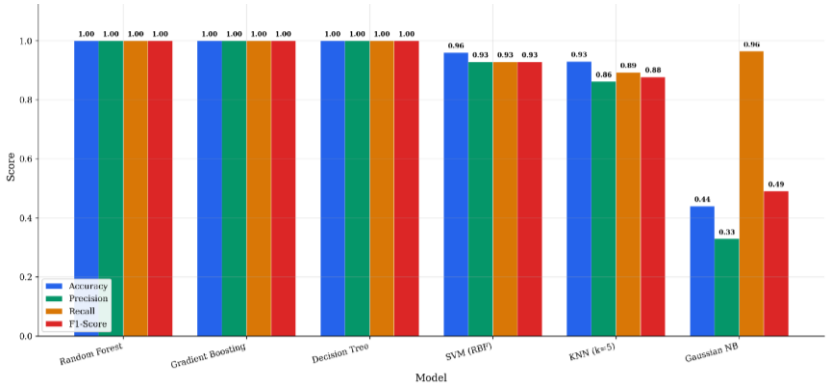



*Figure 2. Comparative model performance: accuracy, precision, recall, and F1-score*

### 4.2. Feature Importance Analysis

Table 4 lists the top ten features ranked by Random Forest importance scores, representing the mean decrease in impurity across all trees attributable to each feature.

*Table 4*

**Top ten features ranked by Random Forest importance score**

| Rank | Feature | Importance | Cumulative |
|---|---|---|---|
| 1 | wafMlAttackScore | 37.61% | 37.61% |
| 2 | wafRceAttackScore | 26.98% | 64.59% |
| 3 | wafMlXssAttackScore | 6.60% | 71.19% |
| 4 | userAgentLength | 6.32% | 77.51% |
| 5 | wafMlSqliAttackScore | 4.84% | 82.35% |
| 6 | clientRequestHTTPProtocol | 3.91% | 86.26% |
| 7 | botScore | 3.71% | 89.97% |
| 8 | clientCountryName | 3.27% | 93.24% |
| 9 | clientAsn | 1.57% | 94.81% |
| 10 | clientASNDescription | 1.52% | 96.33% |

The four WAF ML attack score variables collectively account for 76.03% of total Random Forest feature importance in the constructed dataset: wafMlAttackScore at 37.61%, wafRceAttackScore at 26.98%, wafMlXssAttackScore at 6.60%, and wafMlSqliAttackScore at 4.84%. The top five features explain 82.35% of model decisions. Non-WAF features, including userAgentLength (6.32%), clientRequestHTTPProtocol (3.91%), botScore (3.71%), clientCountryName (3.27%), and network identifiers, account for the remaining 17.65%. This confirms that the model decisions are primarily driven by the same group of variables used to construct the target label, which further supports the interpretation of the framework as a secondary classifier and rule-extraction mechanism rather than an independent detector trained on expert-labelled attacks.

Figure 3 presents the feature importance bar chart for the top fifteen features.

