INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2026, VOL. 72, NO. 3, PP. 1-9
Manuscript received March 13, 2026; revised June, 2026. doi: 10.24425/ijet.2026.157959

Hybrid method for detecting atypical behavior of users of information and educational systems based on Digital Traces analysis

Olena Kryvoruchko, Miroslav Lakhno, Elizaveta Zavhorodnya, Yaroslav Shestak, Valerii Lakhno, Bauyrzhan Tynymbayev, and Oleksii Savon

***Abstract*—The study presents a hybrid multi-level method for analysing the digital traces of users of information and educational systems (IES) in higher education institutions, aimed at detecting atypical behaviour and assessing information-security risks. The method integrates hierarchical structuring of event processes, graph-cluster analysis of interactions, and behavioural risk assessment based on a hybridisation of machine-learning models. The use of real event logs made it possible to demonstrate the method's effectiveness in detecting risk-related deviations that cannot be identified using existing tools for monitoring user behaviour in information and educational systems in higher education institutions.**

***Keywords*—digital traces; information and educational system; atypical behaviour detection; structural event analysis; user graph models; machine learning; information security; behavioural analytics**

I. Introduction

INFORMATION and educational systems of higher education institutions (hereinafter referred to as HEI IES) differ from conventional systems in the degree of integration of digital educational services within the digital learning environment (DLE), as well as in the variety of integrated learning platforms, such as learning management systems and electronic dean's offices. Accordingly, the volume and variability of users' digital traces (DTs) in HEI IES have increased significantly. Besides, DTs include logs generated by components of university information and educational systems (hereinafter IES) during user interactions. The tasks of monitoring information security (IS) in HEIs IES have also become more complex under wartime conditions. These tasks include tracking atypical user behaviour based on DT analysis and preventing incidents. It should be noted that common IS tools (including SIEM, DLP, and IDS/IPS systems) do not respond to events associated with local atypical behavioural indicators of users in IES. Moreover, these IS tools are not adapted to the specifics of educational processes. This creates a need for specialised multi-level methods capable of detecting atypical behavioural patterns customised to the operation of HEI IES and combining structural, graph-based, cluster-based, and behavioural characteristics of user actions, thus underscoring the relevance of the study.

II. Literature review

Issues of collecting, classifying, and analysing DTs have been widely examined in studies across cybersecurity, process mining, digital pedagogy, behavioural analytics, etc. Several research directions related to DT analysis have been established. The first group of works concerns structural and graph-based analyses of user behaviour. In these studies, digital events are viewed as traces of activity in IES. Thus, studies [1], [2], and [3] interpret DTs as sources of information about processes and interactions in information systems. Studies [4] and [5] focus on the role of graph models in representing digital behaviour. Another research direction concerns log analytics and anomaly detection in cybersecurity, particularly through machine learning (ML). For example, study [6] is devoted to rapid behavioural patterns identification while study [7] proposes representation learning for log data. Article [8] analyses the role of instructions in logs, and report [9] introduces a transformer-based model for anomaly detection in logs. Although this model works with unstructured logs, its application to IES log analysis is complicated by the heterogeneity of educational events and insufficient consideration of HEI learning processes.

A separate research direction focuses on detecting user behaviour anomalies using classical ML algorithms. The most common instruments include XGBoost and Isolation Forest (IF). The advantages of XGBoost in classification and risk assessment tasks are demonstrated in [10], [11], and [12], as well as in the study on the selection of classification thresholds [13]. Isolation Forest is also used for anomaly detection, including insider-threat detection [14], atypical business-process behaviour [15], and anomalies in corporate systems [16]. However, IF typically shows limited performance when dealing with complex behavioural patterns (typical for HEI IES), where user behaviour varies depending on time, context, and user role.

Another research direction involves DT analysis in higher education. Numerous studies examine DTs of students, from

Olena Kryvoruchko, Miroslav Lakhno and Valerii Lakhno are with National University of Life and Environmental Sciences of Ukraine, Kyiv, Ukraine (e-mail: ev_kryvoruchko@ukr.net, valss725@gmail.com, lva964@nubip.edu.ua).

Elizaveta Zavhorodnya, Yaroslav Shestak and Oleksii Savon are with State University of Trade and Economics, Kyiv, Ukraine (e-mail: y.zavhorodnya@knute.edu.ua, shestack@knute.edu.ua, o.savon@knute.edu.ua).

Bauyrzhan Tynymbayev is with Al-Farabi Kazakh National university (e-mail: b.tynymbayev@gmail.com).

analysis of individual learning trajectories [17], [18] to assessing risky behaviour in HEI IES [19]. Studies [20] and [21] showed that DTs reflect behavioural patterns of students and can be used to model their activity. However, such works focus mostly on educational tasks (e.g. identifying at-risk student groups, improving learning services etc.) rather than detecting security incidents in HEI IES.

The final research direction includes studies viewing DTs as elements of the HEI information security. Studies [22], [23], and [24] demonstrate the specifics of cyber-threats in higher education institutions.

However, these studies have certain limitations: most of them focus either on learning analytics or on general cybersecurity methods. They do not address the full analysis cycle – from structural modelling of events in HEI IES to risk-oriented behavioural assessment. Besides, they do not consider the hierarchical structure of events in IES, which is essential for modelling user behavioural patterns. Thus, a scientific gap exists: the need to develop hybrid methods for analysing the DTs of HEI IES users that account for structural dependencies between events, behavioural and graph characteristics of users, as well as risk assessment based on combined ML approaches. This confirms the relevance of the study and the need to develop a hybrid method for DT analysis in HEI IES.

III. Aim and Tasks of the Study

The aim of the study is to develop a hybrid multi-level method for analysing the digital traces of users of information-educational systems in higher education institutions.

