Development of a smart personnel security system using machine learning

Maksym Bilychenko[1,*,†], Nataliia Kasianova[1,†], Serhii Smerichevskyi[1,†] and Igor Kryvovyazyuk[2,‡]

[1]The State University "Kyiv Aviation Institute", 1 Liubomyra Huzara ave., Kyiv, 03058, Ukraine
[2]Lutsk National Technical University, 75 Lvivska str., Lutsk, 43018, Ukraine

Abstract

Insider threats remain one of the most challenging aspects of organizational security, particularly in the era of digital transformation and widespread remote access to sensitive data. This study proposes a machine learning–based approach to personnel security that combines Isolation Forest and Local Outlier Factor algorithms with behavioral features enhanced through the use of large language models (LLMs). To improve detection accuracy, user web activity was classified using LLM-generated labels derived from website content analysis. Experimental results demonstrate strong model performance in identifying insider activity at the user level, with high detection accuracy and minimal false classifications. In addition, time-to-detection analysis revealed that most insider threats were identified before or shortly after the onset of malicious behavior. The findings suggest that the proposed system is not only effective in capturing behavioral anomalies but also feasible for real-time deployment in enterprise environments.

Keywords

Insider threat detection, personnel security, anomaly detection, large language models, isolation forest, local outlier factor, behavioral profiling

1. Introduction

Digital technologies have rapidly permeated nearly all domains of social life, including education, mass media, the workplace, daily routines, commerce, sports, healthcare and entertainment. The development of a digital society entails the creation of a complex yet thoroughly transformed ecosystem in which humans and technologies coexist in new forms of interaction and cohabitation. The integration of information and communication technologies (ICT), artificial intelligence, the Internet of Things (IoT), and cloud computing has facilitated the emergence of an interconnected digital environment. Within this environment, business processes can be executed with minimal time and resource expenditure. Digital transformation initiatives are progressively focused on automating repetitive tasks, leveraging big data analytics to inform strategic management decisions, expanding electronic commerce, and strengthening customer interaction channels. Nevertheless, digital transformation remains a complex, resource-intensive, and time-consuming undertaking for most enterprises – particularly for small and medium-sized businesses. In this context, it becomes imperative for companies to articulate clear digital strategies and define priority areas for transformation. Based on these priorities, enterprises must adapt to ongoing changes and implement innovative solutions across both technological and methodological dimensions.

Simultaneously, the advancement of digital transformation introduces not only new opportunities but also a range of emerging risks and threats. These challenges are inherent both to the transformation

CH&CMiGIN-25: Fourth International Conference on Cyber Hygiene & Conflict Management in Global Information Networks, June 20–22, 2025, Kyiv, Ukraine
*Corresponding author.
†These authors contributed equally.
mbilich9@gmail.com (M. Bilychenko); nat_kas@ukr.net (N. Kasianova); s_f_smerichevsky@ukr.net (S. Smerichevskyi); krivovyazukigor@gmail.com (I. Kryvovyazyuk)
0000-0003-4657-1039 (M. Bilychenko); 0000-0001-7729-2011 (N. Kasianova); 0000-0003-2102-1524 (S. Smerichevskyi); 0000-0002-8801-4700 (I. Kryvovyazyuk)

