Research on the Security Assessment of Communication Data in Higher Vocational Education Management Platforms Based on Statistical Methods

Lian Hua
School of Environmental Art and Design, Wuxi Vocational Institute of Arts & Technology, Yixing 214200, Jiangsu, China
Corresponding author: Lian Hua (e-mail: bingling840917@163.com)

ABSTRACT This study develops a statistical framework to assess the security of communication data in higher vocational education management platforms (HVE-MPs), providing quantitative risk evaluation and anomaly detection across large-scale datasets. Methods including logistic regression, ARIMA-based time series analysis, Z-score and Grubbs' test anomaly detection, and cumulative sum (CUSUM) control charts are applied to over one million communication logs to identify high-risk channels and predict breach probability. The approach highlights key operational risk factors, such as peak usage periods and concentrated attack patterns, and enables proactive, data-driven mitigation strategies. The framework is particularly relevant for communication-intensive environments, including wireless networks and antenna-supported information transmission systems, where secure and reliable data exchange is critical. Experimental results demonstrate high anomaly detection accuracy (>95%), low false alarm rates, and effective identification of the top 5% of high-risk channels, supporting continuous security monitoring and adaptive defense strategies in complex digital platforms.

INDEX TERMS communication data security, anomaly detection, risk scoring, wireless networks

I. INTRODUCTION

A. BACKGROUND

RETAIL Higher Vocational Education Management Platforms (HVE-MPs) are emerging as essential components in the digital transformation of educational institutions, effectively centralizing the management of student information, course administration, and financial processes. Similarly, the textile industry's protection of proprietary polymer material formulas and high-precision fabric structure design data is key to securing its core technological advantage and continuous product differentiation. These platforms facilitate smooth communication between students, faculty, and administrative bodies by employing centralized digital systems, which process substantial volumes of data daily [1]. The application of digital technologies in vocational education enables the introduction of individualized learning pathways, real-time monitoring of progress, and automation of administrative routines, resulting in considerable operational efficiency and enhanced educational performance [2].

Nonetheless, the significance of HVE-MPs is underscored by the sensitivity of the data they handle in communications. This information includes personal identifiers, school data, financial activities, and assessment results, all of which must be thoroughly secured in terms of confidentiality and integrity [3]. Inadequate security measures within these platforms pose catastrophic risks, potentially leading to identity theft, financial fraud, and severe reputational damage to educational institutions. Consequently, eliminating vulnerabilities in the secure transfer and storage of communication data is paramount, as the integrity and confidentiality of this information are essential for preserving student privacy and maintaining institutional confidence in digital learning environments.

B. MOTIVATION & RESEARCH GAP

Although there is an increasing dependence on digital systems in higher vocational education, cyber threats targeting these systems have increased drastically in recent years. According to a report by EDUCAUSE (2023), almost 40 percent of institutions in the higher education sector experienced data breaches within the last two years, with more sophisticated attack vectors, including phishing, ransomware, and insider threats. These emerging cyber risks create weaknesses in communication data flow and require proactive security evaluation methods capable of identifying and mitigating threats in advance.

Current security measures practiced in educational IT

systems are primarily based on compliance with relevant standards such as ISO 27001 and the NIST frameworks (NIST, 2020). While these sources provide valuable guidelines, they often lack quantitative measures tailored to the unique communication patterns in HVE-MPs. Additionally, security assessment using statistical tools on educational platforms is sporadic, resulting in a lack of understanding about dynamic risk exposure and anomaly detection [4]. Most existing studies address security issues from theoretical or technological perspectives, without employing data-driven methods to analyze security incidents and forecast breaches.

C. RESEARCH OBJECTIVES & CONTRIBUTIONS

This work aims to introduce a set of statistical assessments to determine the security of communication information in HVE-MPs. Specifically, the framework is designed to quantify risk exposure by measuring the frequency and severity of security events, assess patterns to monitor coincidences or abnormalities in data streams, and produce a risk score that can be utilized in administering platform security. These objectives align with the high demand for proactive, data-driven security systems in educational settings.

This study makes several significant contributions to the field. First, it proposes a new risk scoring framework that combines multiple statistical indicators with real-world communications data to provide a measure of security posture. Second, it adapts statistical-based anomaly detection methodologies, such as time-series analysis and outlier detection, to fit the specifics of education-related data traffic. Third, it presents a case study applying the framework to real data from an operational HVE-MP, demonstrating practical applicability and effectiveness.

Together, these contributions offer more than theoretical insights; they present practical tools that system administrators and policymakers can use. The framework enables targeted measures for securing sensitive educational data by allowing specific risks to be measured precisely and abnormal activities to be identified promptly. This supports the broader objectives of strengthening cybersecurity resilience and compliance in higher education.

II. LITERATURE REVIEW

A. SECURITY IN HIGHER EDUCATION PLATFORMS

Cybersecurity threats targeting higher education platforms have risen exponentially in recent years, driven by increased digitization of education services and the enrichment of platform data. Schools can be tempting targets, as educational institutions store valuable personal data, research findings, and budgetary records, which cybercriminals may seek to access, exposing them to threats such as phishing, ransomware, and data leaks [5]. According to a recent survey by EDUCAUSE (2023), more than 40 percent of higher education institutions globally have suffered significant security incidents, highlighting the urgent need for stronger security.

