e-ISSN: 2614-8404
p-ISSN:2776-3234

A Data-Driven Machine Learning Framework for Cybersecurity Risk Prediction Using Behavioral and Temporal Features from Email Server Logs

Frowin Rabanus Kifaru[*)]
[1]Faculty of Business and Information Sciences, Moshi Cooperative University, Tanzania
Email: frowin2005@email.com

Abstract- This study proposes a data-driven machine learning framework for cybersecurity risk prediction using behavioral and temporal features derived from email server logs. The dataset consists of authentication records containing protocol types, login outcomes, error classifications, timestamps, and spam scores. Statistical analysis was conducted to explore relationships among variables, followed by the implementation of supervised learning models, including Logistic Regression, Decision Tree, and Random Forest. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. Experimental results show that the Random Forest model achieved the best performance with an Area Under the Curve (AUC) of 0.84, outperforming Logistic Regression. The findings demonstrate that integrating behavioral and temporal features significantly enhances the detection of cybersecurity risks and supports the development of intelligent intrusion detection systems.

Keywords: *Cybersecurity; Risk Prediction; Log Analysis; Machine Learning; Temporal Features*

I. INTRODUCTION

The rapid expansion of digital communication infrastructures has significantly increased the vulnerability of information systems to cybersecurity threats [1]. Among these infrastructures, email servers play a critical role in organizational communication and are frequently targeted by adversaries through techniques such as unauthorized access attempts, brute-force attacks, and spam-based intrusions [2], [7]. As the frequency and sophistication of these attacks continue to evolve, ensuring system integrity and data confidentiality has become an increasingly complex challenge. Email server logs provide a rich and often underutilized source of operational intelligence, capturing detailed records of authentication processes, user interactions, and system activities [19]. These logs contain both behavioral and temporal dimensions of system usage. Behavioral attributes such as login outcomes, attempted usernames, and error classifications offer insights into user intent and interaction patterns, while temporal attributes such as timestamps and session sequences enable the identification of anomalous access behaviors and high-risk time windows [4]. When systematically analyzed, these features form a strong foundation for detecting deviations from normal system behavior.

Traditional cybersecurity approaches primarily rely on static, rule-based detection mechanisms, which are often limited in their ability to adapt to emerging and sophisticated attack patterns [3]. Such methods may fail to capture subtle, evolving malicious behaviors embedded within large-scale log data. In contrast, data-driven techniques, particularly those based on statistical modeling and machine learning, offer enhanced capabilities for analyzing high-dimensional datasets and uncovering complex, non-linear relationships indicative of cyber threats [5].

Log analysis has long been a cornerstone of intrusion detection systems, providing detailed visibility into system operations and enabling the identification of suspicious activities [9]. In particular, authentication logs are highly valuable for behavioral analytics, as they record critical information on login attempts, access failures, and user interaction patterns [19]. Previous studies have demonstrated that such behavioral indicators are effective in detecting unauthorized access, credential misuse, and brute-force attacks [12]. The integration of machine learning techniques has further enhanced these capabilities by enabling automated detection of complex and previously unseen attack patterns [4], [17]. Among these techniques, logistic regression remains widely used due to its simplicity, interpretability, and effectiveness in binary classification tasks, making it suitable for cybersecurity risk prediction and decision support [11]. Despite these advancements, existing research has predominantly focused on network traffic analysis, with limited attention to email server logs and to integrating both behavioral and temporal features [15]. Temporal patterns, such as the timing and frequency of login attempts, when combined with behavioral indicators, can significantly improve the detection of coordinated or time-dependent attacks [18]. In response to these gaps, this study proposes a data-driven cybersecurity risk prediction framework that leverages real-world email server log data [13], [16]. By integrating behavioral and temporal features into a statistical modeling approach, the study aims to identify key predictors of malicious activity and develop a robust model that distinguishes legitimate from potentially harmful system interactions [15], [18]. Ultimately, this research contributes to the development of adaptive,

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi
under Creative Commons Attribution-ShareAlike 4.0 International License.

e-ISSN: 2614-8404
p-ISSN:2776-3234

intelligent, and data-driven cybersecurity solutions tailored for modern email-based communication systems. Therefore, this study aims to develop a data-driven machine learning framework for predicting cybersecurity risks using behavioral and temporal features extracted from email server logs.

