Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Deep Vision Guard: Insider Threat Detection Via Deep Learning-Enhanced Image Forensics

Sourav Kumar Upadhyay[1*], Dr. Subhash Chandra Dutta[2]

[1*]Assistant Professor, ICFAI University, Jharkhand; CSE; Cyber Security; Email ID: sourav.cse.rnc@gmail.com; Orcid ID: 0009-0006-2885-1394
[2]Associate Professor (HOD), BIT Sindri, Dhanbad; CSE; Cyber Security and Machine Learning ; Email ID: dutta_subhash@yahoo.com

Abstract
Insider threats remain difficult to detect because malicious activities are often embedded within legitimate organizational behavior and conventional monitoring may overlook subtle temporal deviations. This research introduces DeepVisionGuard, a deep learning-supported behavioral image-forensics framework to discriminate an anomalous insider from a normal cybersecurity event. A total of 71,965 events were cleaned from the SPEDIA dataset and converted into 5,212 chunks of behavioral-image data, each containing a fixed sequence of 32 events, with the help of 84 engineered features. The main architecture was based on CNN for spatial feature extraction, BiLSTM for temporal modelling, and attention, and Grad-CAM and temporal-attention analysis were used for interpretability. In 733 independent test chunks, the directly supervised model had 14 false positives and 14 false negatives and achieved an accuracy of 96.18%, an F1-score of 0.9451, an MCC of 0.9158, an ROC-AUC of 0.9912 and a PR-AUC of 0.9872. The ablation results indicated that by removing the CNN-only part, the performance was significantly lower, while the addition of attention was not statistically significantly better than CNN–BiLSTM, confirming that the CNN-only part was essential and critical to the model's performance. Gains in description were not significant and were small in a secondary experiment that was unsupervised-pretrained. Overall, DeepVisionGuard demonstrates that behavioral-image representation combined with temporal modelling can support accurate, leakage-aware, and interpretable insider-threat detection, while future work should examine cross-organizational validation, sparse anomaly scenarios, richer threat labels, and real-time analyst-assisted deployment.

Keywords: Insider Threat Detection; Deep Learning; Behavioral Image Forensics; CNN–BiLSTM–Attention; Explainable Artificial Intelligence

1. Introduction

Insider threats are a constant cyber security threat as bad actors can have access to organisational systems, data and services by legitimate means. Insider incidents can sometimes be hard to tell apart from ordinary work since the same credentials, devices, applications, and network resources can be used for both good and bad reasons. This means that ebbing from user behavior is one of the major factors that make insider-threat detection different from traditional perimeter defenses. In practice, examples from multinationals have demonstrated that insider threat analytics should be constrained to practical conditions like the heterogeneous data sources, operational noise, high analyst workload, privacy concerns, and the lack of ability to validate alerts in real-time environments (Erola et al., 2022). Intelligent detection systems, therefore, have to be very strong in their prediction capability and very good in their representation of the meaningful behavioral context.

To detect unusual user behavior in audit trails and other cybersecurity logs, an increasing number of machine-learning techniques are used. Such methods can learn complex mappings between behavioral properties that are hard to translate into rules to be manually written, thereby being more adaptable in suspicious activity detection (Bin Sarhan & Altwaijry, 2023). But, insider threat data tends to be unbalanced and diverse, and malicious behavior may represent a small or inconsistent fraction of the data. This can distort classifier learning and create apparently strong overall performance while minority threat behaviors remain poorly detected. CNN-based investigations have therefore emphasized explicitly addressing data imbalance during insider-threat classification (Al-Shehari et al., 2024).

A further challenge concerns how complex behavioral logs should be represented before being supplied to a learning algorithm. Conventional tabular representations may underuse contextual relationships among related behavioral attributes. Image-based insider-threat research has demonstrated that user

activities can instead be organized into visual representations and analyzed with deep neural architectures, allowing contextual patterns to be learned across behavioral dimensions (Duan et al., 2024). This direction is consistent with the wider evolution of deep-learning-based image forensics, where neural networks learn discriminative visual representations that are difficult to specify through handcrafted features alone (Castillo Camacho & Wang, 2021). In DeepVisionGuard, image forensics therefore refers specifically to forensic analysis of behavior-derived images constructed from cybersecurity events rather than classical manipulation detection in photographic evidence.

Several methodological problems nevertheless remain. This balance between false positives and false negatives can be a challenge for analysts, leading to a decline in the effectiveness of automated alerts, while false negatives can permit detrimental insider activity to go undetected, making it essential to control this balance in practical detection (AlSlaiman et al., 2023). Isolated activities can also happen in a temporal sequence – that is, activities themselves may not seem suspicious, but when put together and in context, they may be. Fine-grained methods modeling the temporal dependency among the activities indicate the necessity of modeling the relationship among sequences of activities instead of using coarse summaries at user- or day-level only (Cai et al., 2024). The study of deep-learning based authentication in the field of critical infrastructure further demonstrates the relevance of learned behavioral representations in security sensitive insider-threat detection (Budžys et al., 2024).