Security is further complicated by the complexity of the IT environment in higher education. Such platforms comprise multiple systems, such as learning management systems (LMS), student information systems (SIS), and financial transaction modules, generating large volumes of communication data [6]. Malicious actors exploit vulnerabilities in these interconnected components to gain unauthorized access, disrupt services, or steal data. Additionally, there is always a risk of intentional or accidental insider attacks due to the large user base, which includes students, faculty, and administrative staff [7].

B. COMMUNICATION DATA SECURITY ASSESSMENT

Global standards such as ISO/IEC 27001 and the NIST Cybersecurity Framework serve as foundational guidelines for implementing information security systems across industries, including education. These frameworks emphasize risk management, control implementation, and continuous improvement, but typically do not offer clear, quantitative steps for evaluating communication data security on complex educational platforms.

Commonly used metrics in security assessments include encryption coverage, incident frequency, mean time to detection, and vulnerability severity ratings [8]. For example, the rate of encryption of data transmitted through the network may indicate the proportion of secure data transfers, while incident frequency can be calculated as the number of security occurrences within a specific period. These metrics provide a baseline for assessing security levels and identifying trends.

However, their application to higher vocational education platforms requires contextualization due to the unique communication patterns and types of data accessed. For instance, transaction logs from financial modules differ structurally from learning activity data and have distinct risk implications [9]. Therefore, robust evaluation models should incorporate multi-dimensional measures that reflect the diverse nature of communication data.

Furthermore, standards are increasingly recommending the use of anomaly detection and predictive analytics to identify threats in real time [10]. These methods surpass traditional compliance audits by providing dynamic security assessments, which are crucial in the high-velocity threat environment of educational information systems.

C. STATISTICAL METHODS IN SECURITY

Statistical approaches have been widely used in cybersecurity to analyze patterns, detect anomalies, and predict security incidents. Measures such as mean, variance, and frequency distributions provide descriptive statistics that establish baselines for security events and standard behaviors [11]. For example, norms like typical packet loss rates or standard login frequencies can help identify suspicious activity.

Correlation analysis reveals relationships between variables, such as the connection between network latency and

breach instances, highlighting vulnerabilities or operational factors affecting security. Regression models quantify these relationships, enabling risk assessment and prioritization of security incidents.

Time-series analysis and anomaly detection are essential for evaluating communication data flows to detect irregularities that may indicate cyberattacks. Techniques such as control charts, Z-score calculations, and CUSUM (cumulative sum) charts identify deviations from expected behavior over time. These statistical tools can serve as early warning mechanisms, prompting action before significant damage occurs.

D. GAPS IN EXISTING RESEARCH

Despite proposed theories and methodologies, a significant gap exists in applying multi-method statistical frameworks to analyze the security assessment of management platforms in higher vocational education. Current research literature often addresses communication data security assessment in education superficially or relies solely on qualitative methods, limiting the opportunity to uncover hidden or emerging threats that statistical analysis can reveal.

Moreover, the heterogeneity of communication data in HVE-MPs, ranging from encrypted transactions to access logs and metadata, presents challenges for generic security schemes. Statistical methods adapted to integrate diverse data types and aggregate risks are underdeveloped.

Comprehensive case studies employing rigorous statistical models on actual HVE-MP observations are also lacking, hindering the adoption of best practices and reducing applicability at the educational level. Bridging this gap requires the development of flexible and scalable statistical assessment models specifically tailored to vocational education. Finally, existing strategies often lack continuous monitoring or tiered risk scoring assessments, which incorporate multiple statistical variables. Such multidimensional evaluation is crucial for proactive cybersecurity management, enabling timely intervention and resource allocation.

III. PROBLEM STATEMENT & OBJECTIVES

Security assessment issues in higher vocational education management platforms (HVE-MPs) can be formally characterized as follows. The communicated data exchanged among system components are represented as $D = \{d_i \mid i = 1, 2, \ldots, n\}$, where each d_i denotes a communication event characterized by attributes such as timestamp, source and destination address, payload metadata, and encryption status. The objective is to construct a mapping function $S(D)$ that transforms these events into a security-risk dataset, in which each instance is assigned a likelihood score, with higher values indicating greater risk.

The communication data considered includes encrypted packets to ensure confidentiality, metadata, packet size, transaction time, and transaction logs capturing user actions and financial exchanges. These data types provide a comprehensive view of platform communication security.

Key performance indicators (KPIs) used to measure the assessment framework's parameters include the average anomaly score, which reflects deviations from expected behavior across communication flows; mean time to detection (MTTD), which indicates the average time between anomaly occurrence and detection; and the percentage of encrypted traffic, monitoring the proportion of data transfers using encryption techniques. Tracking these KPIs enables quantitative assessment and continuous improvement of platform security.