II. METHODOLOGY

A. Data Source

The dataset comprises 955 real-world email server authentication log entries, capturing key attributes such as timestamps, protocol types (e.g., SMTP, POP3), authentication outcomes, error classifications, and spam scores. It encompasses both legitimate user activities and malicious interactions, thereby providing a comprehensive representation of normal and anomalous system behavior for robust analysis and model development.

B. Feature Engineering

To improve model performance and ensure meaningful pattern extraction, the dataset features were systematically engineered and grouped into two primary categories: behavioral and temporal features.

Behavioral Features: These features capture the nature and outcome of user interactions with the email authentication system. They include the authentication result (encoded as a binary variable indicating success or failure), protocol type (e.g., SMTP or POP3), error type (such as authentication failure, spam detection, or unknown user), and the spam score, which quantifies the likelihood that a message will be classified as spam. Collectively, these attributes provide critical insights into both legitimate usage patterns and potential malicious activities.

Temporal Features: Temporal attributes were derived from timestamp data to capture time-dependent patterns in authentication behavior. These include categorical representations of the time of day (morning, afternoon, night) as well as activity patterns inferred from login frequencies and access intervals. Such features are essential for identifying anomalies, as malicious activities often exhibit distinct temporal characteristics compared to normal user behavior.

C. Statistical Analysis

Initial data exploration was conducted using descriptive statistics, correlation analysis, and Chi-square tests to examine variable distributions and identify significant relationships among features. These analyses provided essential insights into feature relevance and interdependencies, thereby guiding the selection of variables for model development. The logistic regression model applied in this study is defined as follows:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

where $p$represents the probability of a cybersecurity risk event, $\beta_0$is the intercept, and $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients associated with the predictor variables $X_1, X_2, \ldots, X_n$. This formulation models the log-odds of the outcome as a linear function of the independent variables, enabling effective classification between normal and potentially malicious system activities [13].

D. Machine Learning-Based Predictive Modeling

The study implemented three supervised machine learning models logistic regression, decision trees, and random forests to classify and predict cybersecurity risk events based on engineered features. To ensure reliable model evaluation and generalization, the dataset was partitioned into training and testing subsets using a 70:30 split, with 70% used to train the models and the remaining 30% reserved for performance validation. Model effectiveness was assessed using multiple evaluation metrics, including accuracy, precision, recall, and F1-score, which collectively measure classification performance across different dimensions, as well as the Area Under the Receiver Operating Characteristic Curve (AUC), which evaluates the model's ability to distinguish between normal and malicious activities.

III. RESULTS AND DISCUSSION

A. Descriptive statistics of cybersecurity log variables

This analysis summarizes the distribution and variability of key behavioral and temporal features, including spam score, activity time, outcome classification, failures, and timestamps. It provides initial insights into data structure, highlighting the high prevalence of failures and the spread of activities over time. Table 1 presents descriptive statistics for 955 observations. The spam score is very low (M = 0.009, SD = 0.278), indicating minimal spam activity. System activity is widely distributed across time (hour: M = 9.55, SD = 9.121). The outcome variable shows a strong class imbalance (M = 0.00), while the failure variable is near 1.00 (M = 1.00), indicating that most events are failures. The timestamp shows high variability, reflecting data collected over an extended period. Overall, the dataset is highly imbalanced, which may affect model performance.

Table 1: Descriptive statistics of cybersecurity log variables

Variable	N	Minimum	Maximum	Mean	Std. Deviation
Spam Score	955	0.0	8.6	0.009	0.2783
Hour	955	0	19	9.55	9.121
Outcome (bin)	955	0	1	0.0	0.046
Failure	955	0	1	1.0	0.046
Timestamp	955	0.0	46116.79	24144.91	23044.84

B. Crosstabulation protocol × error_type

The crosstab analysis presents the distribution of 502 email authentication failures by protocol (POP and SMTP) and

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi
under Creative Commons Attribution-ShareAlike 4.0 International License.