To achieve this aim, the following tasks must be addressed to: 1) develop a multi-level hybrid method for analysing the digital traces of IES users in HEIs; 2) conduct a computational experiment using data from the IES of a large HEI and to examine the sensitivity of the proposed method to different types of atypical user actions, comparing it with baseline anomaly-detection methods, namely XGBoost, Isolation Forest (IF), and an ensemble of XGBoost and IF; 3) evaluate the effectiveness of the proposed method using the ROC-AUC, PR-AUC, and Score Distribution metrics.

IV. Methods

This section presents the description of the method that integrates three components to solve the task of analysing user log data (i.e., digital traces) in information and educational systems (IESs), namely: structural hierarchical modelling of processes within IES; cluster-graph analysis of user behaviour; and risk assessment of atypical behaviour. The method forms a system of mappings thus linking raw digital traces of IES users to the final assessment of behavioural atypicality.

We introduce a formal description of a digital event. Events are recorded in the log files of a HEI IES. Let: $U = \{u_1, \ldots, u_N\}$ – be the set of users; $D = \{d_1, \ldots, d_q\}$ – the set of devices or client IP addresses; $O = \{o_1, \ldots, o_r\}$ – the set of bi operation types; and $S = \{s_1, \ldots, s_\omega\}$ – the set of operation execution statuses.

Then each event (e) is defined as follows (1):

$$e = (t_e,\ u_e,\ d_e,\ o_e,\ s_e) \tag{1}$$

where te is the timestamp; ue is the user; de is the device or IP address; oe is the operation executed by the user; se is the execution outcome.

All events over the analysed period T form a stream (2):

$$L(T) = (e_1,\ e_2, \ldots, e_n) \tag{2}$$

Thus, the stream L(T) constitutes the initial representation of user activity in the IES of a higher education institution. To transition from isolated events to an understanding of user behaviour, a hierarchical process model is introduced. At the lowest level, we fix atomic actions and denote their set as (3):

$$B = \{b_1, b_2, \ldots, b_m\} \tag{3}$$

Each atomic event corresponds to an indivisible user operation in the IES, such as: b1 – login; b2 – read of a resource (read); b3 – data modification (write); b4 – lectures viewing (view lecture); b5 – upload file etc.

Then, we define a set of relation types between two atomic events bi and bj as follows:

$$R = \{\prec, ||, Y, \oplus\} \tag{4}$$

where $R = \{\prec, ||, Y, \oplus\}$ denotes strict sequence (i.e. bi precedes bj); $R = \{\prec, ||, Y, \oplus\}$ – parallelism or independence (the events bi and bj occur simultaneously); $b_i Y b_j$ – cyclic dependence (a repeating sequence within a session); $b_i \oplus b_j$ – choice or branching (after event bi, the user proceeds to one of several alternative actions).

Next, we introduce a mapping that assigns each ordered pair of events its corresponding relation type, i.e. we represent the relationship between two events as follows (5):

$$\rho: B \times B \to R \tag{5}$$

For any two events we have $\rho(b_i, b_j) \in R$. This means that between each pair of events we can define one of four relationships. The function ρ forms the basis for the structural analysis of action sequences in the IES.

Moving from atomic to aggregated actions, we introduce an aggregation operator: $Agg: 2^B \to A$, where A denotes aggregated actions such as "module completion" or "work session execution" etc. The process of building the hierarchy is defined recursively (6):

$$H^{(0)} = B, \quad H^{(k+1)} = Agg(H^{(k)}) \tag{6}$$

Recursion continues until one of the stopping conditions (7) is met:

$$\begin{cases} H^{(k+1)} = H^{(k)} \\ k = K_{max} \\ H^{(k)} = 1 \end{cases} \tag{7}$$

As a result, we form a complete hierarchical model (8):

$$H = \bigcup_{k=0}^{K} H^{(k)} \quad (8)$$

where $H^{(k)}$ is the hierarchy level k; K_{max} is the maximum depth; Agg is the algorithm for detecting aggregated subprocesses is HEI IES; H is the full hierarchical model. As a result, expression (8) makes it possible to identify behavioural patterns at different abstraction levels – from individual events to complete processes in the IES.

At the next stage, we construct the graph of user interactions within the IES. Let $S_u \subseteq L(T)$ be the set of events of user u is HEI IES. We define a similarity measure between two users' activities (9):

$$\omega(u_i, u_j) = \frac{\left|S_{u_i} \cap S_{u_j}\right|}{\left|S_{u_i} \cup S_{u_j}\right|} \in [0,1] \quad (9)$$

where $\omega = 1$ indicates identical activity sequences; and $\omega = 0$ indicates no shared actions.

The interaction graph is then defined as (10):

$$G = (U, W) \quad (10)$$

where $W = [\omega(u_i, u_j)]$ is the adjacency (similarity) matrix.

In order to cluster behavioural trajectories of users, spectral clustering is appropriate [25]. User behaviour in the IES is represented by the graph of transitions between atomic actions b_i and b_j. The spectral method operates directly with the graph structure using the Laplacian matrix (11), which encodes the topology of connections between actions or users in the HEI IES [25]. First, we form the Laplacian matrix [26]:

$$L = D - W \quad (11)$$

where $D = diag(d_i)$ is the diagonal degree matrix, whose elements are defined as d_i; $d_i = \sum_j \omega(u_i, u_j)$ characterises the overall similarity of user behaviour u_i with other HEI's IES users.

Clusters are obtained by solving the optimisation problem:

$$\min_{C} Tr(C^T L C) \text{ on condition that } C^T C = I \quad (12)$$

where Tr is the sum of diagonal elements; $C^T LC$ is the quadratic form of the graph Laplacian, measuring the smoothness of the spectral embedding C with respect to the user similarity graph, $C^T C = I$ is the orthonormality of the vectors sought.

Solving problem (12) gives k eigenvectors of the Laplacian matrix, forming the spectral embedding for subsequent clustering. Each user u is thus assigned a cluster (13):

$$C(u_i) \in \{1, \ldots, K\} \quad (13)$$

where K is the total number of clusters detected.