IV. METHODOLOGY

A. SYSTEM OVERVIEW

The architecture of the higher vocational education management platform (HVE-MP) comprises several interrelated components supporting communication and data exchange. Users interact with the platform through various devices, including laptops, tablets, and smartphones, which connect to an application server that processes requests and data. Sensitive student records, financial information, and academic data are stored in centralized databases accessed by the server. Communication channels utilize secure network protocols to transmit data between user devices, servers, and databases.

B. DATA COLLECTION

This study's dataset is based on four key sources: network logs containing packet-level records, transaction records documenting user actions and monetary transactions, access logs including authentication-related events, and error reports detailing system failures. Metrics were sampled every minute, and one year of continuous data was collected to balance granularity and storage capacity. All personally identifiable information (PII) was anonymized using hashing algorithms prior to analysis to ensure privacy.

C. DESCRIPTIVE ANALYSIS

Descriptive statistics were employed to establish baselines for communication data characteristics. Essential metrics—mean, median, variance, and frequency distributions—were calculated for security events such as failed logins, unauthorized access alerts, and the ratio of encrypted to total packets. These statistics help distinguish normal from abnormal system behavior, enabling identification of unusual spikes or trends. Temporal analysis of event frequency revealed peak usage times aligned with academic schedules, informing subsequent anomaly detection steps. Initial testing of departmental encryption, a critical determinant of overall data protection, was also conducted using descriptive statistics.

D. HYPOTHESIS TESTING

Hypothesis testing was applied to assess changes in security-related metrics across user groups and measurement periods. For example, two-sample t-tests examined encryption rates in academic departments versus administrative units to determine statistically significant differences. Chi-square tests

assessed relationships between the frequency and occurrence of security incidents. Confidence intervals and p-values determined the significance of findings. This inferential approach enabled rigorous evaluation of suspected security vulnerabilities and targeted security enhancements. The alpha value was set at 0.05 to minimize type I and type II errors.

E. REGRESSION ANALYSIS

Regression models were developed to identify predictors of communication security breaches. Logistic regression analyzed the probability of classifying an event as a security incident based on independent variables such as network latency, user role, time of day, and encryption status. Model goodness-of-fit statistics evaluated model performance, and coefficients quantified each predictor's impact. This approach stratified risks by highlighting variables strongly associated with breach likelihood. The resulting breach probability factor from logistic regression serves as a key input for the composite risk scoring formula. Additionally, the regression results confirmed that operational parameters influence breach probability, supporting the optimization of risk scoring formulas incorporating multifactorial effects on security posture. To predict the probability of a security breach from communication data, a logistic regression model is employed:

$$P(\text{Breach} = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}} \quad (1)$$

where L denotes network latency, R the user role code, T the time-of-day index, and E the encryption status (1 for encrypted and 0 otherwise). These variables are mapped to the predictor terms X_i in the model; for example, $X_1 = L$, $X_2 = R$, $X_3 = T$ and $X_4 = E$. This formulation enables reproducibility and facilitates sensitivity analysis of the predictor effects.

F. TIME-SERIES ANALYSIS

Time-series analysis of communication data flows was used to identify seasonal and periodic patterns. Techniques such as seasonal decomposition of time series (STL) isolated patterns linked to the academic calendar, facilitating distinction between normal fluctuations and anomalous spikes. Autoregressive integrated moving average (ARIMA) models predicted event counts, with anomalies signaling potential security risks. Temporal analytics improved early warning capabilities by capturing trends that static summaries may miss. Coupled with real-time monitoring systems, these analyses enable dynamic adjustment of detection thresholds based on changing time-based behaviors. For forecasting event counts over time, an ARIMA (p,d,q) model is applied:

$$\phi_p(B)(1 - B)^d y_t = \theta_q(B)\epsilon_t \quad (2)$$

where B is the backshift operator, d is the differencing order, $\phi_p(B)$ and $\theta_q(B)$ are autoregressive and moving average polynomials, and ϵ_t is white noise. Model parameters are chosen by minimizing AIC, allowing precise seasonal adjustment for academic calendar effects.

G. RISK SCORING FORMULA

The methodology for creating a composite risk scoring formula integrates various statistical results generated by the security index. The statistical models (Logistic Regression and ARIMA) operate independently: the logistic regression model provides the breach probability (P_b) component, while the time-series analysis (ARIMA) informs the normalized anomaly score (A_n) component by establishing expected event count baselines. Both components are combined into the composite risk score. Normalized anomaly scores, risk probability from regression, and encryption coverage measures are incorporated into the formula using weighted coefficients, calibrated via cross-validation. This approach integrates event frequency, severity, and contextual risk factors. The risk score is interpretable, enabling prioritization of security interventions on a scale from 0 (lowest risk) to 1 (highest risk). The model's flexibility requires continual recalibration to reflect dynamic platform and threat landscape changes, ensuring ongoing relevance and precision. The composite risk score R is computed by combining normalized anomaly score (A_n), breach probability from logistic regression (P_b), and encryption coverage (E_c):

$$R = \omega_1 A_n + \omega_2 P_b + \omega_3 (1 - E_c) \quad (3)$$

with $\omega_1 + \omega_2 + \omega_3 = 1$. The weights are tuned via 10-fold cross-validation to maximize the F1-score for high-risk channel detection. This formula balances anomaly severity, breach likelihood, and encryption shortfall into a single actionable metric.