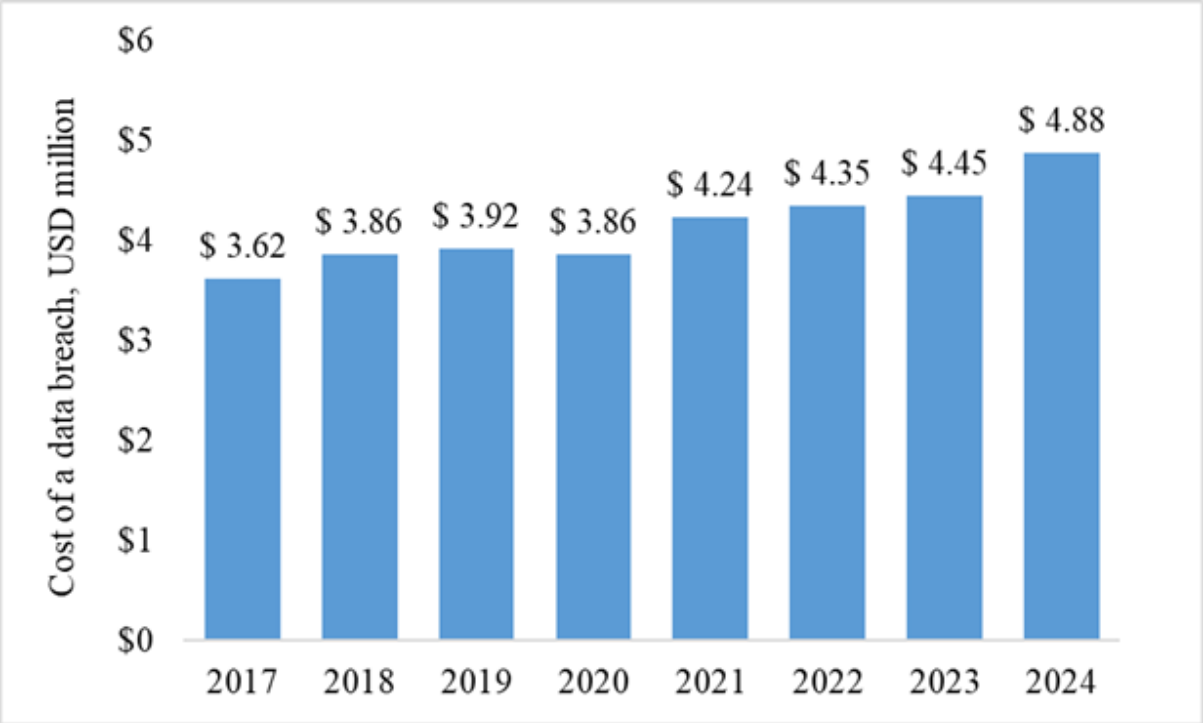

Figure 1: The global average cost of a data breach in 2017-2024, USD million. Source: compiled by the author based on [2]

process itself and to the subsequent digital operations of an organization. As noted in a report by Deloitte [1], one of the key categories of risk involves the exposure of confidential information and data, that is typically associated with improper handling of personal or sensitive information related to customers, employees or business partners. Data breaches, violations of data processing protocols and inadequate storage practices throughout the data lifecycle constitute primary sources of these threats. Moreover, low levels of digital literacy among staff and a lack of organizational culture around information security can generate systemic vulnerabilities.

Insider threats, social engineering attacks, and careless handling of sensitive information frequently lead to serious security incidents. According to IBM's Cost of a Data Breach report [2], the average global cost of mitigating a data breach reached $4.88 million in 2024 – the highest recorded figure in the past eight years. As illustrated in Figure 1, this represents a 10% year-over-year increase, marking one of the most significant annual spikes during the observed period. Additionally, the 2025 Thales Data Threat Report revealed that 45% of surveyed companies worldwide experiencing a data breach, with 14% of these occurring within the past year [3].

Human-related risks have become especially important for enterprises undergoing digital transformation. Ensuring personnel security is essential for stable business operations and requires a structured approach that includes developing employees' digital skills, improving human resource policies, and applying up-to-date cybersecurity standards in day-to-day operations [4]. To achieve this, companies need to put in place clear digital safety policies for employees, regularly monitor how staff handle data, limit access to sensitive information based on user roles and follow modern procedures for data handling and protection. A key part of personnel security is the ability to detect possible threats early – especially identifying employees who might, whether intentionally or by mistake, share confidential business information outside the organization. In other words, timely detection of insider threats and appropriate follow-up actions are critical to protecting enterprise security. In other words, timely detection of insider threats and appropriate follow-up actions are critical to protecting enterprise security. Given the growing complexity and scale of digital environments, traditional control mechanisms are no longer sufficient. This highlights the need for advanced, data-driven approaches capable of identifying subtle

behavioral patterns that may indicate insider risk. The following section provides a review of current methodologies and recent advances in insider threat detection, with a particular focus on algorithmic and ML-driven solutions.

2. Related Work

Traditionally, personnel security has been understood as a set of measures aimed at ensuring the stability of human capital, maintaining adequate employee qualifications, fostering effective motivation and preventing internal threats. However, the rapid advancement of digital technologies necessitates a reassessment of existing approaches to evaluating and forecasting the state of this subsystem. In particular, the use of modern modeling techniques is gaining relevance – ranging from score-based assessments of employee motivation to hierarchical models built on fuzzy logic. These approaches enable the analysis of how changes in motivation systems, training programs, and personnel management influence the likelihood of labor-related risks, while also identifying the most vulnerable areas.

Beyond conventional strategies for ensuring personnel security, insider threats have become especially critical in the digital economy. Despite the availability of extensive software solutions designed to guard against external intrusions, internal threats – posed by employees or individuals with privileged access – remain the most challenging to detect and mitigate. Latest research increasingly shifts from general personnel system management to the development of preventive analytics tools capable of identifying risky behavior before it leads to serious consequences. In this context, particular attention has been directed toward the use of data mining and machine learning methods for building adaptive personnel risk management systems. These systems analyze extensive digital traces left by employees such as activity within internal networks, access to information resources, and behavioral pattern shifts, to detect early signs of potential insider threats with high accuracy.

Insider threat detection systems based on anomaly detection techniques rely on statistical modeling of typical employee interactions with IT resources. Deviations from established behavioral patterns may signal potential personnel-related threats. For instance, M. Raissi-Dehkordi and D. Carr [5] proposed a multi-perspective insider threat detection framework that integrates data from various components of the corporate network. This architecture employs a one-class Support Vector Machine algorithm, which reduces the number of false positives and enhances the accuracy of identifying risky employee behavior. While this method performs well in detecting group-based insider activities, its effectiveness in forecasting individual insider threats remains limited.