Considering this context, this paper suggests the DeepVisionGuard framework for behavioral image forensics, which is enhanced by deep learning and transforms the temporal series of cybersecurity events into structured behavioral images and processes them using a CNN–BiLSTM–Attention structure. CNN captures the local behavioral-image patterns, the bidirectional LSTM models temporal dependencies between the positions of events, and attention offers a transparent way to inspect importance distribution over a sequence. The framework employs leakage-controlled, disjoint training, validation and testing to ensure that a similar behavioral scenario does not span evaluation boundaries. Grad-CAM localization and temporal-attention analysis are used to integrate explainability, so that model decisions can be examined beyond just the uninformative anomaly score.

The study is limited to a binary classification of normal and abnormal insider-related activities, based on a set of structured events, converted into behavioral images. It is not claiming to be able to infer psychological motive, classify attack stages that are not found among the labels provided or to conventional manipulation detection on native screenshots or photographs. The experiments also are not real-time deployments at scale, but rather are a controlled evaluation offline. These constraint are part of the interpretation of the results and ensure the usefulness of the proposed framework in practice.

DeepVisionGuard is important in that it combines the four major components of visual representation learning, temporal modeling, attention and explainable analysis into a single insider-threat framework. Besides accuracy, the study assesses precision, recall, specificity, F1-score, MCC, ROC-AUC, PR-AUC, group-level behavior, architectural ablations, and paired statistical comparison. This gives a more stringent criterion to assess if architectural elements are meaningful for detection, and if unsupervised pretraining actually provides measurable value.

Accordingly, the study pursues three research objectives:

- To develop a leakage-aware behavioral image representation and CNN–BiLSTM–Attention framework for distinguishing anomalous insider activity from normal organizational behavior.
- To evaluate DeepVisionGuard using classification, discrimination, group-level, ablation, and statistical measures, including a secondary unsupervised-pretraining and transfer-learning experiment.
- To investigate model interpretability through Grad-CAM and temporal-attention analysis so that influential behavioral-image regions and temporal positions can be examined alongside predictive performance.

2. Literature Review

The evolution of insider-threat detection has shifted from traditional rule-based monitoring to machine-learning and deep-learning techniques that can detect slight changes in behavior across heterogeneous activities within an organization. An early focus was on choosing a suitable granularity of time and behaviour for modeling user actions. To investigate the varying levels of data granularity for insider-threat detection, Le et al. (2020) explored several levels of granularity for representing user activity and showed that the granularity needs to be fine-grained enough to capture meaningful behavior. Fine-grained modelling has also been used to model the activity of the file system, a model which can be used to enhance temporal user profiles to support the detection of anomalies; instead of considering individual accesses (Mehnaz & Bertino, 2019). These studies laid the foundations for the current insider threat

analytics, through the creation of temporal behavioral modelling.

The challenge of getting a complete set of labels for malicious users has encouraged unsupervised approaches. Le and Zincir-Heywood (2021) explored the unsupervised ensemble-based anomaly detection for insider threats, highlighting the importance of learning the structure of normal behaviour and detecting deviations from the normal without using only labeled attacks. This is especially important as real world organizational settings may have unknown ways of mis-use. At a higher level, Yuan and Wu (2021) surveyed the deep-learning methods for insider-threat detection and identified some common problems associated with insider-threat detection, such as the scarcity of malicious samples, the complexity of insider behavior, the complexity of insider behavior representation, the lack of generalization and interpretability. The difficulties that arise show that it is important to improve the representation of user activity as much as it is important to increase the complexity of the model.

The representation of data by images has therefore become an alternative to the tabular representation. Gayathri et al. (2020) suggested the use of images to represent insider-related behavioral features for convolutional architectures to take advantage of the spatially organized patterns among activity variables. The feasibility of mapping a structured cybersecurity activity into a structure that can be used for computer-vision-inspired learning was also demonstrated by Li et al. (2021), who explored insider threat detection using geometric transformation of images. Randive et al. (2023) continued this line of research by introducing a classification approach that is based on patterns and feature representations of images. These studies present a comprehensive conceptual support for DeepVisionGuard since they are seen from the perspective of transforming the temporally structured activity logs into behavioral images as opposed to simple individual events.

Even if using image-based representations temporal relationships are still important. Li et al. (2022) proposed a memory-augmented framework that also fuses the temporal and spatial information, highlighting the significance of modelling temporal-spatial information jointly. Pal et al. (2023) also explored the temporal feature aggregation for insider-threat detection in activity logs, reinforcing the role of attention mechanisms for assigning different weights to temporal features of the activities spread out in time. Instead, Song et al. (2024) studied the problem using the behavioral rhythm modelling with time awareness and user adaptation approach, which is a trend in developing systems that take into account differences in users' temporal activity profiles. Xiao et al. (2024) also performed a statistical and sequential analysis, but in the context of a wider insider-threat framework, reinforcing the perspective that the detection of insider threats should also take into account the temporal aspects as well as how the information is organized.

The personalization and contextual modelling have been investigated in recent studies too. Racherache et al. (2023) suggested insider-threat detection by cyber-persona identification and profiling, showing the potential of user-centred behaviour characterization in identifying abnormal behaviours relative to expected behaviours. Wang et al. (2023) proposed a digital-twin based framework in combination with self-attention deep-learning models, further highlighting the use of attention-oriented architectures for modelling complex user behaviors. Tian et al. (2025) considered detection in limited insider-threat scenarios, and noted that insider threats are manifested in a variety of different ways; thus, the systems cannot be evaluated based on homogeneous insider anomaly characteristics.