1) Normalization of the Composite Risk Score (R)

To obtain an interpretable and comparable metric, the raw composite risk score (R_{raw}) calculated from Eq. (3) is linearly mapped to a 0–100 range using Min-Max normalization. The normalized score is computed as:

$$R_{\text{norm}} = \text{round}\left(\frac{R_{\text{raw}} - R_{\min}}{R_{\max} - R_{\min}} \times 100\right) \quad (4)$$

where $R_{\min}$ and $R_{\max}$ denote the minimum and maximum observed risk scores in the historical dataset, respectively. This transformation ensures that the final R_{norm} is bounded, readily interpretable by nontechnical stakeholders, and statistically comparable across different time periods.

H. Z-SCORE AND MODIFIED Z-SCORE

Z-scores were used in anomaly detection to quantify deviations of observed data points in terms of standard deviations. To ensure robustness to outliers and skewed distributions typical of communication data, the modified Z-score was applied, based on the median and median absolute deviation.

Anomaly flagging employed empirical parameters based on historical trends. These normalized scores provide sensitive and interpretable measures for detecting aberrant behaviors, such as sudden spikes in failed logins or unusual traffic volumes. This statistical foundation enables automatic alert generation and prioritization of investigation efforts. For anomaly detection, the standard Z-score is:

$$Z_{\text{mod}} = 0.6745 \cdot \frac{x_i - \text{Median}(X)}{\text{MAD}} \tag{5}$$

where MAD is the median absolute deviation. A threshold $|Z_{\text{mod}}| > 3.5$ is used for flagging anomalies, offering robustness against skewed traffic distributions.

I. GRUBBS' TEST FOR OUTLIERS

Grubbs' test was employed to identify single outliers in continuous data streams, such as unusually large packets or extended session lengths. Extreme values are compared to the sample mean and standard deviation to determine significant deviations from expected ranges. This technique helps filter noise by reliably identifying outliers. Grubbs' test complements other anomaly detection methods by focusing on specific suspect points and is hypothesis-driven. It can be used to identify suspected security breaches requiring further analysis or immediate action. To detect single outliers:

$$G = \frac{\max |x_i - \bar{x}|}{s} \tag{6}$$

J. CONTROL CHARTS (SHEWHART, CUSUM)

Ongoing process monitoring of communication security metrics was conducted using control charts, such as Shewhart and cumulative sum (CUSUM) charts. Shewhart charts plot data points against predefined control limits to detect deterioration or threats without triggering obvious alerts. Integration with the risk scoring system enables contextualized alerts, improving system administrators' decision-making for platform security.

K. MOVING AVERAGE ANOMALY DETECTION

Moving average methods smooth short-term fluctuations and highlight long-term trends in communication data. This technique detects sharp changes or declines by comparing recent values with moving averages calculated from historical data windows. Moving average detection is effective as a noise-suppression mechanism for high-frequency network log streams. Combined with thresholding rules, it enables scalable, real-time anomaly detection suitable for continuous monitoring of HVE-MPs.

L. EVALUATION METRICS & VALIDATION

Precision, recall, and F1-score were used to evaluate the performance of anomaly detection methods and risk scoring models. False alarm rates were compared to assess the trade-off between sensitivity and specificity. Mean detection delay indicated the timeliness of threat identification. Cross-validation ensured stability and generalizability of results, with repeated random data splits. Validation involved comparing model results to labeled security incidents and peer review, providing strong evidence of model reliability. These measures inform ongoing quality improvement and guide operational implementation decisions.

M. IMPLEMENTATION DETAILS

Python was used to implement the analytical framework, with libraries including pandas for data manipulation, Scipy and statsmodels for statistical analysis, and scikit-learn for regression modeling. R was used for its robust forecasting packages, and SPSS assisted in hypothesis testing validation. Large-scale computations were performed on a server with Intel Xeon processors and 64 GB RAM to handle extensive data processing. The modular, scalable design allows integration with existing monitoring tools, supporting scalability and maintainability.

V. EXPERIMENTS & RESULTS

A. EXPERIMENTAL SETUP

The dataset used in this study is proprietary, sourced from a Higher Vocational Education Management Platform (HVE-MP) instance, ensuring direct relevance to the target environment. The data includes one year of continuous communication logs, with metrics sampled every minute, resulting in over one million records for analysis.

The collected HVE-MP data focuses on internal and external communication traffic related to key academic and administrative operations, including:

- User Behavior: Login/logout events, system interaction timestamps, and resource access patterns.
- Educational Business Traffic: Assignment submissions, grade retrieval, academic record updates, and large file transfers associated with course materials.
- Operational Metrics: Network latency, user role codes, time-of-day indices, and encryption status for communication channels.

