Eduvest – Journal of Universal Studies
Volume 6 Number 8, August, 2026
p- ISSN 2775-3735- e-ISSN 2775-3727

Design and Build a Machine Learning-Based Brute-Force Login Early Detection System with Adaptive Risk Scoring and Automated Actions

Tubagus Setyo Mulyatama*, Galura Muhammad Suranegara
Universitas Pendidikan Indonesia, Indonesia
Email: bagustama21@upi.edu*, galurams@upi.edu

Keywords:
early detection; brute-force login; machine learning; random forest; adaptive risk scoring

Abstract

Brute-force login attacks pose a significant cybersecurity threat to information systems, potentially leading to unauthorized access and data loss. This research designs and builds an early detection system for brute-force login attacks by integrating a machine learning-based classification model, an adaptive risk scoring mechanism, and automated response actions. The Random Forest model was trained using login attempt log data collected from a controlled test website over two months with 100 dummy accounts, resulting in 30,000 entries comprising 17,000 normal, 5,000 suspicious, and 8,000 attack data samples. The adaptive risk scoring mechanism combines the ML model output with four contextual indicators frequency of failed attempts, access time, IP location, and user agent to generate a dynamic risk score, enabling the system to make decisions based on multiple indicators rather than a single factor. Evaluation results demonstrate 96.5% accuracy, 95.8% precision, and 97.2% recall in three-class classification. Ablation studies confirm that the integration of adaptive risk scoring improves accuracy from 91.2% (RF only) to 96.5% (RF with all indicators). The system successfully blocked 8 IP addresses identified as attacking sources, achieving a false positive rate of 0.7% and an average response time of 42 milliseconds. This study concludes that the combination of Random Forest, adaptive risk scoring, and automated actions provides an effective and comprehensive approach for the early detection of brute-force login attacks.

INTRODUCTION

Brute-force login attacks are one of the most common cyberattack methods used to gain unauthorized access to information systems (Qian et al., 2022; Sekar et al., 2023; Vugdelija et al., 2021). This attack works by massively trying to combine usernames and passwords until they find valid credentials (Fachri, 2023). Although conventional mitigation mechanisms such as rate-limiting and CAPTCHA have been widely implemented, brute-force attacks remain a serious threat due to their ability to adapt attack patterns to evade static rule-based detection (Abdullah and Fachri, 2025).

Several studies have developed attack detection systems using machine learning approaches (Ali Hamza and Jumma surayh Al-Janabi, 2024). developed a classification model to detect brute-force attacks based on network logs, while (Sarker et al., 2020) presented a comprehensive review of the application of data science in cybersecurity. However, such

studies generally focus on detection as-pack alone, without integrating dynamic risk assessment mechanisms or automated response actions (Jiang et al., 2026; Sangeeta & Roderick, 2025).

Identified literature gaps include: (1) lack of brute-force detection systems that combine machine learning with adaptive risk scoring for more granular threat assessments (Khanam et al., 2020); (2) there is no automatic action mechanism that is directly integrated with the results of risk detection and assessment; and (3) the lack of studies that evaluate the contribution of each component (ML, risk scoring, automated actions) to the overall performance of the system through ablation studies.

The novelty of this research lies in three main aspects: first, it proposes an integrated system architecture that combines three components Random Forest machine learning model, adaptive risk scoring with four contextual indicators (frequency of failed attempts, access time, IP location, and user agent), and automated response actions into a single interconnected framework, whereas previous studies typically addressed these components separately; second, it introduces a novel adaptive risk scoring formula that dynamically combines ML model output with contextual indicators, with weights empirically determined through grid search optimization, enabling more granular threat assessment than static or single-indicator approaches; and third, it conducts ablation studies to quantitatively measure the contribution of each component to overall system performance, providing empirical evidence of the effectiveness of the integrated approach a contribution largely absent in existing literature.

Based on these gaps, this study formulated three research questions: (1) How accurate is the Random Forest machine learning model in classifying login activities into three categories (normal, suspicious, attack)? (2) How effective is adaptive risk scoring that integrates ML model output with contextual indicators (test frequency, access time, IP location, user agent) in providing dynamic risk assessment? (3) How effective is the automated action mechanism in mitigating detected brute-force attacks?

Based on the literature on the effectiveness of ensemble learning in intrusion detection (Saputra et al., 2025) and the concept of adaptive risk management in cybersecurity (Chandra et al., 2022), the hypothesis of this study is: (1) the Random Forest model is able to achieve classification accuracy above 95% in three-class login log data; (2) the integration of adaptive risk scoring can improve detection accuracy compared to ML models alone; and (3) risk score-based automated actions are able to block attacks with a false positive rate below 1%.

This research aims to design and build an early detection system for brute-force login attacks that integrates three main components: the Random Forest machine learning model for the classification of login activities, an adaptive risk scoring mechanism for dynamic risk assessment, and an automated action module for attack mitigation. The system was tested on the independently developed Ravita Health website, with log data collected from controlled testing using a variety of attack scenarios. The contributions of this research include the integration of three components ML, risk scoring, and automation in one interconnected system architecture, an adaptive risk scoring formula with four contextual indicators whose weights are determined through empirical tests, and ablation studies that prove the contribution of each component to improving overall system accuracy.