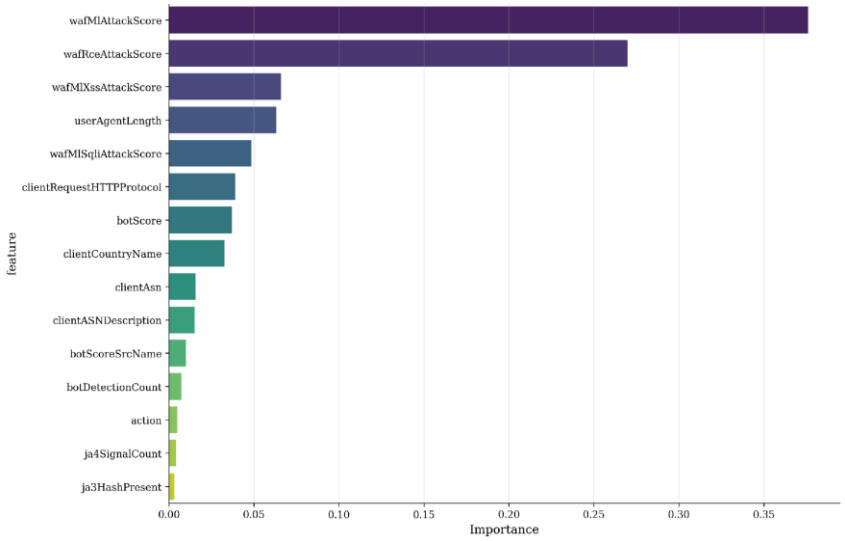



*Figure 3. Top 15 features ranked by Random Forest importance score*

Figure 4 presents the Pearson correlation matrix for the top fifteen features. Pearson correlation was selected because the main variables of interest, especially the WAF score variables, are numerical and continuous after preprocessing. The matrix shows strong positive intercorrelations among several WAF ML attack score dimensions, which is expected because different web attack categories may co-occur or be jointly reflected in Cloudflare's scoring mechanism.

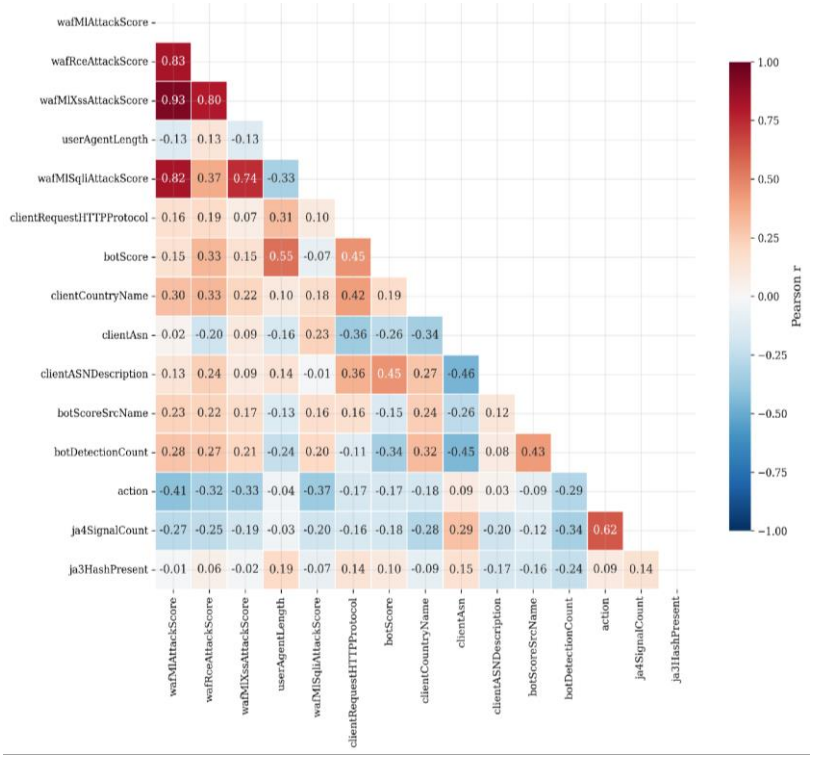



*Figure 4. Pearson correlation matrix for the top 15 features*

However, Pearson correlation is not fully appropriate for interpreting label-encoded categorical variables, such as country, protocol, or ASN description. Therefore, correlations

involving encoded categorical features are treated only as exploratory indicators and are not used as evidence of semantic relationships between categories.

### 4.3. Interpretable Firewall Decision Rules

The DT classifier generated a single interpretable classification rule that reproduced the constructed labels on the test set: a log entry is classified as a threat candidate if the normalized wafMlAttackScore exceeds 0.20; otherwise, it is classified as harmless. This result is theoretically consistent because the composite threat label in equation (1) is derived from WAF score aggregation, and wafMlAttackScore is the most informative component of that aggregation. Therefore, the rule should be interpreted as an interpretable approximation of the constructed WAF-score-based labelling mechanism rather than as an independently validated production rule. This creates a feature space where a single threshold effectively captures the binary class boundary.