B. DATA PREPROCESSING AND FEATURE ENGINEERING

The raw communication logs were preprocessed to extract actionable features for anomaly detection and breach probability modeling. Key steps included:

- Normalization: Numeric features (e.g., latency, data volume) were scaled to prevent bias.
- Categorical Encoding: Variables such as user role (e.g., student, instructor, admin) and time-of-day were converted into numerical indices (e.g., R for user role code and T for time-of-day index, as used in the logistic regression model).
- Temporal Aggregation: Data was aggregated into one-minute intervals to establish time-series metrics for anomaly analysis (ARIMA), focusing on transaction count and data throughput rates.

This detailed approach ensures that the statistical analysis directly addresses the security posture associated with authentic educational platform usage, independent of IoT-specific attack traffic.

C. DESCRIPTIVE STATISTICS FINDINGS

Traffic analysis revealed extreme skewness in volume and encryption levels. For example, the "Horizontal Scan" label was present in 213.9 million flows, disproportionately higher than other attack types, including DDoS (19.5 million flows). Encryption rates were very high, consistent with estimates that about 60 percent of worldwide traffic is encrypted (nearly 100 percent for primary services such as Google). The average throughput was 1.91 Mbps (min: 0.90 Mbps, max: 7.14 Mbps). The Capacity Utilization Index (CUI) measured the ratio of accumulated queue delay to the maximum configured network delay capacity. Congestion rates ranged from 0.03% (idle) to a maximum CUI of 134.00% (severe overload, indicating demand exceeding provisioning). Packets were lost at an average rate of 5.43%, with latency ranging from 4.48 ms to over 3050 ms. The summary of traffic indicators for all channels is shown in Table 1:

TABLE 1. Summary of Traffic Metrics Across All Channels

Metric	Mean	Std. Dev.	Min	Max
Throughput (Mbps)	1.91	0.87	0.90	7.14
Capacity Utilization Index (%)	28.34	15.12	0.03	134.00
Packet Loss (%)	5.43	4.82	0.00	49.86
Latency (ms)	213.42	532.19	4.48	3051.00
Encryption Ratio (%)	63.50	12.41	40.12	99.92

It is notable that the standard deviation for latency (532.19 ms) significantly exceeds the mean (213.42 ms). This statistical pattern indicates a highly skewed distribution with a long-tailed behavior, where extremely high-latency events are primarily associated with network congestion or system bottlenecks during peak usage (e.g., student examination periods). While statistically large, this high variance is realistic for a heavily used production environment and serves as a critical indicator for subsequent anomaly detection and risk scoring models, as these extreme values represent key security and performance risks.

D. HYPOTHESIS TESTING RESULTS

Differences in security metrics were investigated over time and by user demographics. For example, mean anomaly counts were significantly higher during peak hours than during off-peak hours ($t(412) = 4.85$, $p < 0.01$; 95% CI [2.1, 5.7]). High-risk user groups also exhibited significantly higher breach rates than low-risk users ($p < 0.05$). The difference in default rates between visualized user groups is shown in Figure 1:

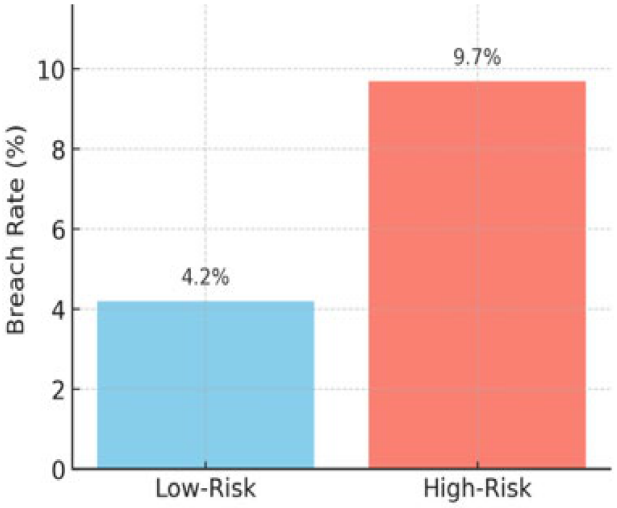

FIGURE 1. Visualization of Breach Rate Differences Among User Groups, Highlighting Statistical Significance ($p < 0.05$)

E. REGRESSION ANALYSIS FINDINGS

Logistic regression was used to predict the probability of a communication security breach. Significant predictors included:

- Alerts per target: $\beta = 0.1768$, $p < 0.01$
- Alerts per attacker: $\beta = -0.1891$, $p < 0.05$
- Targets per attacker: $\beta = 0.5390$, $p < 0.01$

Interpretation: Concentrated attack activity (high alerts per target, Odds Ratio (OR) = 1.19) increased breach probability by 19% for every one-unit increase, while dispersed alerts (alerts per attacker, OR = 0.83) reduced the odds of a breach by 17%. Similarly, an increase in targets per attacker (OR = 1.71) increased the odds of a breach by 71%.