This enables identification of stable behavioural communities within the HEI's IES from a security-oriented perspective. For each user, we construct an individual profile (14):

$$p(u) = \left(f_{stat}(u), f_{dyn}(u), C(u)\right) \quad (14)$$

where $f_{stat}(u)$ represents static account attributes; $f_{dyn}(u)$ is time-varying activity parameters; and $C(u)$ is the behavioural cluster.

The profile serves as input for the stochastic risk assessment of user behaviour based on the XGBoost model [12, 13].

Let $x \in \mathbb{R}^d$ be a feature vector of an event or a user profile in HEI's IES. According to [12, 13], the XGBoost model is an ensemble of trees (15):

$$F(x) = \sum_{k=1}^{M} f_k(x) \quad (15)$$

where M is the number of trees in XGBoost ensemble; $f_k(x)$ is the separate regression tree, k is the index of the leaf where object x ends up when traversing the tree.

Since the ensemble of trees provides an additive model for assessing user behaviour in the IES, we obtain a probabilistic interpretation of the result by applying a sigmoid transformation to the total output of the model. This allows us to move from the values of the regressor tree to the generalised probability of atypical behaviour. We will define the probability of atypical behaviour of an IES user as follows (16):

$$R_{xgb}(x) = \sigma\left(F(x)\right) \quad (16)$$

where $\sigma(.)$ is a sigmoid function, $R_{xgb}(x) \in [0,1]$ is probability of atypical behaviour.

The classification rule [13] is (17):

$$R_{xgb}(x) > \tau \quad (17)$$

Thus, we obtain a probabilistic estimate based on the trained model. Alongside the XGBoost model, it is advisable to take into account the other nature of atypical user behaviour in the IES (namely, rare and abnormal trajectories). These patterns do not necessarily exhibit a high classification risk. To address this, we employ the stochastic Isolation Forest model [16], which evaluates atypicality based on the average path length required to isolate an observation. For the Isolation Forest model (hereinafter IF), the average path length is computed as follows [16]:

$$h(x) = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \quad (18)$$

where T is number of trees in Isolation Forest, $h_t(x)$ is depth at which object x is isolated in tree t.

Based on the average isolation depth, we obtain a normalised estimate of the user's behavioural atypicality, which is independent of the sample size and the properties of random trees. Accordingly, we define atypicality as (19):

$$R_{if}(x) = 2^{-\frac{h(x)}{c(x)}}, \quad c(n) = 2H_{n-1} - \frac{2(n-1)}{n} \quad (19)$$

where *h(x)* is the average isolation path length of object *x* in random trees; and *c(n)* is a normalisation constant that approximates the expected path length for a sample of *n* observations.

In order to combine the strengths of both models (i.e., the predictive accuracy of XGBoost and the sensitivity of IF to rare behavioural structures), we introduce a weighted combination of their output estimates. The IF model, according to [16], enables the detection of rare and non-classical behavioural patterns of the IES users. Then, to integrate the two models, XGBoost and IF, we define the following weighted combination (20):

$$R(x) = \alpha R_{xgb}(x) + (1-\alpha)R_{if}(x),$$
$$0 \le \alpha \le 1 \quad (20)$$

where α is a weighting parameter determining the contribution of each model – XGBoost (see equation (16)) and IF (equation (18)).

Next, we define an integral criterion for assessing the atypicality of HEI IES user behaviour. If the combined score exceeds the predefined threshold, the corresponding record or session is deemed atypical. Thus, a record is considered anomalous when:

$$R(x) > \theta \quad (21)$$

As a result, the obtained function (21) ensures consistency between predictive accuracy and resistance to rare situations of risky behaviour exhibited by HEI IES users.

V. Computational Experiments

A. Methodology for conducting computational experiments

The purpose of the computational experiment is to assess the quality of the developed hybrid method for detecting atypical behaviour in the HEI IES and to compare it with baseline algorithms (namely, the XGBoost gradient boosting model and the ensemble algorithm IF). The experiment was divided into the following stages:

Stage 1. Formation of a dataset of digital events in the HEI based on the data obtained from the website elearn.edu.ua. The dataset contained 200,000 events with 5,000 users of the IES of the State University of Life and Environmental Sciences of Ukraine. The dataset included: event-based features (operation, status, device, IP address); process features (session position, session length); contextual features (day of the week, time of day); risk markers (anomalous operation, suspicious device, high-risk IP addresses); and a final anomaly label with 10% noise.

Stage 2. Aggregation of data at the user level. For each user, the following were calculated: activity intensity; diversity of operations in the HEI IES; share of errors; share of suspicious IPs; graph-based features (centrality, rarity of transitions); and integrated anomaly indicators. In addition, training, validation, and test samples were formed. The split was performed at the user level to prevent information leakage. Accordingly, the training sample comprised 70% of users, the validation sample 15%, and the test sample 15%.

Stage 3. Verification of the viability of the DTs analysis method and evaluation of the quality of the proposed method. The quality assessment was conducted as follows: for each model, ROC-AUC, PR-AUC, and F1 scores were calculated [12]. Thus, to objectively assess the effectiveness of the proposed method and compare it with the baseline algorithms, a comprehensive set of widely accepted binary classification metrics was used. The ROC-AUC (Receiver Operating Characteristic – Area Under Curve) metric reflects the area under the receiver operating characteristic curve and characterises the model's ability to distinguish between classes – that is, between atypical and normal user behaviour in the HEI IES at various threshold levels. ROC-AUC indicated the overall quality of object ranking. Given the class imbalance typical for anomaly detection tasks, the PR-AUC (Precision-Recall – Area Under Curve) metric was additionally calculated. It determined the area under the precision-recall curve. PR-AUC made it possible to evaluate the reliability of detecting rare atypical user events in the HEI IES without distortion caused by a large number of negative examples, i.e., normal sessions. To obtain an integrated accuracy estimate at a fixed decision threshold, the F1-score was applied. The F1 metric is the harmonic mean of Precision and Recall. It served as an indicator of the balance between the algorithms' capability to detect all atypical behavioural patterns and the minimisation of false alarms. Additionally, the share of pre-defined "compromised users" of the HEI IES detected by the models was taken into account.
Accordingly, the experiment compared three algorithms for the task of detecting atypical user behaviour patterns in the HEI IES: the supervised XGBoost gradient boosting model, the unsupervised IF detector, and the proposed DTs analysis method. All algorithms (methods) were evaluated using the same set of aggregated user activity features derived from the dataset.