An alternative approach was introduced by T. Rashid and colleagues [6], who utilized Hidden Markov Models to model habitual employee behavior. These models are well-suited for analyzing temporal sequences but exhibit high computational complexity as model size increases. Another technique, developed by Y. Song et al. [7], applies Gaussian Mixture Models to construct user profiles and detect anomalies. Although this method demonstrates high accuracy, it is primarily geared toward biometric identification rather than targeted detection of insider threats within corporate systems.

In contrast, G. Gavai and colleagues [8] proposed a behavior-based insider threat detection model using digital traces of employee activity. Their framework defined 42 features across five categories: email content and usage, login patterns, and software and web activity. The model applies both Isolation Forest and Random Forest algorithms to detect abnormal behavior and identifies high-risk individuals through an interactive visualization dashboard. This approach supports proactive decision-making in personnel security by offering simplicity, adaptability, and objectivity. However, its performance may degrade in large-scale organizational settings, requiring additional tuning for stability.

In the study by D. C. Le and N. Zincir-Heywood [9], a framework for insider threat detection was introduced based on an ensemble of unsupervised learning algorithms, including Isolation Forest, Autoencoder, Local Outlier Factor (LOF), and Lightweight Online Detector of Anomalies (LODA). Among these, the Autoencoder and LOF models yielded the highest detection performance. Additionally, the use of ensemble voting across models was shown to enhance both the accuracy and robustness of the detection system.

Table 1
Comparative overview of recent studies utilizing modern machine learning methods for insider threat detection

Author & Year	Models	Data	Summary
M. Raissi-Dehkordi & D. Carr, 2011 [5]	One-Class SVM	Simulated data via OPNET	High accuracy for group attacks; moderate performance for individual threats.
T. Rashid et al., 2016 [6]	Hidden Markov Models	CERT	Model is easy to train but becomes computationally intensive with scale.
Y. Song et al., 2013 [7]	Gaussian Mixture Models	RUU Research Dataset	High accuracy, though primarily focused on biometric profiling.
G. Gavai et al., 2015 [8]	Isolation Forest, Random forests	Vegas	Isolation Forest outperformed other models in terms of detection effectiveness.
D. C. Le & N. Zincir-Heywood, 2021 [9]	Isolation Forest, Autoencoder, LOF, LODA	CERT R4.2, R6.2	Autoencoder and LOF achieved the highest detection accuracy.
T. Al-Shehari et al., 2024 [10]	Density-Based Local Outlier Factor (LOF)	CERT R4.2	High accuracy for imbalanced datasets; interpretability remains limited.
C. Song et al., 2024 [11]	LLM-based Multi-Agent System (Audit-LLM)	CERT R4.2, R5.2, PicoDomain	High detection accuracy and improved explainability; scalability constraints noted.

T. Al-Shehari et al. [10] proposed a detection approach using the Density-Based Local Outlier Factor (DBLOF), specifically adapted for imbalanced datasets such as CERT R4.2. Unlike traditional methods, DBLOF focuses on analyzing local density, enabling the precise identification of rare yet potentially harmful behavioral anomalies. The method demonstrated high accuracy and proved effective in detecting insider threats under real-world corporate conditions.

C. Song and colleagues [11] introduced Audit-LLM – a multi-agent insider threat detection system that analyzes event logs using large language models (LLMs). To overcome the limitations of LLMs in processing complex activity logs, the system incorporates three specialized agents alongside a multi-agent debate mechanism designed to improve inference accuracy. The approach demonstrated superior detection performance compared to existing solutions, particularly in terms of explanatory power and handling large-scale log files.

In summary, the most prevalent and effective approaches to insider threat detection within personnel security systems are based on the use of Isolation Forest and Local Outlier Factor algorithms. Their effectiveness lies in analyzing behavioral data collected from corporate networks, such as event logs, access records, and system activity. These algorithms have demonstrated high performance in identifying anomalous employee behavior that may indicate internal security threats.

However, much of the existing research primarily focuses on optimizing model accuracy while paying limited attention to the practical challenges of deploying such solutions in real organizational settings. Key challenges, including the adaptability of detection models to evolving operational environments, the scalability of solutions across enterprise contexts, and their seamless integration into existing IT infrastructures, remain underexplored in the current body of research. Furthermore, the potential of emerging technologies such as large language models (LLMs) to enhance system scalability, computational efficiency, and the interpretability of detection outcomes has received limited scholarly attention. Given these limitations, the present research will focus on combining Isolation Forest and LOF algorithms with LLM-based techniques to develop a modern, adaptive, accurate and scalable insider threat detection system. This approach aims to move beyond academic metrics and prioritize practical applicability in real-world enterprise environments.