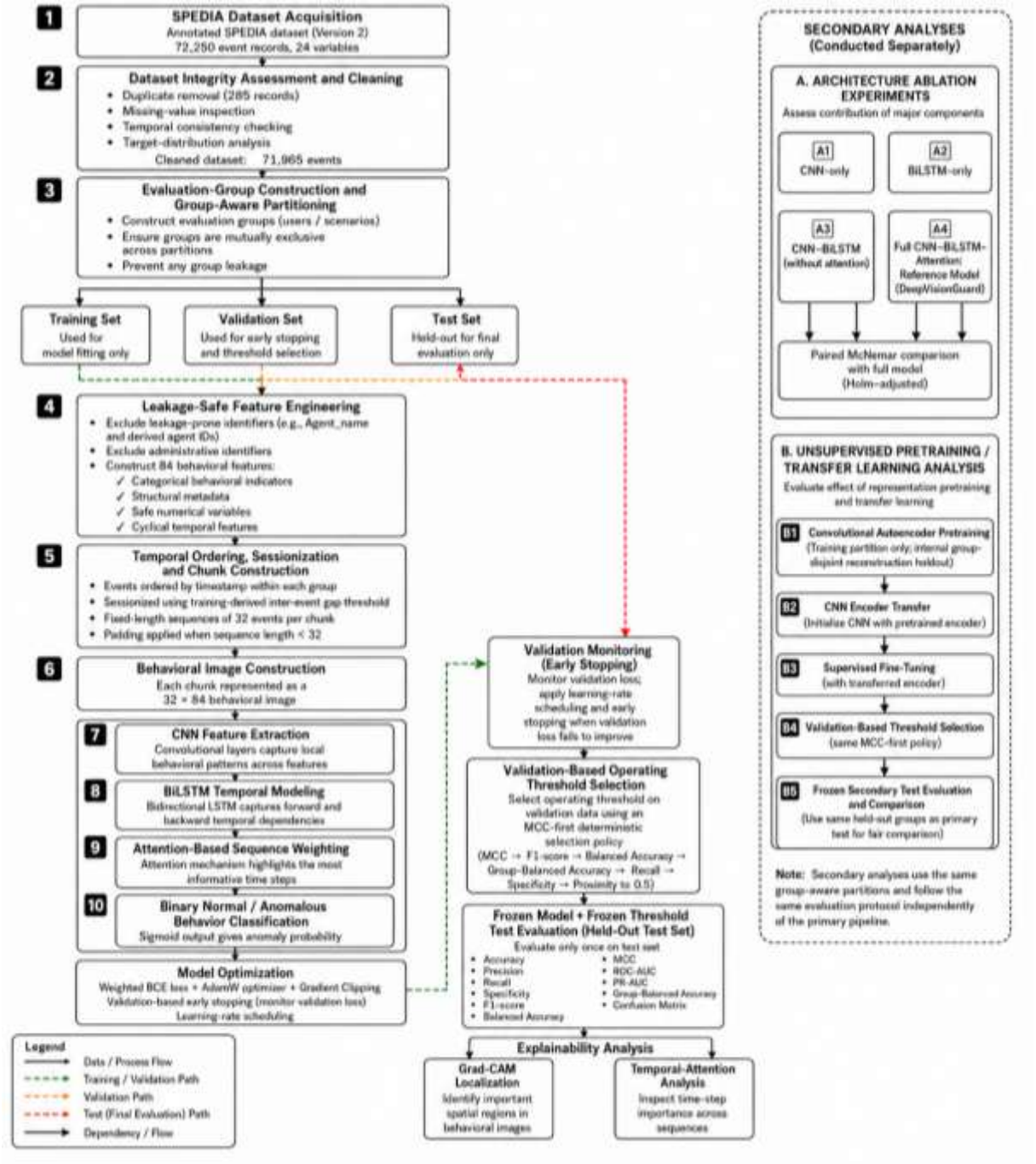
One of the other big issues is the lack of matching malicious behavior, and class imbalance. Gayathri et al. (2024) solved this by proposing a hybrid deep learning system using augmentation techniques based on SPCAGAN, which is gaining popularity in the application of synthetic data for boosting insider-threat analysis in scenarios where there are few observations of the threats. While augmentation can also enhance the availability of data, there are also methodological issues such as the representativeness of the augmented data to real malicious traffic. For this reason, anomaly-oriented and transfer-oriented approaches are desirable to consider as supplementary approaches in the absence of labelled examples.

Literature shows that there has been significant advancement in machine learning, unsupervised anomaly detection, image-based representation, temporal modelling, attention, profiling and data augmentation.In general, the literature shows a great deal of progress in machine learning, unsupervised anomaly detection, image-based representation, temporal modelling, attention, profiling and data augmentation. However, each strand is often analysed individually. Finally, there is a research opportunity for a single framework that integrates the leakage-aware construction of behavioral images, convolutional feature extraction, bidirectional temporal modelling, attention-based sequence weighting, representation pre-training unsupervised by data labels, and explainability of the resulting predictors. DeepVisionGuard bridges this gap by combining these components in just one experimentally controlled framework, as well as by separately evaluating predictive performance and interpretability and by statistically validating the contributions of the architectures using ablation and paired analysis.