Three candidate firewall policy formulations informed by the comparative machine-learning framework are: (i) BLOCK if wafMlAttackScore > 0.20; (ii) BLOCK if wafRceAttackScore > 0.25, which addresses remote code execution risk, particularly relevant to research environments; and (iii) WARN if wafMlXssAttackScore or wafMlSqliAttackScore exceeds 0.20, targeting attack vectors commonly associated with student portal login forms and database-driven academic applications. These rules can be translated into candidate Cloudflare WAF custom rule configurations and evaluated by institutional security administrators. Before production deployment, however, they should be validated against expert-labelled traffic, service availability indicators, and false-blocking statistics to avoid unnecessary disruption of legitimate academic users.

### 4.4. Discussion

The perfect classification accuracy of RF, GB, and DT reflects the high discriminatory power of Cloudflare's ML-generated WAF attack assessment features. These encode the attack probability estimated by the upstream model as continuous numerical signals rather than raw network flow statistics. Although this result is numerically five percentage points higher than the 95% NSL-KDD result reported by Resende and Drummond [5], the two values are not directly comparable because they were obtained using different datasets, feature spaces, and labelling procedures. The high performance observed in the present study is primarily associated with the information density of the ML-enriched CBM features and the relationship between the WAF score variables and the constructed target labels.

The lower accuracy of SVM (96.00%) may reflect the sensitivity of RBF kernel methods to less informative metadata, which introduces noise into the high-dimensional kernel representation. The accuracy of KNN (93.00%) is consistent with the known limitations of distance-based methods in high-dimensional spaces, where non-discriminative features dilute the Euclidean distance metric. The test accuracy of GNB, 44.00%, results from a violation of the assumption of conditional independence, as the features of the WAF attack score are highly correlated (Figure 4), thus structurally invalidating the Naive Bayes factorization.

These results are consistent with the observation by Sommer and Paxson [9] that the selection of a network security classifier must consider both the distributional properties of the feature space and the operational interpretability requirements for security personnel. Extracting the DT rules closes the semantic gap identified in [9]: security administrators gain not only a highly accurate classifier but also an easily understandable threshold rule that is compatible with existing WAF management workflows. This bridges the gap between machine learning research and the practical application of IT security in education, without requiring specialized expertise from the institutions' IT staff.

Compared with manually configured single-threshold WAF rules [4], the proposed comparative ML framework evaluates 24 engineered features simultaneously and provides a basis for analyzing how different classifiers separate constructed threat and benign classes. However, the present study does not include a direct experimental comparison with production Cloudflare rules or real false-blocking statistics. Therefore, the reduction of false positives should be considered a potential benefit to be verified in future deployment-oriented experiments rather than a proven operational outcome. The interpretable DT rule bridges the gap between machine learning research and institutional policy, enabling security administrators to integrate machine learning-based thresholds into standard WAF management workflows.

A practical use case of the proposed model in a university environment is the post-processing of exported CBM firewall logs by the institutional IT security team. For example, the model can be used to identify records whose WAF score patterns correspond to threat-candidate behavior, rank the most influential features, and generate simple candidate rules for review. Security administrators may then compare these candidate rules with existing Cloudflare custom rules, inspect whether legitimate users from student portals or learning management systems would be affected, and test the rules in logging or monitoring mode before applying blocking actions. In this way, the model supports rule interpretation, auditability, and policy refinement, but does not replace expert judgement or production security validation.

This study has certain limitations. The 500-entry dataset is from a single university web application and should be considered a pilot-scale dataset. This sample size is insufficient for drawing general conclusions about large-scale security systems or diverse university traffic patterns. The results therefore demonstrate the feasibility of secondary CBM log classification and rule extraction, but they do not establish production-level robustness. Future experiments should use substantially larger log collections, preferably including more than 10,000 records, multiple academic services, different traffic periods, and expert-verified labels. The quantile-based threat labeling (Equation 2) approximates expert judgments rather than independently verified reference data; the near-perfect accuracy of tree-based classifiers partially reflects the coherence between the feature space and the labeling mechanism. The class ratio of 72/28 may not represent the more extreme imbalance typical of production university WAF environments and therefore requires careful evaluation before widespread deployment.