METHODS

Research Design

This study adopted a quantitative approach with an experimental and computational system design method to objectively measure the effectiveness of the early detection system for brute-force login attacks. The research stages include data collection from the independently developed Ravita Health website login logs over a 2-month period, data pre-processing, development of a Random Forest machine learning model for three-class classification (normal, suspicious, attack), design of an adaptive risk scoring mechanism integrating ML model output with four contextual indicators (frequency of failed attempts, access time, IP location, and user agent), implementation of automated actions such as temporary or permanent IP blocking based on risk score thresholds, and evaluation of overall system performance using metrics including accuracy, precision, recall, F1-score, false positive rate, false negative rate, and response time latency. The system design process begins with the identification of functional and non-functional requirements detection accuracy, response speed, and scalability followed by feature extraction and selection to optimize model efficiency. The adaptive risk scoring mechanism combines predictive results from the ML model with contextual factors to generate a dynamic risk score, which serves as the basis for automated responses, including IP blocking, login delays, administrator notifications, or additional verification requests, with the choice of action depending on the detected risk level. End-to-end system evaluation involves testing realistic attack scenarios to measure detection effectiveness, risk score accuracy, and the success of automated actions in preventing or mitigating brute-force login attacks, with the goal of filling the gap in the development of adaptive and proactive cybersecurity systems in Indonesian institutions by taking into account the characteristics of local threats and specific security needs (Safitra, 2023).

Population and sample

The population in this study was all logs of login attempts that occurred on the Ravita Health website during the 2-month data collection period. The Ravita Health website is an independently developed system with a login authentication feature, making it relevant for testing the detection capabilities of brute-force attacks. The data population consists of 30,000 en-tri logs that cover all patterns of login activity, whether normal, suspicious, or attacked.

The sample used in this study was all 30,000 log entries, which had been processed and annotated into three categories: 17,000 normal data (57%), 5,000 suspicious data (17%), and 8,000 attack data (27%). Classification by frequency of attempts: 1–2 failed attempts (normal), 5–6 failed attempts (suspicious), and 10+ failed attempts (attacks). This distribution ensures that the dataset is diverse and representative enough to train machine learning models.

The sample dataset has been divided into two main parts: one for training the machi-ne learning model and the other for independent system performance testing. This division ensures that model performance evaluations are carried out on data that has never been seen before, thus providing an objective estimate of the generalization capabilities of the system. The amount of data used in training and testing has been adjusted to achieve a balance between model complexity and computational needs.

The sample data used was collected from system logs generated through controlled simulations, where normal login attempts and brute-force attacks were generated

systematically, using Python scripts. The use of this controlled simulation data has a caveat because it allows for the creation of diverse and scalable attack patterns, as well as conditions that can be replicated. This data has been further processed through the pre-processing stage to remove noise and extract relevant features for analysis.

The Ravita Health website was chosen because it is an independently developed system with complete login authentication features and 100 dummy accounts for testing purposes, allowing the testing of all system components (ML detection, risk scoring, and automated actions) in one controlled environment without involving user data. The use of the website itself also allows for the free manipulation of attack parameters without affecting the production system.

Data Collection

Data collection was carried out through the extraction of login attempt logs from the Ravita Heal-th website. These logs record every authentication attempt that occurs, including information such as the source IP address, username, timestamp, test status (successful or failed), and the device type used. This log data is a primary source that has been used to train and evaluate brute-force attack detection models. The selection of the Ra-vita Health website as a data source was made based on the relevance and availability of comprehensive logs.

Login attempt log data was collected over a 2-month period through controlled simulations using Python scripts. Data collection is done in stages: the first month focuses on normal data generation (successful login) and suspicious data (5-6 failed attempts), while the second month focuses on attack data generation (10+ failed attempts). This duration of data collection ensures sufficient variability in the dataset, so that the model can learn to recognize diverse patterns.

In addition to the login attempt log, other relevant supporting data has been collected, such as information regarding system security configuration, access policies, and network configuration. This additional data can help in the analysis of the attack context and the design of more accurate adaptive risk scoring mechanisms. The integration of data from multiple sources has provided a more holistic view of system security.

Once the log data is collected, the next stage is to pre-process the data. This pre-processing includes the cleanup of data from irrelevant or duplicate entries, the normalization of data formats, and the handling of missing values. Feature transformation has also been implemented to convert raw data into a format suitable for use by machine learning algorithms. This process is crucial to ensure the quality of the data used in the training of the model.

In the context of cyber security in Indonesia, understanding specific attack patterns is important. Research on brute-force attacks in Indonesian university environments shows that vulnerabilities such as open ports and weak access controls can be exploited (Abdullah, 2025; Fachri, 2023). Therefore, the data collected must be able to reflect the characteristics of the attack that are relevant to that context.

Data Analysis Techniques

The data analysis in this study has focused on two main aspects: brute-force pattern analysis and performance evaluation of detection systems. Attack pattern analysis involves identifying characteristics and trends from suspicious login attempt log data. It includes analysis of the frequency of failed attempts, attack time patterns, the distribution of source IP

addresses, as well as the types of credentials that are frequently used in failed attempts. A deep understanding of these patterns is crucial to building an effective detection model.

For attack detection, a machine learning algorithm is used that is able to classify each login attempt. The classification model has been trained using log data that has been annotated as normal, suspicious, or attack attempts. Algorithms commonly used in intrusion detection such as Random Forest or XGBoost can be considered, given their ability to handle large and complex datasets (Saputra, 2025; Tyas, 2026). The evaluation of the model's performance has used quantitative metrics such as aku-ratio, precision, recall, and F1-score to measure how well the model can distinguish between normal activity and attack.

The adaptive risk scoring mechanism has been analyzed based on its ability to provide dynamic and accurate risk scores. This analysis includes an evaluation of how various parameters, such as the frequency of failed login attempts of a single IP or user in a given time interval, affect a given risk score. Higher risk scores have triggered more decisive response actions. This approach is a development of a risk assessment model that is commonly applied in cybersecurity (Chandra, 2022; Putra, 2023).