3. Methodology

This study developed DeepVisionGuard, a deep-learning framework for insider-threat detection using behavioral image forensics derived from cybersecurity activity logs. Instead of treating raw event logs only as tabular observations, the methodology transformed them into structured two-dimensional behavioral images that could be processed by convolutional and sequential neural-network components originally inspired by image-recognition tasks. The proposed methodology was designed to limit information leakage, preserve temporal behavioral patterns, enable interpretable anomaly detection, and ensure evaluation reproducibility. The directly supervised CNN–BiLSTM–Attention (CNN–BL–Attention) architecture was selected as the primary model, while an unsupervised convolutional autoencoder (CAE) pretraining stage followed by supervised transfer learning was conducted as a secondary methodological experiment. The overall workflow of the proposed DeepVisionGuard methodology is illustrated in Figure 1.

Figure 1. Overall workflow of the proposed DeepVisionGuard methodology.



3.1 Dataset Acquisition and Integrity Assessment

The experiments used the annotated SPEDIA cybersecurity dataset (Version 2), a publicly available dataset developed for the prediction and early detection of insider attacks (Álvarez et al., 2025). The dataset consisted of 72,250 event records and 24 variables that represent behavioural data related to commands, files, emails, HTTP activity, devices, sessions, timestamps, Anomaly label, etc. Initial integrity analysis included observations of the size of the data sets, missing data, duplicated data, temporal consistency, categorical structure, numerical range, target distribution, and relationships between potentially identifying attributes.

There were 285 behavioral records that were duplicated and removed, leaving a cleaned sample of 71,965 events. The target distribution had 30,478 normal events and 41,487 anomalous events. The fields found to be very deterministic of the target in leakage analysis were Agent_name and derived agent identifiers. The above variables were not included in predictive modelling. Administrative identifiers were also not included and user identifiers were only included for grouping and audit purposes, but not for use as predictive features.

3.2 Leakage-Safe Data Partitioning

A deterministic group-aware splitting strategy was used to avoid repeating scenarios for the same behavioral sequences or its relation across various data partitions. Evaluation groups were built based on baseline days, anomalies and stand-alone behavioral sources. No data was overlapping between the training, validation and test partitions.

The total number of split events were divided into 50,350 events for training, 10,797 events for validation, and 10,818 events for testing. Following temporal sessionization and chunk construction these were 3629 training chunks, 850 chunks for validation and 733 chunks for test. Examples were found to be both normal and abnormal in all the partitions. The test partition was not used to develop the model, to select hyperparameters, to use early stopping, nor to determine the operating-threshold.

3.3 Behavioral Feature Engineering and Image Construction

The raw activity data were transformed into 84 leakage-safe features. They consisted of 60 categorical behavioral indicators, 16 structural metadata features, four safe numerical variables and four cyclical temporal variables. Structural features were defined by the path presence, path depth and log length; while cyclical temporal features were computed as sine and cosine transformations of time of day and day of week.

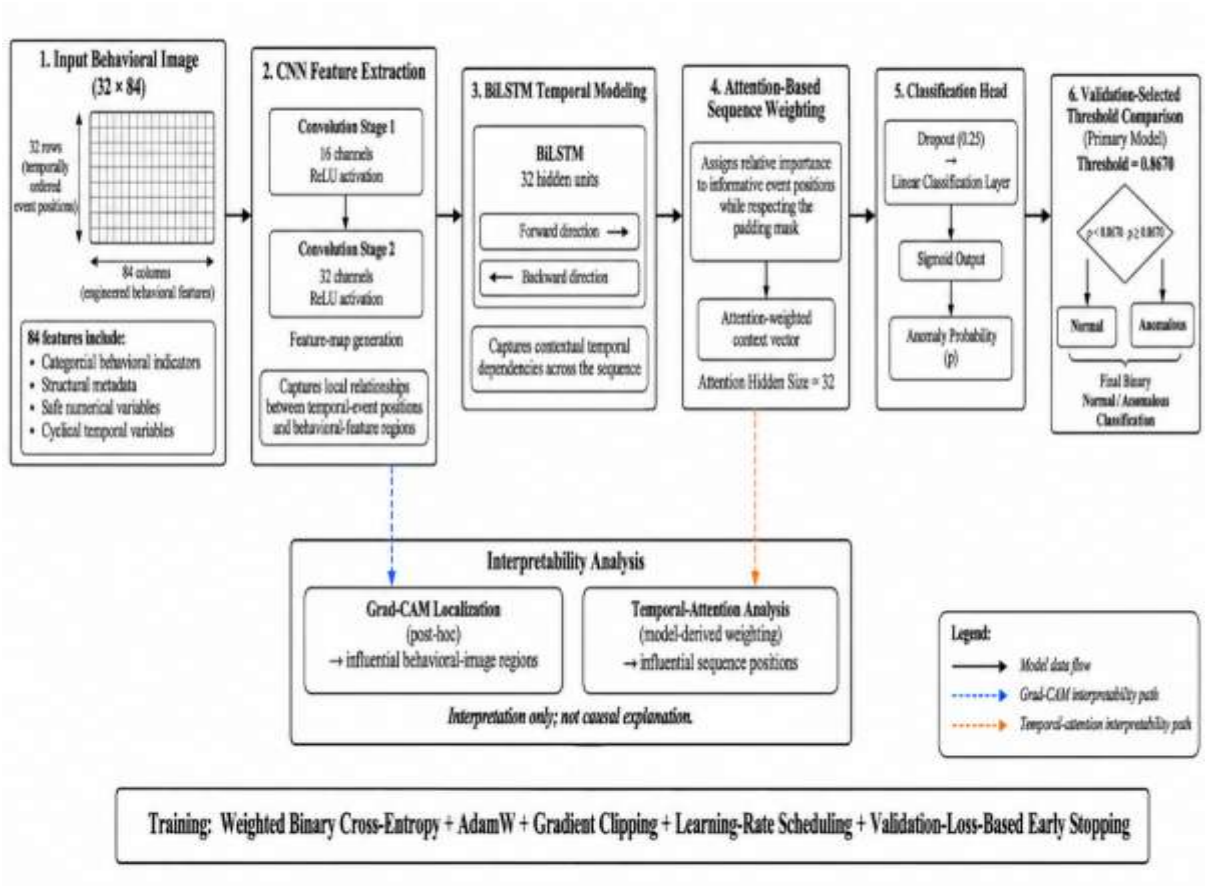
All the continuous variables were well normalized and bounded transformations were performed prior to building up images. Events were ordered temporally within user and evaluation groups. A training-derived inter-event gap threshold of 474 seconds was used for sessionization. Sessions were subsequently divided into fixed-length sequences of 32 events, with zero-padding applied where necessary. Each resulting sample was represented as a 32 × 84 behavioral image, where rows represented temporal event positions and columns represented engineered behavioral features. Padding masks ensured that artificial padded positions did not contribute to sequential processing or explainability calculations.