B. Results of Computational Experiments

The main performance indicators on the test sample are presented below. The key results of the experiment are provided in Table I and in Figures 1-9.

TABLE I
EXPERIMENT RESULTS

№	Method	Metrics				
		RA	AP	YT	BF1T	BF1
1	XGB	0.957	0.889	0.18	0.32	0.8
2	ISO	0.701	0.341	0.281	0.28	0.4
3	HYBRID	0.938	0.881	0.357	0.37	0.85

Note: RA is roc_auc, AP is average_precision, YT is youden_threshold, BF1T is best_f1_threshold, BF1 is best_f1

The plot in Figure 1 illustrates the relationship between the true positive rate (TPR) and the false positive rate (FPR) as the decision threshold varies.

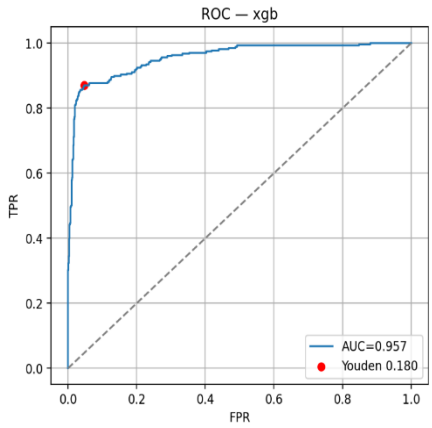

Fig. 1. ROC Curve Visualisation for XGBoost

The result in Figure 1 demonstrated the high sensitivity of the XGBoost model in detecting atypical user behaviour in the HEI IES. The AUC value was 0.957. Thus, on the dataset used, the result confirmed the excellent discriminative capacity of the classifier and its high effectiveness in ranking users by the degree of the risk of atypical behaviour in the HEI IES. The red marker on the curve highlights the point corresponding to the optimal Youden's threshold (Youden's J. statistic) [12, 13]. In this experiment, the optimal probability cut-off value was 0.180. This red point identifies the value at which the difference between sensitivity and the false positive rate (FPR) is maximised.

Figure 2 shows the ROC curve for the unsupervised IF algorithm.

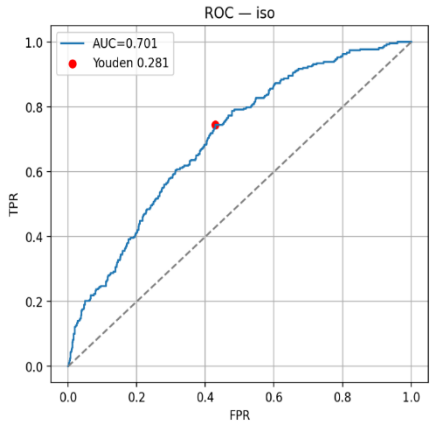

Fig. 2. ROC Curve Visualisation for IF

The AUC value obtained in Figure 2 for the IF algorithm was 0.701. This result indicates a moderate discriminative ability of the anomaly detector for identifying atypical user behaviour in the HEI IES. The flatter shape of the curve, compared with XGBoost, reflects the difficulty of identifying atypical patterns solely on the basis of deviations in data structure without labelled examples. The red marker indicates the optimum according to Youden's criterion. The threshold value was 0.281.

Figure 3 presents the ROC curve of the developed hybrid method.

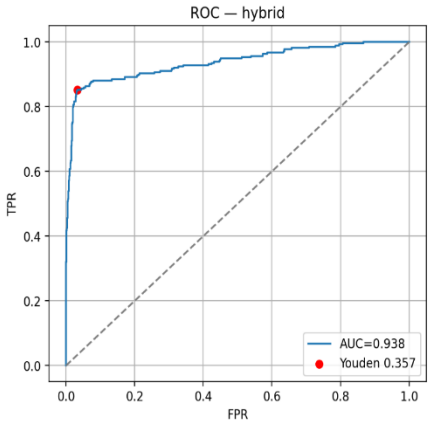

Fig. 3. ROC Curve Visualisation for the Hybrid Method

The AUC value shown in Figure 3 was 0.938. Thus, the result confirmed the high effectiveness of the hybrid method. The method of DTs analysis of user behaviour in the HEI IES retained accuracy comparable to XGBoost while additionally providing resistance to previously unknown forms of atypical user behaviour in the HEI IES. The optimal cut-off threshold determined using Youden's statistic was 0.357 (red marker). This marker represents the recommended operating point for minimising missed atypical events (attacks) at a fixed false alarm rate.

Figures 4, 5, and 6 display the Precision-Recall Curves for XGBoost, IF, and the hybrid method, respectively/

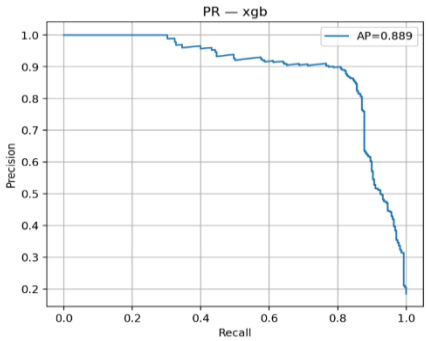

Fig. 4. Precision-Recall Curve for XGBoost

In Figure 4, the Average Precision (AP) value equals 0.889. This result demonstrates the high reliability of the classifier in detecting atypical behaviour of the HEI IES users. The analysis of the curve's shape revealed that XGBoost is capable of achieving a precision level exceeding 0.9. The XGBoost algorithm successfully identified approximately 80% of all the instances of the atypical user behaviour in the HEI IES and, accordingly, exhibited a minimal number of false positives. A sharp decline in precision was observed in the final segment of the curve (see Figure 4), where the Recall value exceeded 0.85. This pattern is typical for the use of the XGBoost algorithm, as it indicates that XGBoost attempted to process complex, borderline cases of atypical HEI IES user behaviour – cases in which the algorithm may mistakenly flag a legitimate HEI IES user, thereby reducing the Precision metric. Several illustrative scenarios are provided below.