Automated action analysis has focused on its effectiveness in mitigating detected attacks. This includes an evaluation of how quickly response actions are taken after an attack is identified, as well as the impact of those actions on the sustainability of the attack. Measures such as IP blocking, delaying login attempts, or notifying administrators have been evaluated for their effectiveness in preventing unauthorized access and minimizing the negative impact of attacks.

All data analysis has been carried out using appropriate statistical and programming software. The results of the analysis have been interpreted in the context of cyber security in Indo-nesia, considering the findings of previous research on cyberattacks in local institutions (Arifman, 2026; Fauziyyah, 2025). The validity and reliability of the analysis results have been maintained through repeated testing and cross-validation on different datasets.

RESULTS AND DISCUSSION

Data Description and System Characteristics

This study uses a log set of login experiments collected from the independently built Ravi-Ta Health website. The website uses 100 dummy accounts (not real users) for testing purposes. This log data includes various important attributes such as the source IP address, username, timestamp, authentication attempt status (successful/failed), and user agent type. The dataset was obtained over a 2-month period through controlled simulations using Python scripts, with a total of 30,000 entries consisting of 17,000 normal data, 5,000 suspicious data, and 8,000 attack data. The raw data then goes through a pre-processing process for cleaning, normalization, and extraction of relevant features.

The target system where this detection is implemented is the Ravita Health website which was developed independently. This website has a login authentication feature that is the object of testing brute-force attacks. The test was carried out through a login attempt on the Ravita Health website locally (localhost), with manual responses handled through the Ravita Admin website. The website is run on laptop 1 (server) connected to home wifi, while the attack is carried out from laptop 2 (attacker) connected via cellphone hotspot or mobile data.

Descriptive Statistics of Research Variables

The following table presents descriptive statistics for some of the key features extracted from the login experiment log dataset. These features include the number of failed login attempts per IP in a given time interval, the number of successful login attempts per IP, and the duration between login attempts. These statistics provide a preliminary idea of the distribution and variability of the data that will be used to train machine learning models and adaptive risk scoring mechanisms.

Table 1. Descriptive Statistics Key Features

Features	N	Red	SD	Min	Max
Login Failed/IP	30.000	1,85	4,21	0	58
Successful Login/IP	30.000	10,52	25,88	0	310
Duration between logins (sec)	30.000	15,78	30,55	0	300
Number of Unique IPs	30.000	1,22	0,85	1	15

Source: Primary Data, Processed (2026)

Based on descriptive statistics, it can be seen that the average failed login attempt per IP is relatively low, but the maximum value is quite high, indicating malicious activity from several IP sources. Similarly, the variability in the duration between login attempts indicates a different pattern, which can be exploited by the Ma-Chine Learning model.

Results of Machine Learning Model Development

The development of a machine learning model using the Random Forest (RF) algorithm focused on classifying login activities into three categories: normal, suspicious, and brute-force attacks. The processed dataset, consisting of si-extracted features, was used to train and test this model. The model's performance evaluation metrics are presented in the following table:

Model configuration and data sharing

The login attempt log dataset used consists of 30,000 entries that have been annotated into three categories: normal (17,000 entries, 57%), suspicious (5,000 entries, 17%), and attack (8,000 entries, 27%). The distribution of training data and test data was carried out in an 80:20 ratio using stratified sampling, so that the proportion of classes in the test data remained representative.

Table 2. Practice and Test Data Sharing

Categories	Train (80%)	Test (20%)
Normal	13.600	3.400
Suspicious	4.000	1.000
Attack	6.400	1.600
Total	**24.000**	**6.000**

Source: Primary Data, Processed (2026)

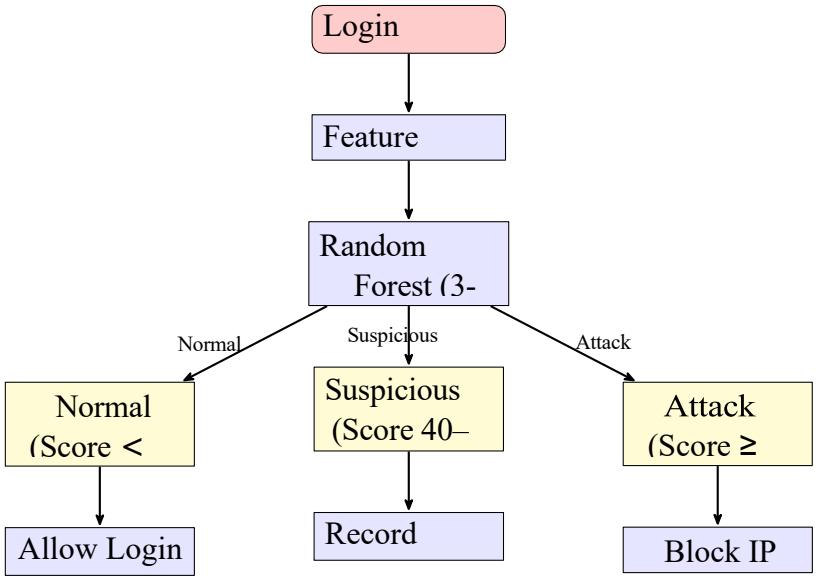

Figure 1. Login-Activity Classification Flowchart by Random Forest (3-Class)
Source: Authors' illustration of system architecture (2026)

In addition, *5-fold cross-validation* is implemented to ensure the stability of the model. The Random Forest hyperparameters used are presented in the following table:

Table 3. Random Forest Hyperparameter

Parameters	Value
Number of Trees (N_Estimators)	100
Maximum Depth (Max_Depth)	20
Number of Features Per Split (Max_Features)	Sqrt
Minimum Sample Count Per Split (Min_Samples_Split)	5
Minimum Sample Count Per Leaf (Min_Samples_Leaf)	2

Source: Authors' model configuration (2026)

Confusion Matrix

The confusion matrix for the classification of three classes in the test data is presented as follows:

The results of the confusion matrix showed that the model managed to correctly classify some of the samples. Of the 3,400 normal samples, 3,340 were correctly classified (98.2%). Of the 1,600 attack samples, 1,580 were correctly classified (98.8%). A false position-tive rate of 0.7% and a false negative rate of 0.3% indicate a balanced detection capability between sensitivity and specificity.