3.4 DeepVisionGuard Architecture

The primary DeepVisionGuard model combined a convolutional neural network, bidirectional long short-term memory network, and attention mechanism. The CNN operated on the structured behavioral images to learn local relationships among temporal events and feature regions. Two convolutional stages with 16 and 32 channels were used to generate compact behavioral representations. The overall working architecture of the proposed CNN–BiLSTM–Attention model is illustrated in Figure 2, showing the progression from behavioral-image input through convolutional feature extraction, bidirectional temporal modelling, attention-based sequence weighting, and final anomaly classification.

The resulting temporal feature sequence was passed to a BiLSTM with 32 hidden units, allowing the model to learn contextual dependencies in both temporal directions. A trainable attention mechanism then assigned relative importance to individual event positions before classification. Dropout was applied to reduce overfitting, and the final classifier produced a binary anomaly logit. The complete architecture contained 24,049 trainable parameters.

Figure 2. Working architecture of the proposed DeepVisionGuard CNN–BiLSTM–Attention model for behavioral-image-based insider-threat detection.



3.5 Model Training and Threshold Selection

The primary model was trained using weighted binary cross-entropy to account for class imbalance. AdamW optimization was used with an initial learning rate of 0.001, weight decay of 0.0001, gradient clipping, learning-rate reduction on validation plateaus, and early stopping based on validation loss. The best primary checkpoint occurred at epoch 20.

The operating threshold was not optimized on the test set. Instead, validation probabilities were systematically evaluated across data-driven threshold candidates. Matthews correlation coefficient (MCC) was the primary threshold-selection criterion, followed by F1-score, balanced accuracy, group-balanced accuracy, recall, specificity, and proximity to 0.5 as deterministic tie-breakers. The frozen primary threshold was 0.8670.

3.6 Ablation and Transfer-Learning Experiments

Three architecture ablations were evaluated: CNN-only, BiLSTM-only, and CNN–BiLSTM without attention. Each model was independently trained and assigned its own validation-selected threshold before one-time test evaluation. Exact McNemar tests with Holm adjustment were subsequently used to compare paired test errors against the complete model.

A secondary transfer-learning experiment was also conducted. A convolutional autoencoder was first trained without class labels using only the training partition. Its pretrained CNN encoder was transferred into a newly initialized CNN–BiLSTM–Attention model and subsequently fine-tuned using supervised training and validation data. This experiment was retained as a secondary analysis and was not allowed to replace or redefine the pre-specified primary model.

3.7 Evaluation and Explainability

Performance was assessed using accuracy, precision, recall, specificity, F1-score, balanced accuracy, MCC, ROC-AUC, PR-AUC, and confusion-matrix statistics. Group-balanced accuracy was additionally reported to evaluate performance across heterogeneous behavioral scenarios.

Model interpretability was examined using Grad-CAM and temporal attention analysis. Grad-CAM localized influential regions within the 32 × 84 behavioral images, while attention weights characterized how the model distributed importance across sequence positions. These explanations were interpreted as model-localization evidence rather than causal attribution.

4. Results

The finalized DeepVisionGuard experiments evaluated the proposed CNN–BiLSTM–Attention architecture under a leakage-controlled, group-disjoint experimental design. Results are reported for the frozen primary model, architecture ablations, the secondary unsupervised-pretraining/transfer-learning experiment, and explainability analyses. All operating thresholds were selected using validation data before test evaluation, and no model or threshold was modified after the test results were observed.

4.1 Dataset Partition and Experimental Sample Distribution

Following data cleaning, sessionization, and fixed-length sequence construction, 71,965 behavioral events were transformed into 5,212 model-ready chunks. The partitions remained group-disjoint, with 43 evaluation groups used for training, 10 for validation, and 9 for final testing. Table 1 summarizes the resulting experimental partitions and demonstrates that both normal and anomalous behavioral chunks were represented in every split.