## 5. CONCLUSIONS AND PROSPECTS FOR FURTHER RESEARCH

This study designed and evaluated a comparative machine-learning framework for secondary classification and interpretation of Cloudflare Bot Management firewall logs in university digital educational infrastructure. The results demonstrate the feasibility of extracting interpretable candidate firewall rules from CBM log features, but they should be interpreted as pilot-scale findings requiring further validation before production deployment. The study produced three main findings.

First, the feature engineering pipeline extracted 24 discriminative security indicators from the raw CBM JSON log entries. These indicators include WAF-ML attack scores, bot detection signals, TLS fingerprint data, and network metadata. Combined threat labelling using equations (1) and (2) yielded 360 benign and 140 threat entries. This provided a sufficiently populated class distribution for the pilot training and evaluation conducted in this study.

Second, the comparative evaluation showed that the classifiers RF, GB, and DT achieved perfect classification performance (accuracy = 1.0000, F1 = 1.0000) with no cross-validation variance, while SVM and KNN achieved test accuracy of 96.00% and 93.00%, respectively. These results should not be interpreted as a direct improvement over classical intrusion detection benchmarks such as NSL-KDD, because the present dataset consists of Cloudflare-

enriched WAF log records and uses a rule-derived target variable. The comparison with earlier intrusion detection studies is therefore contextual rather than benchmark equivalent.

Third, the feature importance analysis identified the wafMlAttackScore as the dominant classifier input, with an importance of 37.61%. The four WAF rating dimensions together contributed 76.03% to the classification decisions. The DT model generated a single interpretable rule (BLOCK if wafMlAttackScore > 0.20) that achieved 100% accuracy and could be directly translated into custom Cloudflare WAF rule policies by university security administrators without specialized machine learning expertise.

These results indicate that machine-learning analysis of CBM log features can support the interpretation of firewall records and the derivation of candidate WAF rules for educational web security. However, independent expert-labelled datasets, comparison with existing Cloudflare rule configurations, and post-deployment analysis of service availability are necessary before the approach can be considered operationally validated. The framework bridges the gap between machine learning research and institutional IT security practice, delivering both precise classification methods and easily understandable policies compatible with existing WAF management workflows. This supports the protection of open digital learning environments, student data repositories, and research platforms upon which modern higher education is built.

Future research should validate the framework using extensive, production-scale CBM log datasets with more than 10,000 entries and expert-verified threat labels. The threshold used for target construction should be estimated on training data only and then evaluated on unseen test data. Additional experiments should compare the proposed classifier with existing Cloudflare custom rules, analyze false-positive and false-negative cases in real service conditions, and assess whether the model affects the availability of learning platforms, student portals, and administrative systems. Integrating deep learning architectures and advanced ensemble strategies such as stacking and soft voting will determine whether higher model complexity, compared to tree-based classifiers, leads to performance improvements in production environments. Temporal analysis of shifts in feature distribution across academic cycles, reflecting differences in traffic patterns during semesters and breaks, will enable adaptive training plans tailored to university operations. Extending the framework to include multi-class threat categorization that distinguishes between SQL injection, XSS, RCE, bot attacks, and harmless traffic would significantly enhance its practical value for developing a detailed IT security policy at universities.

## DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

## FUNDING

The study was conducted without financial support.

## CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

**Authors' contributions**

**Artughrul Gayibov:** Software; Investigation; Data curation; Visualization; Formal analysis; Validation. **Vagif Gasimov**: Conceptualization; Supervision; Project administration. **Islam Islamov**: Methodology; Validation; Writing - review & editing. **Jabir Mammadov**: Methodology; Formal analysis; Validation; Resources; Data curation; Writing - review & editing. **Seyran Gayibov**: Investigation; Writing - original draft; Writing - review & editing; Resources; Project administration.