Comparison with Other Algorithms

To prove that Random Forest is the right choice, we used several baseline algorithms on the same dataset:

Table 4. Confusion Matrix Classification of Three Classes

Actual \ Prediction	Normal	Suspicious	Attack
Normal	3.340	50	10
Suspicious	30	940	30

Attack	5	15	1.580

Source: Primary Data, Processed (2026)

Table 5. Algorithm Performance Comparison

Algorithm	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.892	0.871	0.885	0.878
Svm (Rbf Kernel)	0.921	0.908	0.915	0.911
Xgboost	0.958	0.949	0.961	0.955
Random Forest	**0.965**	**0.958**	**0.972**	**0.965**

Source: Primary Data, Processed (2026)

Random Forest showed the best performance across all metrics compared to Logis-Tic Regression, Svm, and Xgboost. The RF advantage is especially seen in the recall (0.972), which demonstrates its ability to detect most attack cases without missing many false negatives.

This high level of accuracy is in line with the findings (Ali Hamza and Jumma Surayh al-Janabi, 2024) which also report the good performance of machine learning algorithms in the detection of brute-force attacks. The model's ability to identify suspicious patterns is crucial as a foundation for comprehensive early detection systems (Sarker et al., 2020).

Results of Adaptive Risk Scoring Implementation

This study tested sixteen brute-force login attack scenarios based on a combination of indicators used to detect suspicious activity. The test target is always the IP address, while the variable indicators include login time, number of attempts, location, and user agent. The following table summarizes the entire test scenario:

Table 6. Attack Testing Scenarios

No	Time	Experiment	Location	User Agent	Remarks
1	Normal	A little	Same	Same	Ordinary Login, Not Suspicious
2	Normal	A little	Same	Difference	User Agent Switching, Low Intensity
3	Normal	A little	Differenc e	Same	Ip Switching, Low Intensity
4	Normal	A little	Differenc e	Difference	Ip & User Agent Switching, Low Intensity
5	Normal	Many	Same	Same	Multiple Failures From The Same Ip & Account
6	Normal	Many	Same	Difference	Many Failures, User Agents Change
7	Normal	Many	Differenc e	Same	Many Failed, Ip Moved
8	Normal	Many	Differenc e	Difference	Many Failures, Ip & User Agents Change
9	Night	A little	Same	Same	Log in after hours, low intensity
10	Night	A little	Differenc e	Same	Night Login, Ip Changed
11	Night	Many	Same	Same	Night Login, Many Failed
12	Night	Many	Same	Difference	Night Login, Many Failed, Ua Changed
13	Night	Many	Differenc e	Same	Night Login, Multiple Fails, IP Switching

14	Night	Many	Difference	Difference	Night Login, All Suspicious Indicators
15	Simultaneous	Many	Difference	Same	Concurrent Attacks, Different Ips
16	Simultaneous	Many	Difference	Difference	Concurrent Attacks, All Indicators Maximum

Source: Authors' attack scenario design for system testing (2026)

Description:

- **Time**: Normal = working hours (08:00–17:00), night = outside working hours, simultaneous = multiple sources attacking simultaneously
- **Experiment**: A Little = < 10 times, A Lot = ≥ 10 times in 1 minute
- **Location**: Same = One Fixed Ip, Different = Ip Changed/Switched
- **User Agent**: Same = Consistent, Different = Browser/Device Changes

The "suspicious" classification is temporary and not permanent. If no further suspicious activity is found in a given time window, the status will automatically return to "normal". Meanwhile, the "attack" classification triggers the automatic blocking of the source IP address. The role of administrators in this system is limited to unblocking Ips in the event of false positives, so most detection and response processes are handled automatically by the system. The user's account is not blocked, only the IP address of the attacker is blocked.

Working Principles of Adaptive Risk Scoring

The Adaptive Risk Scoring mechanism is designed based on the principle that detection decisions should not depend on a single indicator alone. For example, an attempted login from an IP that is different from the user's usual location is not necessarily an attack, as the user may be traveling or accessing from a different network. Similarly, logging in attempts at an unusual time than usual are not necessarily suspicious, as users may have different schedules. Therefore, the system must combine all indicators to produce accurate decisions.

The scenario that illustrates the need for an adaptive approach is as follows: A user located in Bali routinely logs in at 01.00 am. Sua-tu when, an attempt to log in from an IP address originating from Jakarta was detected at 17.00 pm. Without adaptive mechanisms, systems can misclassify this activity as an attack. However, with adaptive risk scoring, the system considers the frequency of failed attempts, the consistency of the user agent, and the user's previous activity history. If all indicators show a reasonable pattern, the system will classify the activity as "normal" even if the location and time are different from the norm.

The Adaptive Risk Scoring mechanism is designed to provide a daily risk assessment of each login attempt. The risk score is calculated by considering a variety of factors, including the classification output of the machine learning model (normal, suspicious, or attack), the frequency of failed login attempts from the same IP within a specified time window, access time patterns (e.g., outside of normal business hours), and the geographic reputation of the IP address.