Table 1. Data partition and behavioral-chunk distribution

Partition	Events	Chunks	Normal Chunks	Anomalous Chunks	Evaluation Groups
Training	50,350	3,629	2,146	1,483	43
Validation	10,797	850	509	341	10
Test	10,818	733	478	255	9

This distribution preserved an independent final test cohort of 733 behavioral images and prevented evaluation groups from crossing the training, validation, and test boundaries.

4.2 Primary DeepVisionGuard Performance

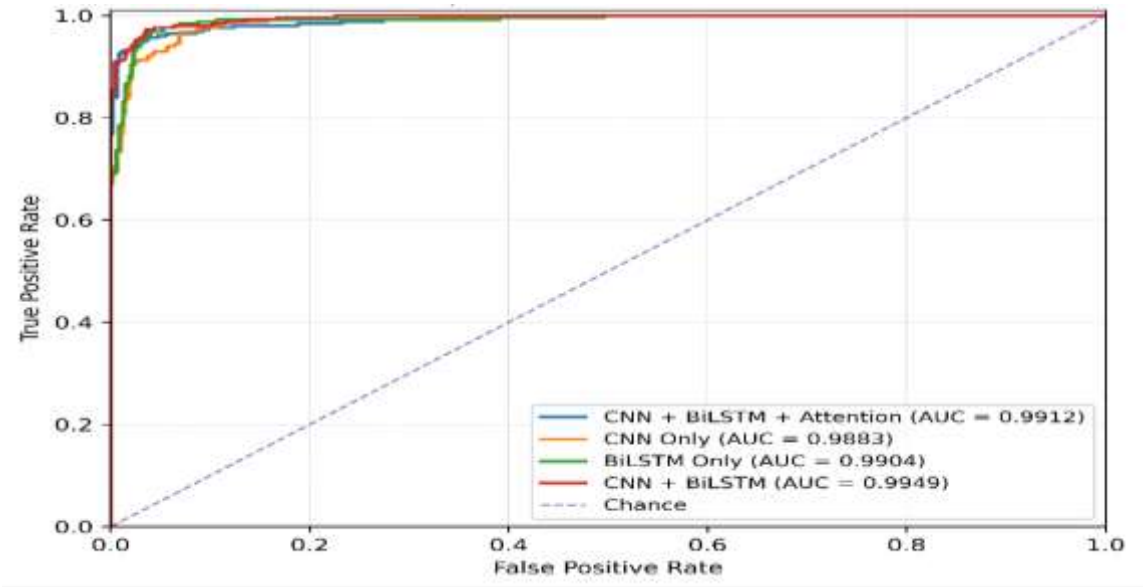
The directly supervised CNN–BiLSTM–Attention model was selected at epoch 20 and evaluated using its validation-selected threshold of 0.867042. The model correctly classified 705 of the 733 test chunks. Table 2 presents the complete primary performance profile.

Table 2. Final test performance of the primary DeepVisionGuard model

Metric	Value
Accuracy	0.961801
Precision	0.945098
Recall/Sensitivity	0.945098
Specificity	0.970711
F1-score	0.945098
Balanced Accuracy	0.957905
MCC	0.915809
ROC-AUC	0.991172
PR-AUC	0.987150
Group-Balanced Accuracy	0.755529
TP / TN / FP / FN	241 / 464 / 14 / 14

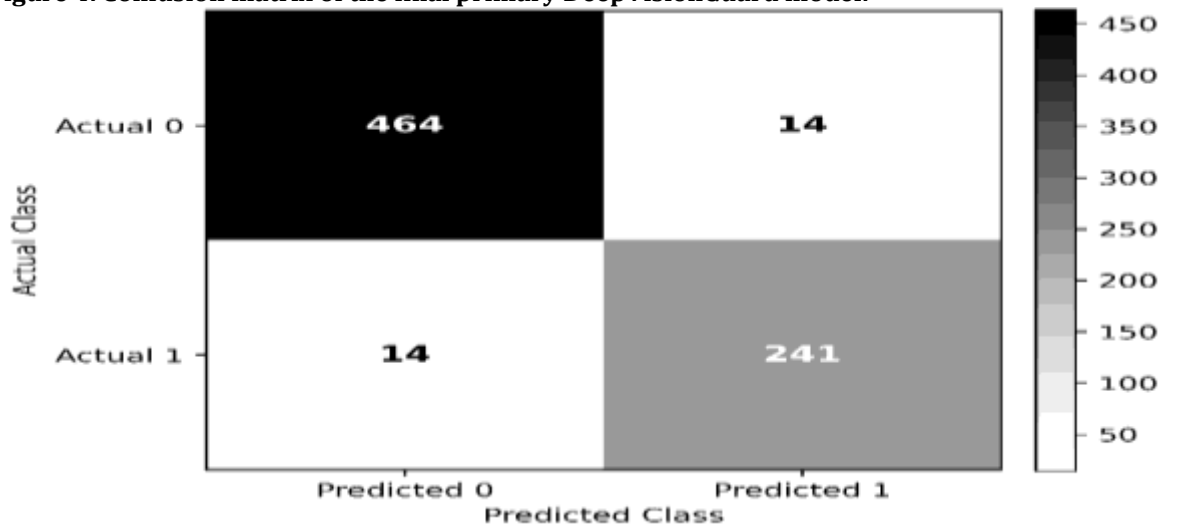
The high MCC of 0.9158 indicates strong agreement beyond chance, while ROC-AUC and PR-AUC values above 0.98 demonstrate strong class discrimination. Figure 3 shows that all evaluated architectures achieved high ROC performance, with AUC values ranging from 0.9883 to 0.9949.