**REFERENCES (TRANSLATED AND TRANSLITERATED)**


[1] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys and Tutorials*, vol. 18, no. 2, pp. 1153-1176, 2016. doi: 10.1109/COMST.2015.2494502. (in English)

[2] Verizon, "*2023 Data* Breach Investigations Report," Verizon Communications Inc., 2023. [Online]. Available: https://www.verizon.com/business/resources/reports/dbir/ (in English)

[3] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, and C. Wang, "Machine learning and deep learning methods for cybersecurity," *IEEE Access*, vol. 6, pp. 35365-35381, 2018. doi: 10.1109/ACCESS.2018.2836950. (in English)

[4] K. Scarfone and P. Mell, "Guide to intrusion detection and prevention systems (IDPS)," *NIST Special Publication 800-94*, National Institute of Standards and Technology, 2007. doi: 10.6028/NIST.SP.800-94. (in English)

[5] P. A. A. Resende and A. C. Drummond, "A survey of random forest based methods for intrusion detection systems," *ACM Computing Surveys*, vol. 51, no. 3, pp. 1-36, 2018. doi: 10.1145/3178582. (in English)

[6] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 1, e4150, 2021. doi: 10.1002/ett.4150. (in English)

[7] X. Gao, C. Shan, C. Hu, Z. Niu, and Z. Liu, "An adaptive ensemble machine learning model for intrusion detection," *IEEE Access*, vol. 7, pp. 82512-82521, 2019. doi: 10.1109/ACCESS.2019.2923640. (in English)

[8] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: Techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, pp. 1-22, 2019. doi: 10.1186/s42400-019-0038-7. (in English)

[9] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in Proc. *2010 IEEE Symposium on Security and Privacy*, Berkeley, CA, USA, pp. 305-316, 2010. doi: 10.1109/SP.2010.25. (in English)

[10] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A deep learning approach to network intrusion detection," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 1, pp. 41-50, Feb. 2018. doi: 10.1109/TETCI.2017.2772792. (in English)

[11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, 2016. doi: 10.1145/2939672.2939785. (in English)

[12] J. R. Quinlan, C4.5: Programs for Machine Learning. San Francisco, *CA: Morgan Kaufmann*, 1993. doi: 10.1016/C2009-0-27846-9. (in English)

[13] Cloudflare, "WAF attack score," *Cloudflare Developers Documentation*, 2026. [Online]. Available: https://developers.cloudflare.com/waf/detections/attack-score/ (accessed Jun. 9, 2026). (in English)

[14] Cloudflare, "Bot Management," *Cloudflare Developers Documentation*, 2026. [Online]. Available: https://developers.cloudflare.com/bots/get-started/bot-management/ (accessed Jun. 9, 2026). (in English)

[15] Cloudflare, "JA3/JA4 fingerprint," *Cloudflare Bot Solutions Documentation*, 2026. [Online]. Available: https://developers.cloudflare.com/bots/additional-configurations/ja3-ja4-fingerprint/ (accessed Jun. 9, 2026). (in English)

[16] V. Gasimov and B. Ismailov, "Analysis of the results of simulative modelling of the information security system in the corporate networks of higher education institutions," *Information Technologies and Learning Tools*, vol. 102, no. 4, pp. 211–223, 2024. doi: 10.33407/itlt.v102i4.5757. (in English)
[17] O. Buinytska, T. Terletska, V. Smirnova, A. Tiutiunnyk, I. Kovalenko, and B. Hrytseliak, "Artificial intelligence in open university ecosystem context," *Information Technologies and Learning Tools*, vol. 105, no. 1, pp. 204–221, 2025. doi: 10.33407/itlt.v105i1.5959. (in English)
[18] Y. V. Shchavinsky, T. M. Muzhanova, Y. M. Yakymenko, and M. M. Zaporozhchenko, "Application of artificial intelligence for improving situational training of cybersecurity specialists," *Information Technologies and Learning Tools*, vol. 97, no. 5, pp. 215–226, 2023. doi: 10.33407/itlt.v97i5.5424. (in English)
[19] O. Burov, O. Butnik-Siversky, O. Orliuk, and K. Horska, "Cybersecurity and innovative digital educational environment," *Information Technologies and Learning Tools*, vol. 80, no. 6, pp. 414–430, 2020. doi: 10.33407/itlt.v80i6.4159. (in English)
[20] V. A. Gasimov, N. F. Mammadzada, J. I. Mammadov, and E. A. Mustafayeva, "Maze based image encryption method constructed by random number generation," *Eurasian Journal of Mathematical and Computer Applications*, vol. 12, no. 3, pp. 35–50, 2024. doi: 10.32523/2306-6172-2024-12-3-35-50. (in English)
[21] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf. (in English)
[22] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324. (in English)