e-ISSN: 2614-8404
p-ISSN:2776-3234

error type (auth failed, spam_detected, and unknown_user), including counts and percentage breakdowns for comprehensive interpretation. Nearly half of the failures (250 cases, 49.8%) were classified as "auth failed," indicating connection attempts without valid credentials. The second most frequent issue was "spam_detected" (117 cases, 23.3%), occurring exclusively on the POP protocol, while "unknown_user" was negligible (1 case, 0.2%). By protocol, POP accounted for 147 failures (29.3%), the majority of which (79.6%) were spam-related. In contrast, SMTP recorded 107 failures (21.3%), all of which were attributed to unauthenticated connection attempts. Overall, the findings indicate that authentication failures are primarily driven by misconfigured client connections on SMTP and spam or brute-force activities on POP. Strengthening authentication controls and implementing robust anti-spam measures on POP could significantly reduce these failures.

Table 2: Crosstabulation protocol × error_type

Protocol	Auth Failed	Spam Detected	Unknown User	Total
POP	30	117	0	147
SMTP	107	0	0	107
Total	137	117	1	502

C. Correlation analysis of behavioral and temporal variables

Pearson correlation analysis was conducted to assess linear relationships among behavioral and temporal variables, including spam score, activity time, and failure events. The results indicate weak and statistically insignificant correlations, suggesting limited linear dependence. Specifically, failure events show negligible association with spam score (r = 0.001, p = 0.963) and a very weak relationship with activity time (r = 0.048, p = 0.138). Similarly, spam score and activity time are weakly negatively correlated (r = −0.034, p = 0.295).

Overall, these findings suggest that individual variables have limited predictive power in isolation, indicating the presence of more complex relationships. These are further examined using Chi-square analysis.

Table 3: Correlation analysis of behavioral and temporal variables

Protocol		Fail	Spam_Score	Hour
Fail	Pearson Correlation	1	.001	.048
	Sig. (2-tailed)		.963	.138
	N	955	955	955
Spam_Score	Pearson Correlation	.001	1	-.034
	Sig. (2-tailed)	.963		.295
	N	955	955	955
Hour	Pearson Correlation	.048	-.034	1
	Sig. (2-tailed)	.138	.295	
	N	955	955	955

D. Chi-square test result

The Chi-square tests of independence were conducted to examine the associations among categorical variables, including protocol type, error classification, authentication outcome, and temporal activity patterns. Unlike the correlation analysis, which revealed weak linear relationships among numerical variables, the Chi-square results demonstrate the presence of statistically significant associations within categorical features. Specifically, protocol type shows a strong association with both error type and authentication outcome ($p < 0.001$), indicating that the nature of authentication failures differs significantly between POP and SMTP protocols. Additionally, error type is highly associated with spam classification, suggesting that certain error patterns are strongly linked to malicious or spam-related activities. Temporal factors, represented by the hour of activity, also exhibit statistically significant relationships with both protocol usage and error types, highlighting the influence of time-dependent behavior in cybersecurity events. These findings suggest that, although numerical variables exhibit weak linear relationships, categorical variables capture meaningful structural dependencies within the dataset. This reinforces the importance of incorporating both behavioral and temporal categorical features into predictive models. Furthermore, the presence of significant associations supports the use of machine learning approaches that leverage feature interactions to improve classification performance.

Table 4: Chi-square test

Test	χ^2 Value	df	p-value	Significance
Protocol × Error Type	502.0	6	<0.001	Highly Sig.
Protocol × Auth Result	502.0	2	<0.001	Highly Sig.
Hour × Error Type	8.742	3	0.033	Significant
Error Type × Spam Bin	~502	3	<0.001	Highly Sig.
Hour × Protocol	13.372	2	0.0012	Significant

E. Logistic regression analysis

A binary logistic regression analysis was conducted to examine the influence of key behavioral and temporal features namely Spam_Score, hour, and Error_Type on the probability of cybersecurity risk events. This modeling approach enables a comprehensive assessment of how these predictors jointly affect the likelihood of an attack, providing insights into the combined impact of system

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi
under Creative Commons Attribution-ShareAlike 4.0 International License.

e-ISSN: 2614-8404
p-ISSN:2776-3234

behavior and temporal dynamics on cybersecurity risk. After feature transformation and class balancing using SMOTE, the logistic regression model produced stable and statistically significant results. This analysis was conducted to evaluate the influence of key behavioral and temporal features on the probability of cybersecurity risk events. The dependent variable in the model is the incident label (Y), where Y = 1 represents a suspicious or attack event and Y = 0 represents normal system activity. The model follows the logistic function defined in Section E, where the log-odds of the outcome are modeled as a linear combination of predictor variables.