Figure 3. ROC comparison of the frozen DeepVisionGuard architectures.



The error structure of the primary system is illustrated in Figure 4, showing balanced false-positive and false-negative counts.

Figure 4. Confusion matrix of the final primary DeepVisionGuard model.



4.3 Architecture Ablation Results

Ablation experiments were conducted to quantify the contribution of the major architectural components. Each ablation received an independently selected validation threshold before final test evaluation. Table 3 shows that the CNN-only model produced the weakest performance, whereas the CNN–BiLSTM model without attention achieved the numerically highest F1-score and MCC.

Table 3. Frozen-test architecture ablation comparison

Architecture	Parameters	Accuracy	F1	MCC	ROC-AUC	Holm-adjusted McNemar p
CNN-only	4,945	0.942701	0.918605	0.874544	0.988326	0.028066
BiLSTM-only	19,809	0.960437	0.943249	0.912886	0.990434	1.000000
CNN–BiLSTM, no attention	21,937	0.964529	0.949807	0.922654	0.994930	1.000000
CNN–BiLSTM–Attention	24,049	0.961801	0.945098	0.915809	0.991172	Reference

The CNN-only model was significantly worse than the full model after Holm correction. In contrast, neither BiLSTM-only nor CNN–BiLSTM without attention showed a statistically significant difference from the complete architecture. Therefore, the attention mechanism did not establish a measurable predictive advantage, although it remained valuable for temporal interpretability.

4.4 Effect of Unsupervised Pretraining and Transfer Learning

The secondary experiment transferred an autoencoder-pretrained CNN encoder into a newly initialized CNN–BiLSTM–Attention model. Its validation-selected operating threshold was 0.652667. Table 4 compares this secondary model with the pre-specified directly supervised primary model.

Table 4. Direct supervised versus pretrained transfer-learning model

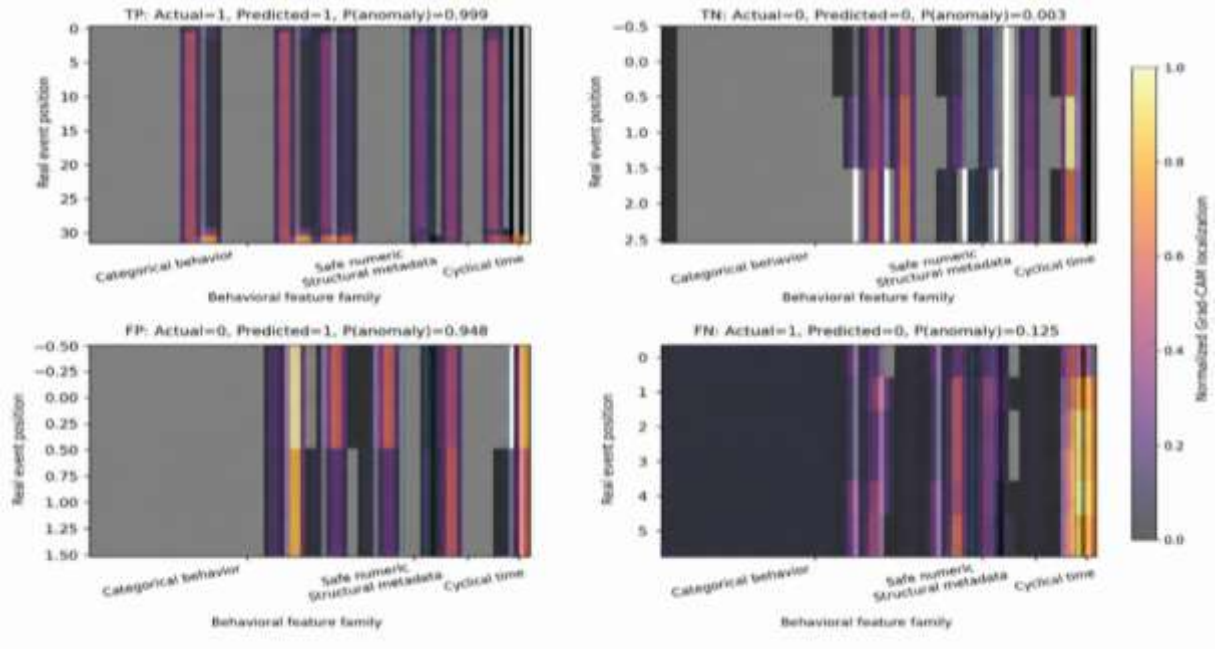
Model	Accuracy	Precision	Recall	Specificity	F1	MCC	ROC-AUC	PR-AUC
Direct supervised primary	0.961801	0.945098	0.945098	0.970711	0.945098	0.915809	0.991172	0.987150
Pretrained secondary	0.963165	0.928571	0.968627	0.960251	0.948177	0.920118	0.993831	0.990619

Pretraining increased recall by 0.023529 and MCC by 0.004309 but reduced precision by 0.016527 and specificity by 0.010460. The two models agreed on 98.23% of test predictions, and the exact paired McNemar test was non-significant (p = 1.000), indicating no statistically established improvement from pretraining.