Scenario 1 – the night before an exam. A higher education

student (hereinafter HES) rapidly opens a large number of lecture materials on elearn.edu.ua, downloads numerous presentations and lecture-related files, and quickly completes tests in a given discipline. The resulting intensity of requests resembles a DoS attack or scraper activity. Additionally, the time of day acts as a risk marker. However, in reality, the HES is urgently preparing for an examination session.

Scenario 2 – change of context. An educator logs into the system from a home IP address. Fifteen minutes later, the login occurs from an IP address of another country due to enabling VPN for accessing scholarly databases or because of the specifics of mobile Internet connectivity. For XGBoost, such actions resemble the classic "impossible travel" anomaly, representing a variant of atypical behaviour suggestive of session hijacking. In reality, this behaviour constitutes legitimate use of technical tools to secure communication.

Figure 5 shows the PR Curve for IF.

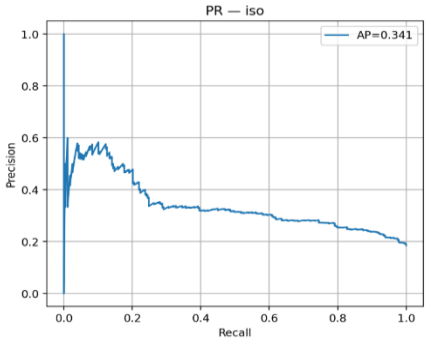

Fig. 5. Precision-Recall Curve for IF

In Figure 5, the Average Precision (AP) equals 0.341. The shape of the curve indicates that the Isolation Forest (IF) is unable to ensure an adequate level of precision. This is due to the limited expressiveness of the IF detector when confronted with complex behavioural patterns exhibited by users in the HEI IES. The AP value in Figure 5 amounted to 0.881, which is comparable to the XGBoost indicator of 0.889. Thus, the result confirms the correctness of the hybrid method's operation. The slight difference in precision is offset by the hybrid method's ability to detect atypical user behaviour in the HEI IES in an unsupervised manner.

Figure 6 presents the Precision–Recall Curve for the proposed hybrid method.

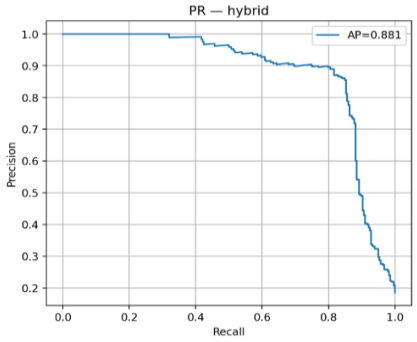

Fig. 6. Precision–Recall Curve for the Hybrid Method

The curve in Figure 6 exhibits a stable plateau of high precision in the Recall range of 0.4–0.8. The hybrid method maintained "Precision" above 0.9, with "Recall" values reaching up to 0.8. This indicates that the system is capable of identifying 80% of all incidents while maintaining an extremely low level of false alarms. The decline in the graph in Fig. 6 for very high recall values (Recall > 0.85) is an expected outcome of the IF component of the method. It is this component that detects suspicious yet ambiguous events that require additional analysis by the HEI IES administrator.

Figure 7 shows the histogram of probability score distribution assigned by the XGBoost model to the test data.

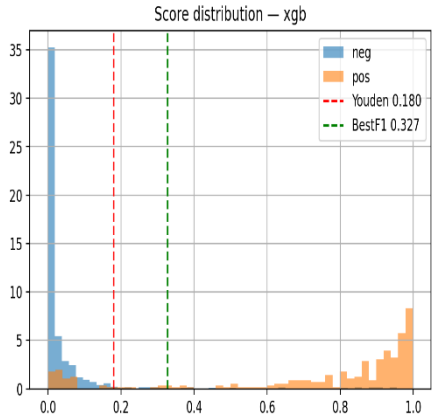

Fig. 7. Score Distribution Histogram Assigned by XGBoost to the Test Data

In Figure 7, the density of normal sessions (neg) is shown in blue, while atypical events (pos) are presented in orange. The histogram demonstrates a bimodal structure, confirming the discriminative capability of the XGBoost classifier. The vast majority of the legitimate HEI IES user actions are concentrated in the low-probability interval [0.0; 0.1], whereas anomalies are shifted to the high-risk interval [0.8; 1.0]. The noticeable sparsity of observations in the central part of the scale [0.3; 0.7] indicates that the XGBoost algorithm minimised the uncertainty zone and effectively separated the classes. The vertical lines in Figure 7 denote the computed optimal decision thresholds: the Youden threshold of 0.180 (red dashed line) and the F1-maximising threshold of 0.327 (green dashed line). The placement of both thresholds in the region of minimal class overlap ensures system robustness to minor data fluctuations and predictable user behaviour when implementing XGBoost in practical analysis of DTs in the HEI IES.

Figure 8 presents the histogram of score distribution generated using the IF algorithm.