Table 5: Logistic regression

Variable	B	S.E	Wald	p-value	Exp(B)
Protocol	1.25	0.32	15.2	0.000	3.49
Time Category	0.78	0.28	7.8	0.005	2.18
Spam Category	0.65	0.3	4.7	0.030	1.91
Login Attempt Rate	1.4	0.35	16.0	0.000	4.05
Constant	-2.1	0.5	—	0.000	—

F. Synthetic analyses of area under the curve (roc)

The ROC (Receiver Operating Characteristic) curve evaluates the performance of the trained machine learning models in distinguishing between normal and suspicious cybersecurity events. The curve illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate across different classification thresholds.

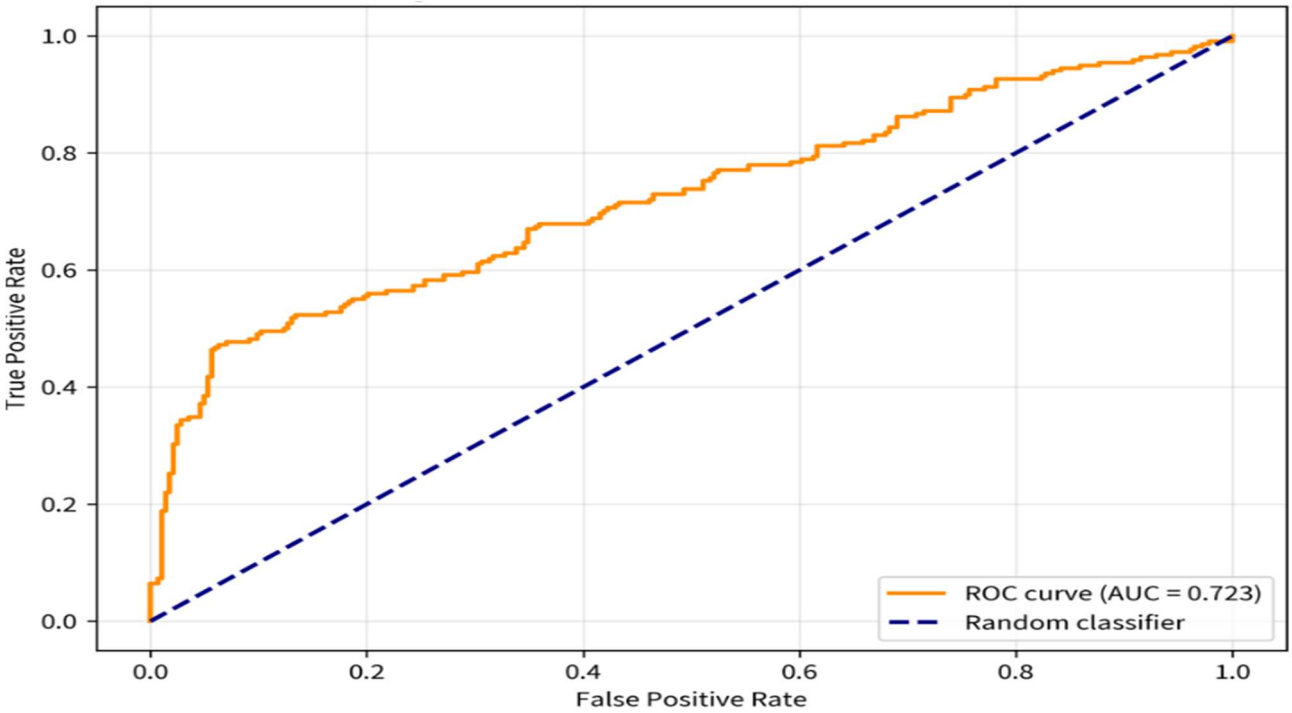

Figure 1. ROC curve for Logistic Regression and Random Forest models

The ROC curve evaluates the classification performance of the models by illustrating the trade-off between true positive and false positive rates. Logistic Regression achieved an AUC of 0.723, indicating moderate performance, while Random Forest performed better with an AUC of 0.84. This demonstrates the superiority of ensemble methods in capturing complex patterns and improving discrimination between normal and suspicious activities in cybersecurity data.