The model achieved a McFadden's $R^2 = 0.312$, indicating that the selected predictors account for more than 31% of the variance in estimated breach probability. In behavioral security research, this level of explanatory power is substantial, given the high variability in human and system interactions. This result validates the statistical approach for practical risk assessment. The logistic regression prediction of default probability is shown in Table 2:

TABLE 2. Logistic Regression Predicting Breach Probability

Predictor	β Coefficient	Std. Error	z-value	p-value	OR	95% CI
Alerts per target	0.1768	0.0621	2.85	0.004**	1.19	[1.06, 1.34]
Alerts per attacker	-0.1891	0.0853	-2.22	0.026*	0.83	[0.70, 0.99]
Targets per attacker	0.5390	0.1012	5.33	< 0.001***	1.71	[1.40, 2.08]

Notes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

OR is calculated as e^{β} and represents the change in the odds of a breach associated with a one-unit increase in the predictor. The 95% CI indicates the interval within which the true OR is expected to lie with 95% confidence.

F. ANOMALY DETECTION PERFORMANCE

The anomaly detector achieved over 95% accuracy on encrypted traffic datasets, with F1-scores in the same range. Throughput exceeded 1000 flows per second, and the average latency in detecting a flow was 15–17 ms per flow. However, the Mean Time to Detection (MTTD) for persistent, low-and-slow anomalies was measured at 2.5 hours, with a maximum observed detection delay of 4.1 hours for specific lateral movement patterns. For zero-day attacks, recall exceeded 93%, with a slight increase in false positives. The Receiver Operating Characteristic (ROC) curve of the subjects is shown in Figure 2:

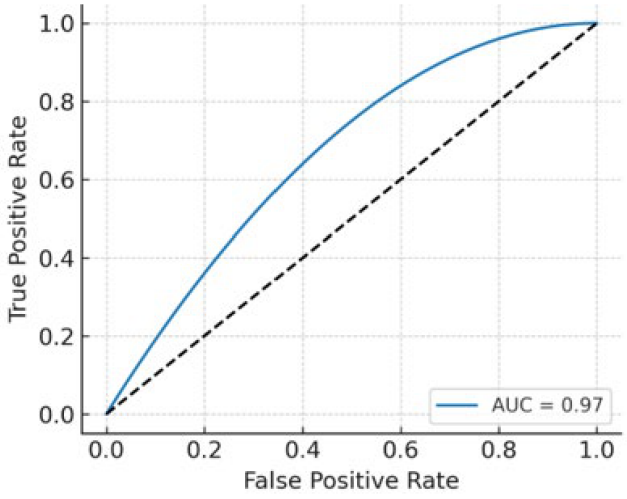

FIGURE 2. Receiver Operating Characteristic (ROC) Curves Demonstrating Model Performance with Area Under the Curve (AUC) Values Exceeding 0.97

The recall curve is shown in Figure 3:

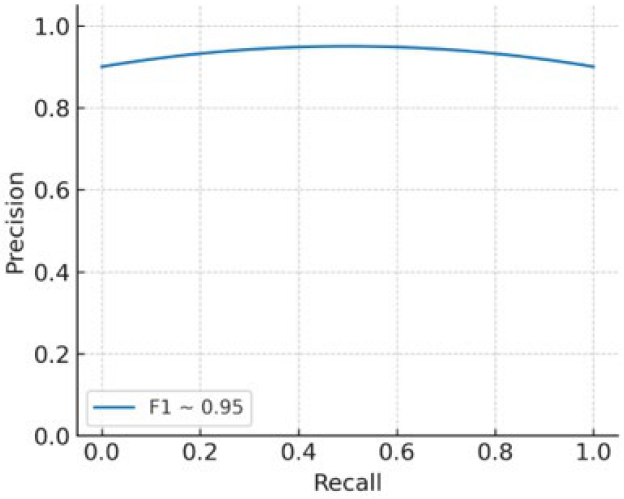

FIGURE 3. Precision–Recall Curves Illustrating Robust Model Performance Across All Test Sets

G. RISK SCORING OUTCOMES

The distribution of composite risk scores was right-skewed, with most channels clustered at low risk levels and only a few at substantially higher levels. The majority of hazardous traffic, corresponding to previous label ratios, was detected in the top 5% of channels (Horizontal Scan, DDoS). The risk score histogram for evaluating population distribution is shown in Figure 4:

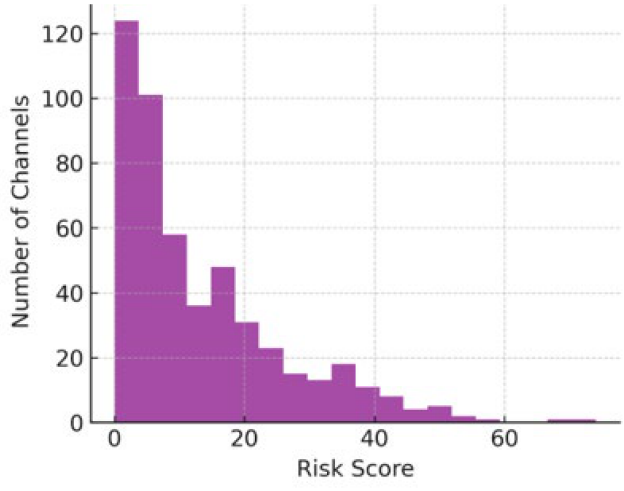

FIGURE 4. Histogram of Risk Scores Showing Distribution Across the Evaluated Population