3. Data and methodology

Building upon the insights from the literature and addressing the identified research gaps, this study proposes a hybrid insider threat detection approach that integrates both established anomaly detection techniques and recent advancements in data-driven modeling. Specifically, two algorithms were selected for model development: Isolation Forest and Local Outlier Factor (LOF), both of which have demonstrated high accuracy in previous studies while offering complementary strengths in detecting anomalous behavior.

The LOF algorithm is based on the concept of local density. Unlike global models that assess each data point in relation to the entire dataset, LOF evaluates an observation in the context of its immediate neighborhood. The underlying assumption is that an anomalous instance will exhibit significantly lower local density compared to its neighbors — indicating its presence in a relatively sparse region of the feature space. Mathematically, the LOF algorithm is implemented through several sequential steps. The first involves calculating the distance to the k-nearest neighbors k-distance(p).

This parameter defines the radius of the local neighborhood around an observation p within which the analysis is conducted. Subsequently, the algorithm calculates the reachability distance, a measure designed to prevent unrealistically low density estimates that may result from the presence of extremely close neighboring points.

$$\text{reach\_dist}_k(p, o) = \max\left(k\text{-}distance(o), dist(p, o)\right) \tag{1}$$

The next step involves computing the local reachability density (lrd) of the observation:

$$lrd_k(p) = \left(\frac{\sum_{o \in N_k(p)} \text{reach\_dist}_k(p, o)}{\mid N_k(p) \mid}\right)^{-1} \tag{2}$$

where $N_k(p)$ is the collection of the *k* closest points to *p*, based on a distance metric which is typically Euclidean distance. This metric quantifies how densely the point *p* is surrounded by its *k*-nearest neighbors. In the final step, the LOF score is computed:

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{\mid N_k(p) \mid} \tag{3}$$

The LOF score represents the ratio of the average local reachability density of a point's neighbors to the local reachability density of the point itself. Formally, if $LOF_k(p) \approx 1$, the behavior of point *p* is considered normal. However, if $LOF_k(p) \gg 1$, the point is identified as anomalous, indicating that its local density is significantly lower than that of its surrounding neighbors.

The second algorithm selected for this study is Isolation Forest. Unlike most traditional methods that model normal behavior and identify deviations, Isolation Forest is based on the principle of isolation. The core idea is that anomalous instances are more susceptible to isolation – they tend to be located further away from dense clusters of points and can therefore be separated more quickly through recursive partitioning of the feature space. The model consists of an ensemble of isolation trees (iTrees). The construction of each isolation tree involves the random selection of a feature, followed by the choice of a split value within the domain of that feature. This recursive partitioning continues until the instance is fully isolated in a leaf node, with the resulting isolation depth serving as an empirical indicator of separability – where shorter path lengths typically correspond to anomalous instances due to their greater isolation from the general data distribution.

The overall structure of the algorithm can be summarized as follows: the isolation forest builds a tree-based structure capable of efficiently isolating each individual instance (see Figure 2). Due to the sensitivity of this method to isolation characteristics, outliers tend to appear closer to the root of the tree, while normal points are usually isolated deeper in the tree. This recursive isolation mechanism underpins the effectiveness of the approach, and the resulting structure is referred to as an isolation tree

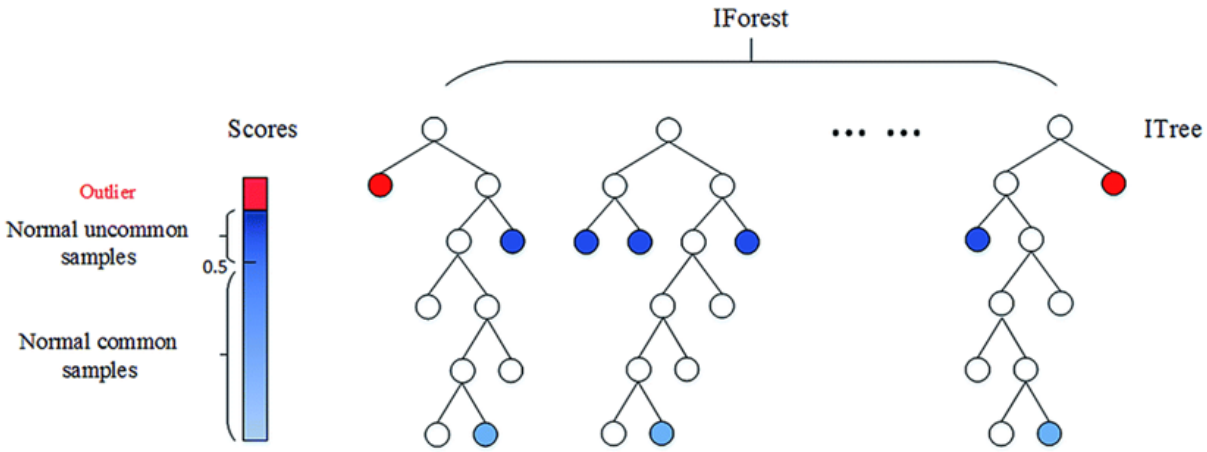

Figure 2: General architecture of the Isolation Forest algorithm. Source: [12]

or iTree. An isolation forest consists of an ensemble of iTrees built from the dataset, where anomalies are identified as those instances with a significantly shorter average path length across the trees.