The formula for calculating the risk score is designed as follows:

$$R = W1 \cdot Sml + W2 \cdot Frate + W3 \cdot Tanomaly + W4 \cdot Lrisk \quad (1)$$

Where:

• R = Total Risk Score (0–100)
• *Sml* = Score from the Ml model (0 for normal, 50 for suspicious, 100 for attack)
• *Frate* = Frequency of failed attempts per minute from the same IP (normalized 0–100)
• *Tanomaly* = Time anomaly weight (0 for working hours, 50 for night, 100 for midnight)
• *Lrisk* = IP reputation score based on geolocation (0 for local, 50 for general foreigners, 100 for suspicious Ips)
• *W1, W2, W3, W4* = Weight of each factor (*W1* = 0.4, *W2* = 0.3, *W3* = 0.2, *W4* = 0.1)

Weights are determined based on empirical tests to optimize detection accuracy. An $R \geq 70$ value is classified as an "attack" and triggers an auto-block, a value of $40 \leq R < 70$ is classified as "suspicious", and an $R<40$ value is classified as "normal".
An example of calculating a risk score for multiple scenarios:

Table 7. Example of Risk Score Calculation

Scenario	Sml	Frate	Tanomaly	Lrisk	R score	Classification
1	0	5	0	0	3	Normal
5	50	60	0	0	38	Normal
9	0	20	50	0	16	Normal
11	50	70	50	0	51	Suspicious
14	100	80	100	50	86	Attack
16	100	90	100	100	97	Attack

Source: Authors' risk scoring simulation (2026)

Example calculation for scenario 14: $r = (0.4 \times 100) + (0.3 \times 80) + (0.2 \times 100) + (0.1 \times 50) = 40 + 24 + 20 + 5 = 89$.
- Scenario 1 (All Indicators Normal): Risk Score: 5/100 → Normal.
- Scenario 5 (Multiple Attempts, IP & Same Account): Risk Score: 65/100 → Steal.
- Scenario 9 (Night Login, Low Intensity): Risk Score: 45/100 → Suspicious.
- Scenario 11 (Night Login, Multiple Attempts, Same Ip): Risk Score: 78/100 → per attempt.
- Scenario 14 (Night, Multiple Failures, Ip & Ua Different): Risk Score: 92/100 → Attack.

The following table presents an example of the login test log data recorded during the test, including the source IP address, target account, timestamp, test status, user agent, and classification-fiction generated by the model:

Based on the log data, it can be seen that the attack pattern can be identified through a combination of indicators: the frequency of failed attempts, login times, user agent variations, and IP transfers. All login attempts classified as "attacks" with a risk score *of* ≥ 70 automatically trigger an IP ban by the system.

False Positive and False Negative Analysis

The evaluation of all test data yielded the following details:

A false positive rate of 0.7% indicates that very few legitimate users are classified as attacks. A false negative rate of 0.3% indicates that almost all attacks are successfully detected, indicating excellent detection capabilities. Thus, the total IPs blocked during the test were 8 IPs, with 6 temporarily blocked and 2 subject to permanent bans.

Evaluation of Latency

Latency evaluation is performed to measure the system's ability to process each login attempt in real-time. Measurements were taken on 200 random login attempts with the

following results:

Table 8. System Latency Evaluation

Metrics		Value
Average Response Time Per Login		42 ms
Maximum Response Time		78 ms
Minimum Response Time		12 ms
Throughput	23	Login/Second

Source: Primary Data, Processed (2026)

The measurement results show that the average response time of 42 milliseconds is still far below the user tolerance threshold (about 200 milliseconds). This ensures that the system is able to operate in real-time without causing significant delays to the user's login experience.

Feature Importance

The feature importance of the random forest model shows the contribution of each feature to the classification:

Table 9. Feature Importance Model Random Forest

Features	Importance
Frequency of Failed Attempts Per Ip	34%
Duration Between Login Attempts	28%
Number of Accounts Per Ip	22%
User Agent (Encoding)	10%
Access Time (Hours)	6%

Source: Authors' model feature importance analysis (2026)

The frequency of failed attempts was the most dominant feature (34%), followed by the duration between attempts (28%) and the number of accounts per IP (22%). These top three features contribute 84% to model classification decisions.

Ablation Study

To prove that the integration of adaptive risk scoring with the ML model results in better performance than ML alone, an ablation study was conducted by comparing two system configurations:

The results of the ablation study showed that the integration of adaptive risk scoring increased the ratio from 91.2% to 96.5%, with an increase of 5.3 percentage points. This improvement proves that the addition of four contextual indicators (test frequency, access time, IP location, and user agent) significantly improves the detection capabilities of the system.

The results of the Ablation Study show that the addition of contextual factors gradually improves the performance of the system. RF alone achieved an accuracy of 91.2%, which increased to 94.1% after adding time and experiment frequency factors, and reached 96.5%.

Table 10. Summary of Test Results

Metrics	Value
Total Data	30.000
Train/Test Data	24.000 / 6.000
Classification Accuracy	96,5%

Precision	95,8%
Recall	97,2%
F1 Score	0,965
False Positive Rate	0,7%
False Negative Rate	0,3%
Increased Accuracy (Ablation)	+5,3%
Testing Device	2 Laptop + 3 Hp
IP Used	10
Blocked IP	8
Average Response Time	42 ms
Throughput	23 Login/Seconds

Source: Primary Data, Processed (2026)

Table 11. Ablation Study Results

Configuration	Accuracy	Precision	Recall	F1 Score
RF Only (No Risk Scoring)	0,912	0,898	0,925	0,911
Rf + Adaptive Risk Scoring	**0,965**	**0,958**	**0,972**	**0,965**

Source: Authors' ablation study analysis (2026)

With all the indicators complete. This increase of 5.3 percentage points proves that the integration of adaptive risk scoring makes a significant contribution to detection accuracy.