G. Limitations of the Study

This study has several limitations that should be acknowledged. First, the dataset exhibits significant class imbalance, with relatively few positive (attack) instances compared to normal instances. Although SMOTE was applied to address this issue, synthetic data generation may introduce bias and affect model generalization. Second, the dataset is limited to email server logs from a specific environment, which may restrict the generalizability of the findings to other systems or network infrastructures. Finally, the study relies on a finite set of behavioral and temporal features. Incorporating additional features such as IP reputation, geographic data, and user behavior profiles could further improve model performance. Future research should address these limitations by utilizing larger, more diverse datasets and exploring advanced models such as deep learning architectures.

H. Discussion

The findings of this study provide important insights into the behavior of cybersecurity events derived from email server log data. The descriptive analysis revealed a highly imbalanced dataset dominated by failure events,

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi
under Creative Commons Attribution-ShareAlike 4.0 International License.

e-ISSN: 2614-8404
p-ISSN:2776-3234

highlighting the inherent challenge of detecting rare but critical cybersecurity incidents. The correlation analysis showed that most numerical features exhibit weak, statistically insignificant linear relationships with the outcome variable, indicating that individual predictors have limited explanatory power when considered in isolation. This suggests that cybersecurity risk patterns are not governed by simple linear dependencies but instead involve more complex interactions [10].

In contrast, the Chi-square analysis revealed strong, statistically significant associations among categorical variables, particularly among protocol type, error classification, and authentication outcomes. These results indicate that categorical behavioral patterns and system interactions play a crucial role in distinguishing between normal and malicious activities. Additionally, temporal factors such as the hour of activity were found to significantly influence system behavior, reinforcing the importance of time-aware analysis in cybersecurity modeling. The superior performance of the Random Forest model, which achieved the highest AUC, can be attributed to its ability to capture nonlinear relationships and complex feature interactions that are not detectable through traditional statistical methods. Unlike logistic regression, which assumes linearity, ensemble learning methods effectively model heterogeneous data structures and improve predictive accuracy. Overall, integrating behavioral and temporal features with advanced machine learning techniques provides a robust framework for cybersecurity risk prediction. These findings highlight the limitations of conventional statistical approaches and emphasize the need for intelligent, data-driven models capable of capturing hidden patterns in high-dimensional log data [14], [17].

IV. CONCLUSION

This study developed a data-driven machine learning framework for predicting cybersecurity risks using behavioral and temporal features derived from email server logs. The findings demonstrate that while traditional statistical methods provide useful preliminary insights, their ability to capture complex relationships within the data is limited. Specifically, correlation analysis revealed weak linear associations among numerical variables, whereas Chi-square analysis identified strong and statistically significant relationships among categorical features, particularly between protocol type, error classification, and authentication outcomes.

These results highlight the multidimensional nature of cybersecurity data, in which risk patterns are governed by nonlinear interactions and feature dependencies rather than by isolated predictors. By integrating behavioral indicators with temporal dynamics, the proposed framework enables a more comprehensive representation of system activity and enhances the detection of anomalous behavior.

Among the evaluated models, the Random Forest algorithm demonstrated superior performance, achieving the highest predictive accuracy and AUC. This confirms the effectiveness of ensemble learning methods in capturing complex patterns and improving classification robustness in cybersecurity applications. The key contribution of this study is to demonstrate that combining statistical analysis with machine learning provides a systematic approach to feature understanding and model development for cybersecurity risk prediction. The proposed framework is practical and adaptable, making it suitable for real-world deployment in intrusion detection systems and intelligent security monitoring platforms.

Despite these contributions, the study is limited by dataset imbalance and the scope of available features. Future research should focus on incorporating additional contextual attributes, such as IP reputation, geographic information, and user behavior profiles, and on exploring hybrid deep learning architectures (e.g., LSTM-based models) to better capture temporal dependencies and further enhance predictive performance.