The mathematical formalization of the Isolation Forest algorithm can be introduced as follows. In the first step, the expected path length for isolating a data point x in a tree constructed from n observations is computed:

$$E[h(x)] \approx log_2(n) + \gamma \tag{4}$$

where $h(x)$ denotes the path length required to isolate the point x, and γ is a correction constant (the Euler's constant). Following this, the anomaly score is calculated as:

$$s(x, n) = 2^{\frac{-E[h(x)]}{c(n)}} \tag{5}$$

where $c(n)$ represents the average path length, which, for a binary search tree, can be approximated as follows:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \tag{6}$$

where $H(i)$ denotes the i-th harmonic number, defined as the sum of the reciprocals of the first i positive integers.

The interpretation of results obtained using the Isolation Forest algorithm is based on the anomaly score $s(x)$, which ranges from 0 to 1. Values approaching 1 (i.e., $s(x) \to 1$) indicate a high likelihood that the observation is anomalous, suggesting that the behavior of the corresponding instance significantly deviates from the norm observed in the overall dataset. Conversely, values of $s(x) < 0.5$ are indicative of typical, non-anomalous behavior consistent with the majority of the observations. Scores near 0.5 require further investigation, as they may represent borderline or potentially risky cases that have not yet escalated into overt violations.

Due to the limited availability of actual corporate data for insider threat research, this study utilizes the publicly accessible CERT Insider Threat Dataset [13]. Although the CERT dataset does not originate from real-world organizational environments, it is synthetically generated to support the development and evaluation of insider threat detection models and the analysis of personnel security risks. Despite its artificial nature, the dataset incorporates behavioral patterns that closely resemble real employee activities within an organizational information system. The CERT Insider Threat Tools include logs of employee interactions with computer systems, complemented by selected organizational attributes. The dataset structure consists of relational tables containing attributes such as user identifiers, event timestamps, and descriptions of observed actions. For this study, version R4.2 of the CERT dataset was selected, as it provides a sufficient number of observations for both normal users and insider threat cases. Specifically, this version includes data on approximately 1000 employees, among whom about 70 are labeled as potential insider threats to organizational personnel security.

The first step of the analysis involved constructing a focused sample of users for more detailed investigation. Since the full dataset is quite large and contains more information than necessary for initial model development, a decision was made to reduce its size while preserving a balanced distribution between classes. Specifically, 20 out of the 70 users flagged as insider threats were selected, along with 30 users from the remaining 930 who exhibited no suspicious behavior. This resulted in a working sample of 50 employees, with insiders making up 40% of the group. Class balance was a key factor in building this sample. In the original dataset, insider cases were relatively rare accounting for less than 10% of all observations. Using the original distribution would have led to a strong class imbalance, which could hinder the model's ability to detect insider threats effectively. By selecting a more balanced subset, we aimed to improve the model's capacity to learn from and identify unusual behavioral patterns.

A key innovation in this study is the enhancement of existing behavioral data through the integration of additional features generated by large language models. This approach aims to increase the informativeness of features relevant to personnel security analysis, particularly in relation to employees' web activity. The analysis focuses on two primary risk scenarios:

1. Visiting job search websites, which may be linked to attempts to take sensitive data when changing jobs.
2. Accessing websites that could be used to share confidential company information without permission.

Traditional threat classification methods often rely on fixed lists of domains (such as job search or data leak sites), but these are inflexible and cannot detect new or unlisted threats. To address this, we propose a content-based approach that uses the text of visited websites to determine their type. Based on the *content* and *url* fields in the CERT dataset, we created two binary indicators, which were evaluated using LLM prompts to classify the intent behind each web visit. This LLM-driven feature enrichment allows for more nuanced and adaptive threat identification beyond rigid rule-based systems.

Given the large volume of data it was essential to select a fast and resource-efficient model for generating LLM-based responses. For this purpose, the Phi-3.5-mini-instruct model [14] was chosen. With fewer than 4 million parameters, this lightweight model can be deployed locally on a personal computer, enabling high-speed processing without significant loss in output quality. To further enhance processing throughput, the vLLM library [15], a state-of-the-art tool for optimized text generation, was integrated into the pipeline. In total, more than one million web activity records were processed within approximately five hours using GPU resources provided by Google Colab A100 [16]. As a result, two new variables were generated based on the model's responses indicating whether the user was likely searching for a job and whether the user's activity suggested potential information leakage.

One of the key steps in feature engineering for modeling personnel security threats was the segmentation of employee activity based on time, distinguishing between actions occurring during standard working hours and those outside of them. To support this, an analysis of user activity distribution across the 24-hour period was conducted, allowing for the identification of core working hours (see Figure 3). Based on this, two separate groups of variables were created: one representing activity within working hours and the other representing activity beyond that period. This approach enabled differentiation of behavioral patterns and contributed to improved accuracy in anomaly detection.