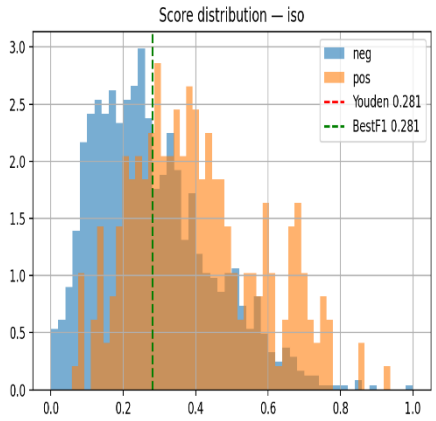

Fig. 8. Score Distribution Histogram Assigned by IF to the Test Data

In Figure 8, unlike XGBoost, we observe substantial overlap between the distributions of normal events (neg, represented by blue bars) and atypical events (pos). This result confirms the low degree of class separation. The highest concentration of normal events is not located near zero but is instead shifted to the 0.2-0.3 interval. This gives us the low confidence in the ability of the IF model to distinguish even legitimate aspects of the HEI IES user behaviour. Both optimal thresholds – the Youden index and the F1-maximising threshold – remain at 0.281. These thresholds lie within a zone of extensive distributional overlap. Thus, the results in Figure 8 confirm the limited capacity of the IF model to detect complex behavioural patterns of the HEI IES users.

Figure 9 shows the histogram of score distribution for the hybrid DTs analysis method.

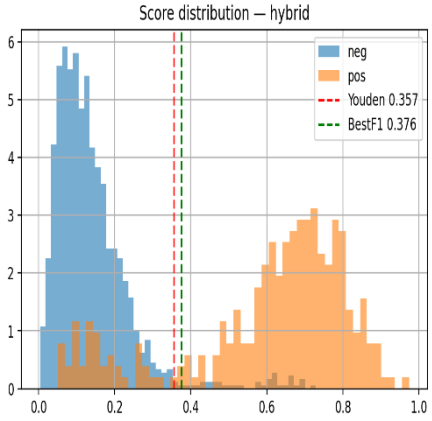

Fig. 9. Score Distribution Histogram Assigned by the Hybrid Method to the Test Data

The distribution in Figure 9 is bimodal. This bimodal structure demonstrates the model's ability to separate the classes effectively. Specifically, normal events (neg), shown as blue bars in Figure 9, are clustered in the low-probability zone [0.0; 0.2], while atypical events (pos), marked by orange bars, are concentrated in the high-risk zone [0.6; 0.9]. The pronounced trough between the neg and pos distributions confirms the ability of the hybrid method in classifying events. The optimal thresholds are: Youden – 0.357 (red dashed line) and F1-measure – 0.376 (green dashed line). Their placement within a narrow zone of minimal class overlap indicates that the hybrid method preserves the discriminative quality of XGBoost while ensuring a balance between predictive accuracy and robustness to rare cases of dangerous behaviour among the HEI IES users.

To summarise the experimental validation of the proposed hybrid method, we note the following. The XGBoost algorithm demonstrated the best ranking of the HEI IES users by risk level according to the ROC metric, achieving AUC = 0.957 and Average Precision = 0.888. The IF algorithm performed significantly worse, with an ROC AUC of approximately 0.70 and AP of about 0.34, which is attributable to the limited expressiveness of a simple unsupervised detector when faced with complex behavioural patterns in the HEI IES environment. The hybrid method, by contrast, exhibited high effectiveness: its ROC AUC was nearly 0.938, AP = 0.8806, and Best-F1 was approximately 0.85. Although the hybrid method demonstrated a slight decline in ROC AUC compared to the fully supervised XGBoost algorithm, it remained close to XGBoost across all major metrics and substantially outperformed IF. This confirms the research paradigm that combining data from supervised and unsupervised sources yields a more accurate estimation of atypical behaviour risk. Thus, the hybrid method mitigated the false positives of XGBoost and compensated for the insufficient sensitivity of IF to previously unseen atypical behavioural patterns during the HEI IES operation.

It should be noted that when interpreting F1-threshold metrics, it is advisable to consider the objective of the study. In the conducted experiments, the F1 metrics achieved high values (>0.85) for both the XGBoost algorithm and the proposed hybrid method. However, for practical tasks involving the detection of atypical user behaviour, the ROC AUC metric is more informative. It is specifically the ROC AUC indicator that enables an information security analyst of the HEI IES to determine whether a detected threat is real. According to this metric, the hybrid method also demonstrated operational suitability: its ROC AUC value remained high at approximately 0.938.

The scientific novelty of the study lies in the development and validation of a hybrid method for analysing digital traces (DTs) of HEI IES users. The method integrates: the hierarchical structure of IES processes; graph-cluster analysis of behavioural relationships; and risk assessment of atypical user actions based on the hybridisation of ML methods. Unlike approaches that predominantly rely on classification or isolation models alone, the proposed method integrates structural, behavioural, and statistical characteristics of user activity. This ensures high-quality detection of atypical actions and allows the multi-layered nature of user behaviour scenarios in the HEI IES to be taken in.

The proposed method and its software implementation are used as a behavioural monitoring module of the IES at the National University of Life and Environmental Sciences of Ukraine. The method and its accompanying software can potentially operate as an add-on to existing IES logging systems. When integrated into the IES of Ukrainian HEIs, the method enables the prioritisation of high-risk events, thereby reducing the workload on IES analysts. Additionally, the method may be used to study interactions between higher education students and the digital resources of the HEI IES.

VI. Research Limitations

The limitations of the study were primarily associated with the dependence of the hybrid method on the quality of the preliminary structuring of the digital traces. This is due to the fact that any changes in IES log formats or event registration procedures would require additional adjustments. The hybrid behavioural risk assessment relied on statistical and machine-learning characteristics of individual events, which enabled to achieve high accuracy. However, emotional, social, and motivational factors that may also influence user behaviour were not taken into account.

The study did not include deep neural network architectures, which are therefore considered a direction for further improvement of the method's accuracy through the application of more sophisticated models. Another limitation is that the data (logs) were collected from a single IES, which affected the

generalisability of the results to other higher education institutions with structurally different processes.