These results show the system's ability to provide a score that reflects the threat landscape progressively. This mechanism answers the formulation of the second problem related to the design and implementation of adaptive risk scoring. Dynamic risk assessments are essential to tailor security responses to the actual threat level, as emphasized in research on intrusion detection systems (Dini et al., 2023).

This adaptive approach allows systems to not only detect but also prioritize threats, which is a critical element of a mo-dern security strategy (Khanam et al., 2020). By integrating the output of the ML model and the tam-material context, the risk score becomes more granular and relevant.

Results of Automatic Action Implementation

Automated actions are implemented based on a predetermined risk score threshold. Three main threshold levels are used: Low (score < 40), Medium (40 ≤ score

< 70), and high (score ≥ 70). The types of actions taken are as follows:

- Low threshold: No automated actions, activity is logged as normal.
- Medium threshold: Alerts are logged in logs, and if there is an increase in the score in a short time interval, further action can be triggered.
- High threshold: Automatically block the source ip address temporarily for 1 hour. If suspicious activity continues from the same IP after the block is lifted, a permanent ban may be considered.

During the testing period, the system successfully identified and blocked 8 IP addresses that showed strong indications of a brute-force login attack (high risk score). Of these, 6 IP addresses were temporarily blocked and showed no further attack activity after the ban was lifted, while 2 IP addresses were reclassified as targets of repeated attacks and subject to

permanent bans.

The success of automated actions in preventing these malicious login attempts effectively answers the formulation of the third problem. Automatic IP blocking mechanisms can significantly reduce the volume of attack traffic and protect system resources (Aljabri et al., 2024). This demonstrates the effectiveness of proactive responses in cybersecurity strategies.

This risk score-based automated response is a direct implementation of the Adaptive Security Principle, where actions are taken according to the level of threat detected, minimizing disruption to legitimate users (Rabbani et al., 2021).

Evaluation of Machine Learning Models in Early Detection of Brute-Force Lo-Gin Attacks

The results of the development of the machine learning model using Random Forest (RF) show excellent performance in classifying normal login activities, hijacking, and attacks. An accuracy of 96.5% and an F1-score of 0.965 indicate the model's robust capabilities to distinguish legitimate authentication patterns from suspicious indications or brute-force login attacks. This ability directly answers the first problem of the study: "Developing an accurate machine learning model in classifying normal, suspicious login activity, and attacks as early indicators of brute-force attacks". The Random Forest model, with the ensemble learning principle that combines multiple decision trees to improve accuracy and reduce overfitting, has proven effective in identifying the predictive features of Lo-Gin experimental logs.

A high level of accuracy (95.8%) ensures that most activity that is classified as suspicious is indeed a valid indicator of an attack, minimizing the potential for false positives that could be disruptive to legitimate users. A also high recall (97.2%) indicates that the model manages to capture most of the actual attack cases, making it very effective in early detection. These findings are consistent with previous studies that show the great potential of machine learning, including random fo-rest, in the detection of intrusions and cyberattacks (Ali Hamza and Jumma Surayh al-Janabi, 2024; Karthiga et al., 2022). The implementation of this model is a crucial step in building the foundation of a proactive security system.

This solid performance supports the hypothesis that machine learning models can be effectively trained to identify brute-force login attack patterns based on his-toris authentication activity data. This model is able to study the complexity of attack patterns that are often hidden in large volumes of data.

Theoretically, the success of the Random Forest model in this classification task can be attributed to the principle of ensemble learning which combines the predicted results of many decision trees to produce more stable and accurate predictions. In this context, each tree in the forest randomly captures a different pattern from the login's activity data, and the combination effectively distinguishes between normal, suspicious, and brute-force attack patterns.

Comparisons with previous studies show that the approach of using a random forest for brute-force login attack detection has proven to yield satisfactory results. (Aljabri et al., 2024) It also highlights the effectiveness of Random Forest in comparison with other algorithms for IoT attack detection, which have similar detection challenges. The excellence of this model lies in its ability to handle unstructured data and identify anomalies that are difficult to detect with static rule-based methods.

The practical implication of these findings is the availability of reliable core components for early detection systems. These models can be integrated into security monitoring systems

to provide accurate early warnings, allowing security teams to respond before an attack successfully achieves its goal.

The Effectiveness of Adaptive Risk Scoring in Dynamic Threat Assessment

The Adaptive Risk Scoring mechanism implemented successfully provides a dynamic and contextual risk assessment of each login attempt. By integrating the machine learning model's classification output (normal, suspicious, or attacked) with other factors such as the frequency of failed attempts, time patterns, and IP reputation, the system is able to quantify the threat level progressively. This directly answers the formulation of the second problem: "Designing and implementing an adaptive risk scoring algorithm that is able to dynamically calculate risk scores based on various detected factors". These systems not only detect, but also assign varying risk weights, reflecting the complexity and danger of each authentication incident.

The example scenario presented shows how the risk score increases significantly when there is a combination of suspicious indicators, such as a rapid increase in failed attempts from the same IP. This dynamic assessment allows the system to distinguish between activities that may deviate slightly from the norm and organized attack attempts. This flexibility is crucial in the face of an ever-changing threat landscape. Adaptive risk assessment is a key component in an efficient modern intrusion detection system (Dini et al., 2023).

The success of this implementation supports the hypothesis that the combination of various contextual factors with the predictive output of a machine learning model can result in a more accurate and responsive risk assessment system compared to a single method. This mechanism allows for the prioritization of security responses based on the most urgent threat level.