Particular attention was given to the development of novel variables not previously used in insider threat research. For example, the variable *num_site_jobs_1_growth_total_40_days* was introduced to capture the maximum increase in visits to job search websites over the past 40 days. This feature was motivated by the assumption that seeking new employment is typically a prolonged process and not limited to a short time window, which may significantly elevate the risk of insider activity. Another important variable, *num_connect_after_job_search*, was designed to reflect the multiplicative interaction between the increase in the number of connected devices and the growth in job search activity over the same 40-day period. The rationale behind this feature is to detect a specific behavioral pattern in which an employee first intensifies job search efforts and subsequently connects additional devices to transfer or store sensitive corporate data, indicating high likelihood of data leakage. The final dataset comprised

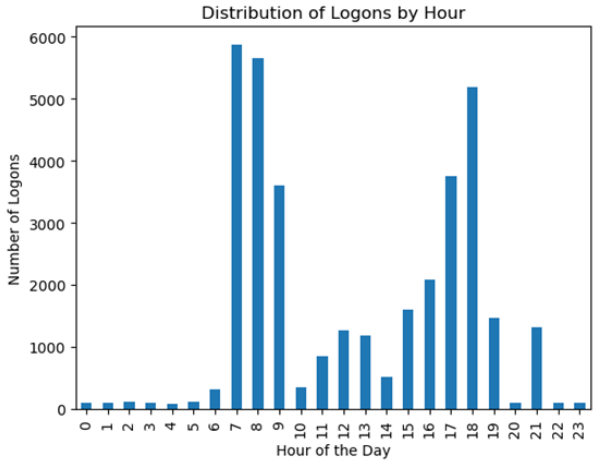

Figure 3: Distribution of user logins by hour of the day. Source: compiled by the author.

of 5 input advanced variables over 15000 instances, where each instance represents the behavioral profile of a specific user on a given day.

The complete dataset was then split into training and test sets. Model training was performed on the training subset, while evaluation was conducted using the test set. Following best practices, the training data was composed primarily of normal users, whereas the test set included a higher proportion of users exhibiting anomalous or risky behavior. Specifically, the split was performed at the user level: 19 users were assigned to the test set, 68% of whom were identified as insiders, while 31 users were included in the training set, of which only 22% were labeled as insiders. The distribution of data instances was similarly stratified, with approximately 65% of records allocated to the training set.

The next logical step in the study involved selecting appropriate evaluation metrics to assess the performance of the developed models. Three commonly used classification metrics from machine learning practice were chosen: Detection Rate (DR), Accuracy, and F1 Score.

The primary and most critical metric is the Detection Rate, which reflects the proportion of true insider threats correctly identified by the model out of the total number of actual insiders. This metric is particularly important in the context of personnel security, where the main objective is to minimize the number of undetected threats. The formula for calculating the Detection Rate is as follows:

$$DR = \frac{\text{Number of detected insiders}}{\text{Total number of actual insiders}} \tag{7}$$

The second evaluation metric is Balanced Accuracy, which reflects the overall correctness of the model's predictions. It measures the proportion of all correctly classified instances relative to the total number of observations. The formula for calculating Balanced Accuracy is:

$$BalancedAccuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

where: *TP* – correctly identified insiders, *TN* – correctly identified non-insiders, *FP* – normal employees incorrectly labeled as insiders, *FN* – actual insiders missed by the model.

The third key evaluation metric is the F1 Score, which represents the harmonic mean of Precision and Recall. This metric provides a more balanced assessment of model performance, particularly in scenarios where the dataset exhibits significant class imbalance. The F1 Score is especially useful when both false positives and false negatives carry important consequences. The formula for calculating the F1 Score is:

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{9}$$

where P denotes precision and calculated as follow:

$$P = \frac{TP}{TP + FP} \tag{10}$$

Table 2
Performance of anomaly detection models on daily-level activity data

Metric	Isolation Forest	Local Outlier Factor (LOF)
Balanced Accuracy	65.74%	63.24%
F1 Score	42.25%	33.55%

Table 3
Performance of anomaly detection models on user-level aggregated data

Metric	Isolation Forest	Local Outlier Factor (LOF)
Detection Rate	100%	100%
Balanced Accuracy	100%	58%
F1 Score	100%	84%

and R means recall, i.e. the proportion of actual insiders correctly identified by the model:

$$R = \frac{TP}{TP + FN} \tag{11}$$

Accordingly, the evaluation of these metrics provides a multifaceted understanding of the model's performance, capturing both its ability to promptly detect insider threats and its overall classification accuracy with respect to employee activity.

4. Experimental results

Based on the engineered features, anomaly detection models were trained to identify potential insider threats. The evaluation was conducted at two levels: individual daily activity records and aggregated user-level behavior. This section presents the results of model performance at the daily level, where each data point represents a single day of activity for a specific user. The evaluation metrics are summarized in Table 2.