Additionally, the present study did not employ autoencoders [27], recurrent neural networks [28], or graph neural networks [29], as their applicability within a university information and educational system (IES) is limited. This is due to the fact that autoencoders require a substantial volume of homogeneous data, stability of statistical properties, and a large number of training cycles. The latter contradicts the nature of events in an IES, where the structure of the trace changes depending on academic semesters, workload, and organisational processes.

Recurrent neural networks presuppose the availability of long temporal sequences for each user. However, the actual digital traces of higher education students are characterised by fragmentation, irregularity, and uneven temporal distribution throughout the academic year. Consequently, these conditions did not allow for the construction of complete time series.

Graph neural networks, despite their potential effectiveness in modelling network structures, require a stable and coherent graph representation. Yet events within an IES generate incomplete and context-dependent graphs that demand the development of complex pre-processing procedures.

Given these constraints, the use of complex pre-processing procedures classical anomaly detection methods combined with structural-behavioural modelling proved to be more effective and more representative of the real data provided by HEI IES.

Conclusion

A hybrid multi-level method for analysing the digital traces (DTs) of users of information and educational systems (IES) in higher education institutions (HEIs) has been developed. The method integrates hierarchical structural modelling of events, graph-cluster analysis of behavioural interactions, and risk assessment based on a combination of classification and isolation models. Implemented as a software application built upon the models incorporated in the hybrid method, it enabled the transition from the raw HEI IES log-based DTs to generalised behavioural profiles of the users. The method provided the possibility to factor in both frequent and rare activity patterns in the HEI IES environment.

The computational experiment conducted on the IES dataset demonstrated the following: 1. The developed method ensured high accuracy and stability in detecting atypical user actions. 2. The results showed that the hybrid method preserved the discriminative power of the supervised gradient boosting algorithm while simultaneously compensating for its vulnerability to rare behavioural scenarios through the integration of isolation-based analytical components. 3. According to ROC-AUC, PR-AUC, and F1-score metrics, the method proposed in the article approached the best values of the baseline XGBoost model. Moreover, the hybrid method provided improved stability to the behavioural heterogeneity of the IES users.

A comparative evaluation with the baseline XGBoost and Isolation Forest methods confirmed the viability of integrating supervised and unsupervised information sources, as this enabled the generation of a reliable risk assessment of behavioural atypicality based on the analysed digital traces. The conducted sensitivity analysis with respect to different types of atypical events demonstrated the method's suitability for the use in operational HEI IES environment.

References

[1] A. Vaccari, F. Calabrese, B. Liu, and C. Ratti, "Towards the SocioScope", in *17th ACM SIGSPATIAL International Conference*, Seattle, Washington, Nov. 4–6, 2009. New York, New York, USA: ACM Press, 2009, pp. 52–61. [Online]. Available: doi: 10.1145/1653771.1653782

[2] B. T. Pentland, J. Recker, J. Wolf, and G. Wyner, "Bringing Context Inside Process Research with Digital Trace Data", *Journal of the Association for Information Systems*, vol. 21, no. 5, pp. 1214–1236, Sep. 2020. [Online]. Available: doi: 10.17705/1jais.00635

[3] T. Grisold, W. Kremser, J. Mendling, J. Recker, J. vom Brocke, and B. Wurm, "Generating Impactful Situated Explanations through Digital Trace Data", *Journal of Information Technology*, vol. 39, no. 1, pp. 2–18, Oct. 2023. [Online]. Available: doi: 10.1177/02683962231208724

[4] J. Howison, A. Wiggins, and K. Crowston, "Validity Issues in the Use of Social Network Analysis with Digital Trace Data", *Journal of the Association for Information Systems*, vol. 12, no. 12, pp. 767–797, Dec. 2011. [Online]. Available: doi: 10.17705/1jais.00282

[5] M. Ollagnier–Beldame, "The use of digital traces: a promising basis for the design of adapted information systems?", *International Journal on Computer Science and Information Systems. Special Issue "Users and Information Systems"*, no. 6(2), 2011, Art. no. hal-00669019. [Online]. Available: https://hal.science/hal-00669019v1/document

[6] H. Hamooni, B. Debnath, J. Xu, H. Zhang, G. Jiang, and A. Mueen, "LogMine: Fast pattern recognition for log analytics", in *CIKM'16: the 25th ACM Conference on Information and Knowledge Management*, Indianapolis Indiana USA. New York, NY, USA: ACM, 2016, p. 1573–1582. [Online]. Available: doi: 10.1145/2983323.2983358

[7] I. Arnaldo, A. Cuesta-Infante, A. Arun, M. Lam, C. Bassias, and K. Veeramachaneni, "Learning Representations for Log Data in Cybersecurity", in *Lecture Notes in Computer Science*. Cham: Springer Int. Publishing, 2017, pp. 250–268. [Online]. Available: doi: 10.1007/978-3-319-60080-2_19

[8] J. Bogatinovski, G. Madjarov, S. Nedelkoski, J. Cardoso, and O. Kao, "Leveraging Log Instructions in Log-based Anomaly Detection", in *2022 IEEE International Conference on Services Computing (SCC)*, Barcelona, Spain, Jul. 10–16, 2022. IEEE, 2022. [Online]. Available: doi: 10.1109/scc55611.2022.00053

[9] H. Guo, S. Yuan and X. Wu, "LogBERT: Log Anomaly Detection via BERT", in *2021 International joint conference on neural networks (IJCNN)*, Mar 7, 2021. IEEE, 2021, pp. 1-8. [Online] Available: doi: 10.48550/arXiv.2103.04475

[10] X. Shi, Y. D. Wong, M. Z.-F. Li, C. Palanisamy, and C. Chai, "A feature learning approach based on XGBoost for driving assessment and risk prediction", *Accident Analysis & Prevention*, vol. 129, pp. 170–179, Aug. 2019. [Online]. Available: doi: 10.1016/j.aap.2019.05.005