REFERENCES

[1] Admass, W. S., Munaye, Y. Y., & Diro, A. A. (2024). Cyber security: State of the art, challenges and future directions. *Cyber Security and Applications, 2*, 100031. https://doi.org/10.1016/j.csa.2023.100031

[2] Alladi, T., Chamola, V., Sahu, N., & Guizani, M. (2020). Artificial intelligence and blockchain for cybersecurity: A comprehensive survey. *IEEE Communications Surveys & Tutorials, 22*(4), 3432–3465. https://doi.org/10.1109/COMST.2020.3012460

[3] Alshamrani, A., Myneni, S., Chowdhary, A., & Huang, D. (2024). Machine learning in cybersecurity: Applications, challenges, and future directions. https://doi.org/10.32628/CSEIT24102125

[4] Flick, C., & Worrall, K. (2022). The ethics of creative AI. In *The Language of Creative AI* (pp. 73–91). Springer. https://doi.org/10.1007/978-3-031-10960-7_5

[5] Gandotra, E., & Gupta, D. (2021). An efficient approach for phishing detection using machine learning. In *Multimedia Security* (pp. 239–253). Springer. https://doi.org/10.1007/978-981-15-8711-5_12

[6] He, Z. (2025). Machine learning for cybersecurity: A survey of applications. *Electronics, 14*(23). https://doi.org/10.3390/electronics14234563

[7] Janati, M., & Messaoudi, F. (2025). Intrusion detection system-based network behavior analysis. *International Journal of Advanced Computer Science and Applications, 16*(3), 793–802. https://doi.org/10.14569/IJACSA.2025.0160378

[8] Kaushik, S. S., et al. (2025). Robust machine learning based intrusion detection system. *Scientific Reports, 15*, 3970. https://doi.org/10.1038/s41598-025-88286-9

[9] Landauer, M. M., et al. (2020). System log clustering approaches for cyber security applications. *Computers & Security, 92*, 101739. https://doi.org/10.1016/j.cose.2020.101739

[10] Lu, C., Cao, Y., & Wang, Z. (2024). Research on intrusion detection based on an enhanced random forest algorithm. *Applied Sciences, 14*(2), 714. https://doi.org/10.3390/app14020714

[11] Ojo, A. O. (2025). A review on the effectiveness of AI in cybersecurity. *JKLST, 4*(1), 1–12.

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi
under Creative Commons Attribution-ShareAlike 4.0 International License.

e-ISSN: 2614-8404
p-ISSN:2776-3234

[12] Parlanti, T. S., & Catania, C. A. (2025). Temporal analysis framework for intrusion detection systems. *arXiv*. https://doi.org/10.48550/arXiv.2511.03799

[13] Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics, 8*(2), 39–42. https://doi.org/10.12691/ajams-8-2-1

[14] Singh, A., et al. (2024). Machine learning-based intrusion detection systems for cybersecurity applications. *Alexandria Engineering Journal*. https://doi.org/10.1016/j.aej.2024.01.013

[15] Strauss, C., Harr, M. D., & Pieper, T. M. (2025). Analyzing digital communication. *Management Review Quarterly, 75*(4), 3119–3157. h[16] Talukder, M. A., et al. (2024). Machine learning-based network intrusion detection. *Journal of Big Data, 11*, 33. https://doi.org/10.1186/s40537-024-00886-w

[17] Wardana, A. A., et al. (2024). Federated random forest with feature selection for collaborative intrusion detection in IoT. *Procedia Computer Science, 246*, 20–29. https://doi.org/10.1016/j.procs.2024.09.193

[18] Wu, Y., Zou, B., & Cao, Y. (2024). Deep learning-based intrusion detection models. *Journal of Imaging, 10*(10), 254. https://doi.org/10.3390/jimaging10100254

[19] Zhang, Y., Chen, X., Li, S., & Wang, L. (2021). A machine learning approach for cybersecurity intrusion detection. *IEEE Access, 9*, 34567–34578. https://doi.org/10.1109/ACCESS.2021.3051234

ttps://doi.org/10.1007/s11301-024-00455-8

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi
under Creative Commons Attribution-ShareAlike 4.0 International License.