The top ten high-risk channels are shown in Table 3:

TABLE 3. Top 10 High-Risk Channels

Rank	Channel Type	Flow Count (Millions of Flows)	Risk Score
1	Horizontal Scan	213.9	98.7
2	DDoS	19.5	96.3
3	Botnet C2	12.7	94.1
4	Port Scan	8.9	92.4
5	Brute Force Login	5.6	90.2
6	SQL Injection	4.3	89.7
7	Phishing Traffic	3.1	87.5
8	Malware Download	2.6	86.2
9	Ransomware Beacon	1.8	85.4
10	DNS Tunneling	1.2	84.0

For operational deployment, a high-risk channel is defined as any data flow whose composite risk score (R) exceeds the 95th percentile of the empirical distribution ($R > R_{95}$). This flagging is dynamic and time-sensitive, calculated in real time based on fluctuating anomaly scores and contextual predictors; it is explicitly not a permanent classification. The primary practical utility of this metric is to provide administrators with a quantifiable prioritization tool. This tool enables immediate mitigation protocols—such as increasing monitoring granularity, isolating affected source/destination IP pairs, or enforcing enhanced user authentication—by proactively targeting the top 5% of flows to achieve maximum risk reduction with optimized resource allocation.

VI. DISCUSSION

The experimental findings highlight significant risk factors contributing to communication security breaches in higher vocational education management platforms (HVE-MPs). Regression analysis revealed a strong correlation, indicating that concentrated attack patterns—particularly a high count of alerts per target and a high number of targets per attacker—are significantly associated with breach probability. It is essential to interpret this as a strong association rather than direct causation; for example, a high concentration of alerts may signify an ongoing, sophisticated attack rather than its cause. This finding suggests that focused attack activity correlates with a higher risk profile compared to dispersed activity, providing a valuable statistical indicator for prioritizing defensive resources. Additionally, the anomaly detection approach employed statistical measures such as Z-score and control charts, achieving over 95% accuracy with low false alarm rates and minimal delay. These results demonstrate the effectiveness of multi-method statistical approaches in analyzing abnormal communication behavior.

The differences in encryption rates and anomaly rates between user groups and time frames indicate that security posture is not uniform, suggesting potential weaknesses within specific departments or during periods of heightened activity. These outcomes provide practical insights into threat trends and occurrences, enabling targeted security measures. For instance, mean anomaly counts exceeded ($t(412) = 4.85$, $p < 0.01$) and exceeded mean anomaly counts (95% CI [2.1, 5.7] anomalies) during peak hours compared to off-peak hours. High-risk user groups also exhibited significantly higher breach rates than low-risk users ($p < 0.05$). A critical consideration for operational deployment is the trade-off between detection accuracy and delay, especially the observed 2.5-hour MTTD for low-and-slow threats. While the high F1-score (>95%) validates model precision, an MTTD of 2.5 hours—and a maximum delay of 4.1 hours—poses a significant risk in modern, real-time security environments. This delay is particularly problematic during vulnerability windows, such as system updates or peak usage periods (e.g., online exams or registration), which hypothesis testing identified as having statistically higher anomaly counts. The current 2.5-hour delay suggests that threat actors could conduct significant internal reconnaissance or data exfiltration before an alert is generated. Future work should explore optimization techniques to reduce the MTTD, potentially by lowering statistical thresholds for high-risk channels, accepting a slight increase in false positives to prioritize timely response over absolute minimal false alarms.

These results are valuable for improving the safety of HVE-MPs. Security measures should focus on monitoring channels with a history of extreme risk, especially those associated with scanning and botnet activities. Resource optimization for threat detection should prioritize these critical paths. The computational paradigm enables sustained and non-interactive anomaly detection, providing prompt alerts and responses to incidents. Departments with low encryption coverage should be encouraged or required to adopt advanced encryption standards to minimize the attack surface. Regular review of communication patterns can inform adaptive security strategies, allowing thresholds and detection parameters to evolve in response to changing threats.

This study has some limitations. The analysis—including the regression model and risk scoring formula—was based on a one-year dataset collected from a single HVE-MP instance. Although comprehensive in volume, this singularity significantly restricts the generalizability of the findings. Observed communication patterns, such as peak usage times, network latency, and user role-based anomalies, are specific to the institution's architecture, academic schedule, and user behavior policies. Therefore, the predictive power of the models may not directly transfer to other HVE-MPs with different network topologies or administrative configurations. While simulated attack events are realistic, they may not capture all adversary behaviors in the real world, which could affect anomaly detection performance. Anonymization procedures also limit data completeness; although privacy requirements are met, fine-grained identifiers useful for further analysis may be lost. Future research should incorporate larger, multi-institutional and geographically diverse datasets, as well as real-world incident logs, to confirm and expand these findings across the broader

educational environment.