[11] W. Wang et al., "A User Purchase Behavior Prediction Method Based on XGBoost", *Electronics*, vol. 12, no. 9, p. 2047, Apr. 2023. [Online]. Available: doi: 10.3390/electronics12092047

[12] Y. Zhang et al., "Comparison of Prediction Models for Acute Kidney Injury Among Patients with Hepatobiliary Malignancies Based on XGBoost and LASSO-Logistic Algorithms", *International Journal of General Medicine*, Volume 14, pp. 1325–1335, Apr. 2021. [Online]. Available: doi: 10.2147/ijgm.s302795

[13] R. Aznar-Gimeno, L. M. Esteban, G. Sanz, and R. del-Hoyo-Alonso, "Comparing the Min–Max–Median/IQR Approach with the Min–Max Approach, Logistic Regression and XGBoost, Maximising the Youden Index", *Symmetry*, vol. 15, no. 3, p. 756, Mar. 2023. [Online]. Available: doi: 10.3390/sym15030756

[14] T. Al-Shehari, M. Al-Razgan, T. Alfakih, R. A. Alsowail, and S. Pandiaraj, "Insider Threat Detection Model using Anomaly-Based Isolation Forest Algorithm", *IEEE Access*, p. 118170–118185, 2023. [Online]. Available: doi: 10.1109/access.2023.3326750

[15] N. Fang, X. Fang, and K. Lu, "Anomalous Behavior Detection Based on the Isolation Forest Model with Multiple Perspective Business Processes", *Electronics*, vol. 11, no. 21, p. 3640, Nov. 2022. [Online]. Available: doi: 10.3390/electronics11213640

[16] L. Sun, S. Versteeg, S. Boztas and A. Rao, "Detecting Anomalous User Behavior Using an Extended Isolation Forest Algorithm: An Enterprise Case Study", *ArXiv*, 2016. Available: doi: 10.48550/arXiv.1609.06676

[17] D. Azcona, I.-H. Hsiao, and A. F. Smeaton, "Detecting students-at-risk in computer programming classes with learning analytics from students' digital footprints", *User Modeling and User-Adapted Interaction*, vol. 29, no. 4, pp. 759–788, Apr. 2019. [Online]. Available: doi: 10.1007/s11257-019-09234-7

[18] M. E. Buitrago-Ropero, M. S. Ramírez-Montoya, and A. C. Laverde, "Digital footprints (2005–2019): a systematic mapping of studies in education", *Interactive Learning Environments*, pp. 1–14, Sep. 2020. [Online]. Available: doi: 10.1080/10494820.2020.1814821

[19] V. Lakhno et al., "Analysis of digital footprints associated with cybersecurity behavior patterns of users of University Information and Education Systems", *International Journal of Electronics and Telecommunications*, pp. 673–682, Jul. 2024. [Online]. Available: doi: 10.24425/ijet.2024.149596

[20] A. Nugumanova, M. Mansurova M., and Y. Baiburin, "Exploration of student behavior patterns through digital footprints", *Journal of Mathematics, Mechanics and Computer Science*, vol. 103, no. 3, Oct. 2019. [Online]. Available: doi: 10.26577/jmmcs-2019-3-25

[21] Y. Sürmelioğlu and S. S. Seferoğlu, "An examination of digital footprint awareness and digital experiences of higher education students", *World Journal on Educational Technology: Current Issues*, vol. 11, no. 1, pp. 48–64, Feb. 2019. [Online]. Available: doi: 10.18844/wjet.v11i1.4009

[22] D. S. Dolliver, A. K. Ghazi-Tehrani, and K. T. Poorman, "Building a robust cyberthreat profile for institutions of higher education: An empirical analysis of external cyberattacks against a large University's computer network", *International Journal of Law, Crime and Justice*, vol. 66, p. 100484, Sep. 2021. [Online]. Available: doi: 10.1016/j.ijlcj.2021.100484

[23] M. A. Haque et al., "Cybersecurity in Universities: An Evaluation Model", *SN Computer Science*, vol. 4, no. 5, p. 569, Jul. 2023. [Online]. Available: doi: 10.1007/s42979-023-01984-x

[24] N. S. Fouad, "Securing higher education against cyberthreats: from an institutional risk to a national policy challenge", *Journal of Cyber Policy*, pp. 1–18, Sep. 2021. [Online]. Available: doi: 10.1080/23738871.2021.1973526

[25] G. Obadi *et al.*, "Using Spectral Clustering for Finding Students' Patterns of Behavior in Social Networks", in *DATESO 2010*, Štědronín-Plazy, Czech Republic, Apr 21-23, 2010. Prague: MATFYZPRESS, 2010, p. 118-130. Available: https://ceur-ws.org/Vol-567/dateso2010.pdf

[26] I. Ben-Gal, S. Weinstock, G. Singer, and N. Bambos, "Clustering Users by Their Mobility Behavioral Patterns", *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 13, no. 4, pp. 1–28, Aug. 2019. [Online]. Available: doi: 10.1145/3322126

[27] R. Bouman and T. Heskes, "Autoencoders for Anomaly Detection are Unreliable", *ArXiv*, 2025. Available: doi: 10.48550/arXiv.2501.13864

[28] B. Lindemann, B. Maschler, N. Sahlab, and M. Weyrich, "A survey on anomaly detection for technical systems using LSTM networks", *Computers in Industry*, vol. 131, p. 103498, Oct. 2021. [Online]. Available: doi: 10.1016/j.compind.2021.103498

[29] Z. Feng, R. Wang, T. Wang, M. Song, S. Wu, and S. He, "A Comprehensive Survey of Dynamic Graph Neural Networks: Models, Frameworks, Benchmarks, Experiments and Challenges," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–20, 2025. [Online]. Available: doi: 10.1109/tkde.2025.3621291