As shown in Table 2, both models demonstrate relatively low F1 Scores, suggesting that accurately identifying anomalous behavior based solely on daily patterns remains challenging. This limitation is likely due to the subtle and context-dependent nature of insider activity, which may not manifest clearly within a single day's behavioral footprint.

Nevertheless, the Balanced Accuracy values are moderately higher, particularly for the Isolation Forest model, which outperforms LOF by more than two percentage points. Balanced Accuracy, which accounts for both sensitivity and specificity in imbalanced datasets, indicates that Isolation Forest has a slightly better ability to distinguish between normal and anomalous classes. This result highlights its relative robustness in handling behavioral data where insider activity represents a minority class.

A substantially different outcome was observed when the analysis shifted from daily-level activity to user-level behavior. In this approach, all behavioral data were aggregated per user, allowing for a more holistic representation of activity patterns. The results of this evaluation are presented in Table 3.

As shown in Table 3, the Isolation Forest model achieved outstanding performance across all metrics, including a 100% Detection Rate, meaning that it correctly identified all insider threats in the test set. Moreover, it achieved perfect Balanced Accuracy and F1 Score, indicating a high degree of precision and recall with no false positives or negatives. This level of performance suggests that, when user activity is considered in aggregate rather than on a daily basis, distinct behavioral patterns become more detectable and easier to classify. Although the LOF model also achieved a 100% Detection Rate, its Balanced Accuracy (58%) and F1 Score (84%) were substantially lower than those of Isolation Forest. This indicates that while LOF was able to detect all insiders, it misclassified a larger portion of non-insiders, resulting in reduced precision and overall classification stability.

Figure 4: Dynamics of key Isolation Forest features for user 'KRL0501'. Source: compiled by the author.

It is important to emphasize that in insider threat detection tasks, timeliness of detection is as critical as the detection itself. The earlier a threat is identified, the greater the likelihood of implementing preventive measures before confidential information is leaked or organizational harm occurs. To illustrate this aspect, the modeling results for an individual user are presented in Figure 4, which shows the temporal dynamics of several key features used by the Isolation Forest model for user "KRL0501" over the analyzed period.

The Figure 4 reveals several noteworthy behavioral signals. The purple line shows intermittent spikes, suggesting periods where the user connected additional devices shortly after job search activity which is a potentially suspicious pattern. Red vertical dotted lines indicate the actual periods of insider activity, as defined by the original dataset authors. Overall, the user's activity pattern appears irregular, characterized by alternating periods of high engagement and inactivity. Especially concerning are the timeframes where elevated job search behavior coincides with increased access to suspicious files or tools. This convergence of risk indicators may signal a heightened probability of insider threat – for instance, a user actively seeking new employment while accessing or exfiltrating sensitive resources. Visualizations of this kind are critical tools for insider threat detection systems, as they capture the interplay between behavioral indicators and provide actionable insights into user intent and risk escalation over time. Such patterns can help security teams prioritize monitoring efforts and trigger early interventions.

To further assess the effectiveness of the proposed insider threat detection model, an additional analysis was conducted to examine the timing of the model's first alert relative to the actual onset of insider activity. Given the critical importance of early detection in mitigating risks such as data leaks or internal sabotage, a key objective of this analysis was to evaluate how promptly the model responds to anomalous behavioral patterns. A comparative study was carried out, measuring the time between the model's first detection of suspicious activity and the verified beginning of insider behavior. The findings are summarized below:

- 6 insiders were flagged either on the exact day of their first suspicious activity or several days prior, indicating the model's high sensitivity and its capacity to detect emerging threats at a very early stage.
- 4 users were identified within the first 31 days after the onset of insider activity, confirming the model's effectiveness under medium- and short-term monitoring scenarios.
- All insiders were detected before the end of their malicious activity period, demonstrating the

model's reliability in completing the detection-response cycle before significant harm was done.

These results provide strong evidence that the model not only performs well in detecting anomalies but also offers timely intervention, which is essential for proactive threat mitigation. Therefore, the developed approach can be considered a robust and practical solution for real-time monitoring of employee activity in enterprise environments. It is well-suited for supporting timely, evidence-based decision-making in the context of personnel security.

5. Conclusions

This study introduced a data-driven approach to insider threat detection by combining unsupervised machine learning algorithms such as Isolation Forest and Local Outlier Factor with behavioral data and novel features generated through LLMs. Using the CERT R4.2 dataset, the models were evaluated at both daily and user-aggregated levels. While detection performance on daily records was limited, the user-level analysis yielded highly promising results, with the Isolation Forest model achieving 100% detection rate, perfect balanced accuracy, and no false positives.

A key innovation was the integration of LLM-based content classification to dynamically assess web activity associated with job search and data leak risks. These enriched features improved the model's ability to capture complex behavioral patterns indicative of insider threats. Time-to-detection analysis further demonstrated that the model could identify risks before or shortly after the start of malicious activity, supporting its use for early intervention. Overall, the results confirm the high effectiveness, robustness, and practical utility of the proposed approach in real-world personnel security contexts.