*Text of the article was accepted by Editorial Team 04.05.2026*

# ЗАСТОСУВАННЯ МАШИННОГО НАВЧАННЯ ДЛЯ КЛАСИФІКАЦІЇ ПОТЕНЦІЙНИХ ЗАГРОЗ У ЖУРНАЛАХ CLOUDFLARE BOT MANAGEMENT ЦИФРОВОЇ ОСВІТНЬОЇ ІНФРАСТРУКТУРИ УНІВЕРСИТЕТУ

**Артугрул Гайібов**
Бакинський інженерний університет, м. Хирдалан, Азербайджан
ORCID ID 0009-0009-7349-0286
*agayibov@beu.edu.az*

**Вагіф Гасімов**
Бакинський інженерний університет, м. Хирдалан, Азербайджан
ORCID: 0000-0003-3192-4225
*vaqasimov@beu.edu.az*

**Іслам Ісламов**
Бакинський інженерний університет, м. Хирдалан, Азербайджан
ORCID: 0000-0001-8645-0640
*isislamov@beu.edu.az*

**Джабір Мамедов**
Бакинський інженерний університет, м. Хирдалан, Азербайджан
ORCID: 0000-0003-3939-4708
*camammadov@beu.edu.az*

**Сейран Гайібов**
Бакинський інженерний університет, м. Хирдалан, Азербайджан
ORCID ID: 0000-0001-7434-6248
*sqayibov@beu.edu.az*

**Анотація.** Об’єктом цього дослідження є автоматизована класифікація записів журналу брандмауера Cloudflare Bot Management (CBM), зібраних з вебзастосунку університету. Дослідження розглядає проблему розрізнення потенційно шкідливих запитів від легітимного академічного трафіку в цифровій освітній інфраструктурі вищого навчального закладу. Статичні правила брандмауера на основі порогів часто недостатні для аналізу багатовимірних показників безпеки, включно з оцінками атак машинного навчання WAF,

сигналами, пов’язаними з ботами, інформацією про відбитки TLS та метадані рівня запитів. Порівняльну структуру машинного навчання було розроблено з використанням шести контрольованих класифікаторів: Випадковий ліс (RF), Градієнтне підвищення (GB), Дерево рішень (DT), Машина опорних векторів (SVM), K-найближчі сусіди (KNN) та Гаусівський наївний байєсівський метод (GNB). Моделі були навчені та оцінені на 24 інженерних ознаках, витягнутих з 500 записів журналу СВМ, за допомогою стратифікованого поділу навчальних тестів 80/20 та 5-кратної перехресної перевірки. RF, GB та DT досягли 100% точності тестування на побудованому наборі даних, тоді як SVM, KNN та GNB досягли 96,00%, 93,00% та 44,00% відповідно. Ці результати слід інтерпретувати в контексті запропонованої процедури маркування, оскільки цільова змінна була отримана з агрегації оцінок WAF і та сама група показників WAF також використовувалась серед вхідних ознак. Аналіз важливості ознак показав, що оцінки атак машинного навчання WAF формують домінуючу групу класифікаційних показників. Модель дерева рішень створила інтерпретоване порогове правило на основі нормалізованого «wafMlAttackScore». Це правило представлено не як повністю розгорнута політика брандмауера виробничого рівня, а як правило-кандидат, яке може підтримувати подальшу експертну перевірку та налаштування політик безпеки університету WAF. Отже, запропонований підхід сприяє інтерпретації та вторинній класифікації журналів брандмауера СВМ в управлінні безпекою освітніх ІКТ.

**Ключові слова:** безпека університетської мережі; ансамблеве машинне навчання; виявлення загроз WAF; аналіз журналів брандмауера; освітня ІКТ-інфраструктура.

(cc) BY-NC-SA

This work is licensed under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.