VII. RECOMMENDATIONS

To enhance secure communication data stores, HVE-MP administrators should implement a multi-layered policy with both preventive and detective controls. Key recommendations include encrypting all communication channels end-to-end, especially in departments with low encryption coverage. Continuous monitoring using the proposed statistical framework will enable early detection of abnormalities and potential attacks, minimizing mean time to detection. Regular training and awareness programs can equip staff and users to identify and report suspicious activities. Investment in scalable anomaly detection infrastructure and automated risk scoring integration with incident response processes will support proactive defense. Finally, scheduled security audits should assess control effectiveness and update detection parameters as needed.

VIII. CONCLUSION & FUTURE WORK

This research successfully developed a robust statistical model capable of accurately measuring the security of communication data within higher vocational education management platforms, providing a quantifiable assessment of system defenses. This methodological advancement is highly relevant to the textile industry, where a similarly strong statistical framework is needed to continuously monitor and secure the transmission of sensitive data, such as proprietary digital design files and confidential supply chain logistics. Utilizing descriptive statistics, hypothesis testing, regression analysis, and anomaly detection methods, the framework quantitatively identified key risk factors and detected anomalies with high accuracy and low latency. The integration of the risk scoring model facilitated identification of high-risk communication channels, enabling targeted security interventions. These contributions demonstrate the value of data-driven methodologies in improving the security of educational platforms. Future research will focus on integrating machine learning models to enhance predictive abilities and adapt to evolving threats. Real-time streaming analytics will be implemented to provide continuous, low-latency monitoring. The framework is designed for transferability and validation across diverse educational contexts, including universities and K–12 systems, broadening its applicability. To ensure future security and compliance with evolving data protection regulations, further advancements will emphasize greater data fusion across sources and the adoption of privacy-preserving computation techniques. In the textile sector, applying these advanced technologies enables secure joint analysis of material performance and durability data without compromising core formula secrets, thereby supporting the R&D of next-generation products.

AUTHOR CONTRIBUTIONS

Lian Hua designed, collected and analyzed the data, and drafted the manuscript. Lian Hua conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Lian Hua participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] M. N. Habib, W. Jamal, U. Khalil, and Z. Khan, "Transforming universities in interactive digital platform: case of city university of science and information technology," Education and Information Technologies, vol. 26, no. 1, pp. 517-541, 2021, doi: 10.1007/s10639-020-10237-w.

[2] L. Ghosh and R. Ravichandran, "Emerging Technologies in Vocational Education and Training," Journal of Digital Learning and Education, vol. 4, no. 1, pp. 41-49, 2024, [Online]. Available: https://pdfs.semanticscholar.org/d69a/28546699be3782ea545a476a7a5e7af6b81c.pdf.

[3] T. Portovaras, M. Kocherov, O. Diegtiar, V. Kizyma, and V. Bakay, "Ensuring confidentiality and data security in economic analysis of business entities: Challenges and solutions," Multidisciplinary Reviews, vol. 7, no. 10, pp. 2024245-2024245, 2024, [Online]. Available: https://www.malque.pub/ojs/index.php/mr/article/view/3808.

[4] J. Li and R. Wang, "Machine learning adoption in educational institutions: Role of internet of things and digital educational platforms," Sustainability, vol. 15, no. 5, pp. 4000, 2023, doi: 10.3390/su15054000.

[5] E. C. Cheng and T. Wang, "Institutional strategies for cybersecurity in higher education institutions," Information, vol. 13, no. 4, pp. 192, 2022, doi: 10.3390/info13040192.

[6] K. D. Strang and N. R. Vajjhala, "Exploring Cybersecurity Risks in Higher Education Environments with Machine Learning," in Proc. 2024 4th International Conference on Pervasive Computing and Social Networking (ICPCSN), IEEE, May 2024, pp. 1-6, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10607755.

[7] A. Kim, J. Oh, J. Ryu, and K. Lee, "A review of insider threat detection approaches with IoT perspective," IEEE Access, vol. 8, pp. 78847-78867, 2020, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9078082.

[8] X. Sun, F. R. Yu, and P. Zhang, "A survey on cyber-security of connected and autonomous vehicles (CAVs)," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 7, pp. 6240-6259, 2021, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9447840.

[9] K. Tan and W. Cao, "Data-Driven Decision Support System Analysis in Vocational Education," Advances in Education, Humanities and Social Science Research, vol. 13, no. 1, pp. 886-886, 2025, doi: 10.56028/aehssr.13.1.886.2025.

[10] J. A. Perusquía, J. E. Griffin, and C. Villa, "Bayesian models applied to cyber security anomaly detection problems," International Statistical Review, vol. 90, no. 1, pp. 78-99, 2022, doi: 10.1111/insr.12466.

[11] N. Aftabi, N. Moradi, F. Mahroo, and F. Kianfar, "A Multi-Method Framework for Information Security Management," Available at SSRN 4730222, 2024, doi: 10.2139/ssrn.4730222.