Based on the outcomes of this study, the following directions are recommended for practical implementation and broader adaptation of the proposed model:

- Integration into real-world monitoring systems. The results support the effective integration of the Isolation Forest model into corporate employee monitoring platforms. Its ability to detect suspicious behavior in a timely manner makes it a valuable tool for mitigating the risk of internal data leaks.
- Strengthening internal control and risk management. Real-time detection of insider threats enables organizations not only to respond to current risks but also to proactively prevent future incidents. The model can be used as a key part of internal control systems, helping quickly detect unusual behavior that may point to data misuse, access abuse, or attempts to leak confidential information.
- Adaptation across sectors and organizations. Due to its flexibility, the Isolation Forest model can be adapted for use in various industries, including finance, government and technology. Its deployment can be tailored to the specific operational needs and risk profiles of different organizations, making it a scalable solution for enhancing personnel security in high-stakes environments.

In conclusion, the proposed approach demonstrates high potential as a practical tool for real-time monitoring and early detection of insider threats. It offers a balanced combination of detection accuracy, timeliness, and operational applicability, making it well-suited for deployment in enterprise security systems operating in dynamic, data-rich environments.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4o in order to: Grammar and spelling check. After using these service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Deloitte, Managing Risk in Digital Transformation, 2018. URL: https://www2.deloitte.com/content/dam/Deloitte/za/Documents/risk/za_managing_risk_in_digital_transformation_112018.pdf.

[2] IBM Security, Cost of a Data Breach: A Million-Dollar Race to Detect and Respond, 2024. URL: https://www.ibm.com/reports/data-breach.

[3] Thales Group, Thales Data Threat Report: Global Edition, 2025. URL: https://cpl.thalesgroup.com/data-threat-report.

[4] V. G. Goulart, L. B. Liboni, L. O. Cezarino, Balancing skills in the digital transformation era: The future of jobs and the role of higher education, Industry and Higher Education 36 (2022) 118–127. doi:10.1177/09504222211029796.

[5] M. Raissi-Dehkordi, D. Carr, A multi-perspective approach to insider threat detection, in: 2011 - MILCOM 2011 Military Communications Conference, 2011, pp. 1164–1169. doi:10.1109/MILCOM.2011.6127457.

[6] T. Rashid, I. Agrafiotis, J. R. Nurse, A new take on detecting insider threats: Exploring the use of hidden markov models, in: Proceedings of the 8th ACM CCS International Workshop on Managing Insider Security Threats, MIST '16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 47–56. doi:10.1145/2995959.2995964.

[7] Y. Song, M. B. Salem, S. Hershkop, S. J. Stolfo, System level user behavior biometrics using fisher features and gaussian mixture models, in: 2013 IEEE Security and Privacy Workshops, 2013, pp. 52–59. doi:10.1109/SPW.2013.33.

[8] G. Gavai, K. Sricharan, D. Gunning, R. Rolleston, J. Hanley, M. Singhal, Detecting insider threat from enterprise social and online activity data, in: Proceedings of the 7th ACM CCS International Workshop on Managing Insider Security Threats, MIST '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 13–20. doi:10.1145/2808783.2808784.

[9] D. C. Le, N. Zincir-Heywood, Anomaly detection for insider threats using unsupervised ensembles, IEEE Transactions on Network and Service Management 18 (2021) 1152–1164. doi:10.1109/TNSM.2021.3071928.

[10] T. A. Al-Shehari, D. Rosaci, M. Al-Razgan, T. Alfakih, M. Kadrie, H. Afzal, R. Nawaz, Enhancing insider threat detection in imbalanced cybersecurity settings using the density-based local outlier factor algorithm, IEEE Access 12 (2024) 34820–34834. doi:10.1109/ACCESS.2024.3373694.

[11] C. Song, L. Ma, J. Zheng, J. Liao, H. Kuang, L. Yang, Audit-llm: Multi-agent collaboration for log-based insider threat detection, 2024. URL: https://arxiv.org/abs/2408.08902, arXiv preprint.

[12] V. Chouksey, Automatic Selection of Outlier Detection Techniques, Master's thesis, Eindhoven University of Technology, Netherlands, 2018. URL: https://pure.tue.nl/ws/portalfiles/portal/109406381/CSE663_Vishal_Chouksey_31_aug.pdf.

[13] B. Lindauer, Insider threat test dataset, 2020. URL: https://kilthub.cmu.edu/articles/dataset/Insider_Threat_Test_Dataset/12841247.

[14] M. Abdin, X. Zhou, Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL: https://arxiv.org/abs/2404.14219, arXiv preprint.

[15] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, I. Stoica, Efficient memory management for large language model serving with pagedattention, 2023. URL: https://arxiv.org/abs/2309.06180, arXiv preprint.

[16] Google, Google Colaboratory, 2025. URL: https://colab.google/.