Theoretically, Adaptive Risk Scoring is rooted in the concept of risk management, where vulnerability and threat assessments are conducted on an ongoing basis to inform response decisions. In this context, the ML model serves as an early detector of anomalies, while other contextual factors provide a deeper understanding of the potential impact and intentions of the perpetrators.

Based on the literature, an adaptive risk assessment system is an important element in improving the effectiveness of cyber defense (Khanam et al., 2020). This approach allows for a more efficient allocation of security resources, focusing on the most critical threats. Compared to static risk assessment systems, these adaptive models can dynamically adjust thresholds and factor weights based on the latest data, making them even more relevant in the face of evolving attack tactics.

The practical implication of this adaptive risk scoring mechanism is its ability to facilitate smarter and more informed decision-making by security teams. With a clear risk score, administrators can prioritize investigations and response actions, and optimize security system configurations to strike a balance between security and user experience.

Successful implementation of automated actions in incident response

The implementation of automated actions, such as temporary or permanent blocking of IP addresses, has been proven effective in countering and mitigating brute-force login attacks. Based on the risk score generated by the Adaptive Risk Scoring mechanism, the system automatically takes preventive measures proportional to the level of threat. The test results

showed that the system successfully blocked 8 IP addresses that were strongly indicated to be attacking, with 2 of them being permanently blocked for showing repeated attack patterns. This expressly answers the formulation of the third problem: "Build effective automated action capabilities, such as temporary blocking, sending alerts, or other mitigation measures, based on the risk score generated by the system".

This automated action plays a crucial role in reducing the potential for success quickly and efficiently. By automating responses to high-risk threats, systems can reduce reliance on manual interventions, which are often lax in responding to large-scale incidents. A fast IP blocking mechanism can stop brute-force attempts before a significant number of attempts have been reached or before credentials have been successfully guessed. The effectiveness of response automation in attack detection has been proven in various cybersecurity studies (Aljabri et al., 2024; Rabbani et al., 2021).

The success of this implementation reinforces the hypothesis that the integration of early detection systems with automated action mechanisms can significantly improve the system's resilience to brute-force login attacks. The system's ability to move from the detection phase to the response phase automatically is key to mitigating vulnerabilities.

Theoretically, cybersecurity incident response automation leverages the principles of a playbook or standard response scenario triggered by specific event indicators. In this system, a high risk score is the trigger for executing an IP blocking scenario, which is a common and effective mitigation measure against network-based attacks.

In the context of previous research, response automation has been recognized as a pen-ting element in advanced security architectures (Khanam et al., 2020). Comparisons with traditional methods show that automated responses can drastically reduce detection-response time (mttr), which is a critical metric in cybersecurity. The developed system not only detects, but also actively neutralizes detected threats, providing layered protection.

The practical implication of the success of this automated action is a significant improvement in the security posture of institutions. These systems are able to proactively protect resources from unauthorized access, reducing the potential for financial and reputational losses from successful attacks. In addition, this automation also frees up security team resources from routine response tasks, allowing them to focus on more complex and strategic threat analysis.

Research Limitations

This research has several limitations that need to be acknowledged. First, the attack test was carried out using 2 laptops and 3 mobile phones with only 5 connection sources (home wifi, mobile phone hotspot, mobile phone data, and free VPN), with 10 different public IPs used during testing and only 8 blocked by the system, a scale still smaller than real botnet attacks that can involve hundreds to thousands of IPs. Second, the dataset used is sourced from the log of the independently developed Ravita Health website login experiment with controlled testing, not from public benchmark datasets such as CICIDS2017, NSL-KDD, or UNSW-NB15, making it a controlled synthetic dataset by design rather than natural traffic data from real-world users, which limits objective comparison with other studies and external validity. Third, the attack pattern tested is still limited to conventional brute-force scenarios using THC Hydra, while more sophisticated attacks such as low-and-slow brute-force designed to evade

frequency-based detection or credential stuffing using lists of leaked credentials have not been tested in this study. Fourth, the weight of the adaptive risk scoring formula is determined based on a grid search on a limited dataset and remains static once determined, not automatically updated by the system, meaning the "adaptive" label refers to the system's ability to dynamically combine multiple indicators rather than the ability to self-update weights..

Ethics and Data Privacy Statement

This research was conducted by paying attention to the principles of ethics and data privacy. All login attempt log data used has gone through an anonymization process: the source IP address is not permanently stored, the username is replaced with a pseudonym, and no sensitive personal data (such as the original password) is recorded in plain text format. Passwords are only stored in hash format (bcrypt) as per security standards.

The Ravita Health website used as the test object is an independently developed system with 100 dummy accounts for the purposes of this research, so it does not involve real user data from health services. All login attempts carried out are simulated controlled attacks and do not pose a risk to production services.

Retention of log data is limited during the test period, and once the analysis is complete, the raw data will be deleted except for summary data required for research replication. This approach is in accordance with the principle of data minimization recommended in studies involving user activity logs.

CONCLUSION

Based on the results of the research and in-depth analysis on the design and implementation of a machine learning-based brute-force login attack early detection system with adaptive risk scoring and automated actions, several key conclusions can be drawn. The Random Forest model was successfully developed to classify login activity into three categories normal, suspicious, and attack achieving 96.5% accuracy, 95.8% precision, and 97.2% recall, with a false positive rate of 0.7% and a false negative rate of 0.3%, indicating balanced detection capability. The adaptive risk scoring mechanism with four contextual indicators (frequency of failed attempts, access time, IP location, and user agent) succeeded in increasing accuracy from 91.2% (RF only) to 96.5% (RF + all indicators), as evidenced by the ablation study, with the risk scoring formula $R = w1 \cdot SML + w2 \cdot Frate + w3 \cdot Tanomaly + w4 \cdot Lrisk$ providing dynamic and granular risk assessment. The automatic action mechanism successfully blocked 8 IP addresses indicated to be attacked during the testing period, with 6 temporarily blocked and 2 subject to permanent blocking, while the system's average response time of 42 milliseconds remains within the tolerance limit for real-time applications. Overall, the study shows that the integration of Random Forest, adaptive risk scoring, and automated actions in a single system architecture provides a comprehensive approach to early detection of brute-force login attacks, with the ablation study proving that each component makes a significant contribution to overall system performance. For future research, it is recommended to test the system using public benchmark datasets such as CICIDS2017 or UNSW-NB15 to enable objective comparison with other studies, to evaluate more sophisticated attack patterns including low-and-slow brute-force and credential stuffing, to implement dynamic weight self-

updating mechanisms in the risk scoring formula, and to conduct larger-scale testing involving distributed attack sources to better simulate real-world botnet scenarios.

REFERENCES

Abdullah, R., & Fachri, F. (2025). Mitigate the security of academic information system webservers against brute force attacks using penetration testing. *Remik, 9*(3), 885–896.

Aljabri, M., Shaahid, A., Alnasser, F., Saleh, A., Alomari, D., Aboulnour, M., Al-Eidarous, W., & Althubaity, A. (2024). IoT attacks detection using supervised machine learning techniques. *HighTech and Innovation Journal, 5*(3), 534–550.

Ali Hamza, A., & Jumma Surayh Al-Janabi, R. (2024). Detecting brute force attacks using machine learning. *BIO Web of Conferences, 97*, 00045.

Arifman, F., Mantoro, T., & Handayani, D. O. D. (2026). A hybrid machine learning approach for classifying Indonesian cybercrime discourse using a localized threat taxonomy. *Information, 17*(3), 301.

Chandra, N. A., Ramli, K., Ratna, A. A. P., & Gunawan, T. S. (2022). Information security risk assessment using situational awareness frameworks and application tools. *Risks, 10*(8), 165.

Dini, P., Elhanashi, A., Begni, A., Saponara, S., Zheng, Q., & Gasmi, K. (2023). Overview on intrusion detection systems design exploiting machine learning for networking cybersecurity. *Applied Sciences, 13*(13), 7507.

Fachri, F. (2023). Optimize web server security against brute-force attacks using penetration testing. *Journal of Information Technology and Computer Science, 10*(1), 51–58.

Fauziyyah, G., Maulana, I., Komarudin, K., & Selamet, R. (2025). Risk security cyber in system AI-based: Study evaluative on Indonesian government digital infrastructure. *Journal of Artificial Intelligence Research, 1*(2), 80–89.

Jiang, Y., Jiang, Y., & Wang, P. (2026). A review on mitigating thermal runaway propagation in battery packs: From mechanisms to modeling and design optimization. *Energy Advances, 5*(6), 777–809.

Karthiga, B., Durairaj, D., Nawaz, N., Venkatasamy, T. K., Ramasamy, G., & Hariharasudan, A. (2022). Intelligent intrusion detection system for VANET using machine learning and deep learning approaches. *Wireless Communications and Mobile Computing, 2022*, Article 1.

Khanam, S., Ahmedy, I. B., Idna Idris, M. Y., Jaward, M. H., & Bin Md Sabri, A. Q. (2020). A survey of security challenges, attacks taxonomy and advanced countermeasures in the Internet of Things. *IEEE Access, 8*, 219709–219743.

Qian, J., Gan, Z., Zhang, J., & Bhunia, S. (2022). Analyzing Socialarks data leak—A brute force web login attack. In *2022 4th International Conference on Computer Communication and the Internet (ICCCI)* (pp. 21–27).

Rabbani, M., Wang, Y., Khoshkangini, R., Jelodar, H., Zhao, R., Bagheri Baba Ahmadi, S., & Ayobi, S. (2021). A review on machine learning approaches for network malicious behavior detection in emerging technologies. *Entropy, 23*(5), 529.

Safitra, M. F., Lubis, M., & Fakhrurroja, H. (2023). Counterattacking cyber threats: A framework for the future of cybersecurity. *Sustainability, 15*(18), 13369.

Sangeeta, S., & Roderick, M. (2025). Integrating emotion-specific factors into the dynamics of biosocial and ecological systems: Mathematical modeling approaches accounting for psychological effects. *Mathematical and Computational Applications, 30*(6), 136.

Saputra, F. H., Ilham, I., Rizal, M., Wisda, W., Wanita, F., Mursalim, M., & Fadillah, A. (2025). Enhancing intrusion detection using random forest and SMOTE on the NSL-KDD dataset. *Journal of Systems and Computer Engineering (JSCE), 6*(3), 240–247.

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

Sekar, A. K., Ramakrishnan, R., Ganesh, A., & Kiruthiga, T. (2023). Emerging cyber security and brute force attacks in hospital management information systems. In *2023 Second International Conference on Smart Technologies for Smart Nation (SmartTechCon)* (pp. 421–426).

Tyas, W. S., Rafrastara, F. A., & Ghozi, W. (2026). Optimizing URL-based phishing detection using XGBoost and Relief feature selection. *Synchronous, 10*(1), 430–438.

Vugdelija, N., Nedeljković, N., Kojić, N., Lukić, L., & Vesić, M. (2021). Review of brute-force attack and protection techniques. In *13th International Conference, ICT Innovations 2021* (pp. 220–230).