4.5 Explainability Results

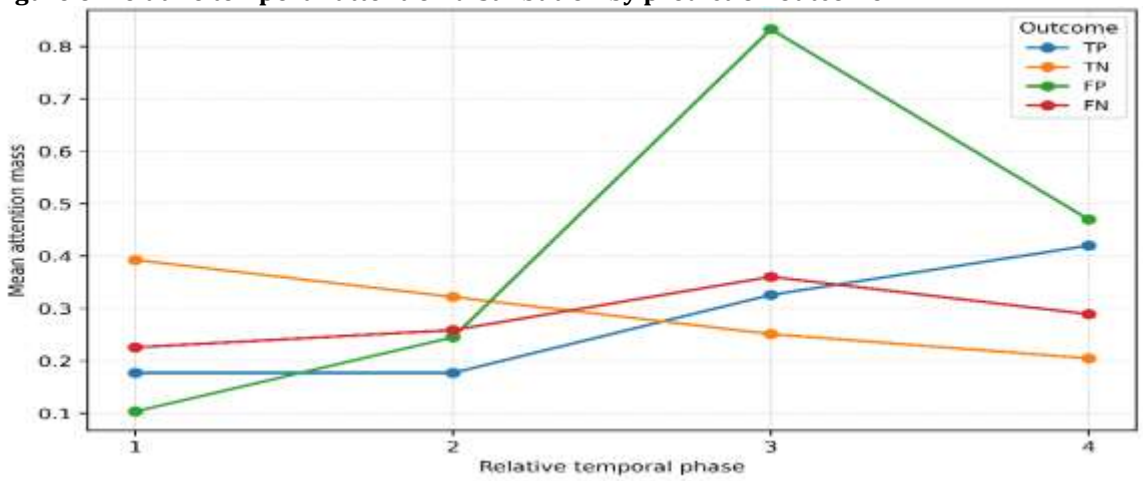
Grad-CAM analysis yielded the highest global localization score for cyclical temporal variables (0.288950), followed by structural metadata (0.129885), safe numerical variables (0.104622), and categorical behavior (0.082271). The most important individual features were weekday cosine, weekday sine, depth of path, device activity, e-mail activity, hour cosine and hour sine. Representative Grad-CAM visualization examples are shown in Figure 5, indicating the regions of the images where the model paid more attention to during classification for true-positive, true-negative, false-positive and false-negative predictions, respectively; this is providing an interpretable view of the different spatial patterns involved in the various prediction outcomes.

Figure 5. Representative Grad-CAM explanations for behavioral-image predictions.



Temporal weighting also varied according to prediction outcome. Figure 6 shows that the true-positive sequences put the highest mean attention mass on the last phase of the relative phase (0.420107), while the true negatives focused on the first phase (0.392704). The false negative and false positive patterns had a higher attention in phase 3 (0.360486), but note that error category estimation was based on relatively limited number of chunks and thus is not intended to represent population-level effects, but rather diagnostic patterns.

Figure 6. Relative temporal-attention distribution by prediction outcome.



Overall, the results demonstrate strong insider-threat discrimination while showing that the principal incremental value of attention lies in interpretable temporal weighting rather than statistically superior predictive accuracy.

5. Discussion

The findings demonstrate that DeepVisionGuard can discriminate anomalous insider-related behavior from normal organizational activity with strong predictive performance while maintaining a leakage-controlled evaluation design. The primary CNN–BiLSTM–Attention model showed an accuracy of 0.9618, an F1-score of 0.9451, a MCC of 0.9158, an ROC-AUC of 0.9912, and a PR-AUC of 0.9872 on the independent test set, with 14 false positives and 14 false negatives out of a total of 733 behavioral-image chunks. The results indicate that these temporally structured cybersecurity activities can be captured in a helpful way in 32 × 84 behavioral images, which can be used to capture complementary behavioral patterns by convolutional and sequential components. The specificity is relatively high at 0.9707, which is significant in insider-threat monitoring since too many false alarms can clutter the security analyst's mind. Meanwhile, the overall accuracy of 0.7555 suggests that the results of the group don't necessarily reflect the same level of accuracy across all behavioral contexts. It should be interpreted as a very discriminatory model overall, but with variation across groups of insider-behaviours.

The results are broadly consistent with previous research emphasizing the value of image-based behavioral representation and temporal modelling. Gayathri et al. (2020) demonstrated that cybersecurity activities can be reorganized into image-like representations suitable for convolutional classification, supporting the central design principle adopted in DeepVisionGuard. The present findings extend that idea by combining image-based representation with bidirectional temporal modelling and sequence-level attention rather than relying only on spatial feature extraction. Similarly, Pal et al. (2023) showed that temporal feature aggregation and attention can improve the modelling of activity-log sequences. In the present study, however, the ablation analysis provides a more nuanced interpretation: the CNN-only architecture was significantly worse than the complete model after Holm correction, whereas BiLSTM-only and CNN–BiLSTM without attention were not significantly different from the full architecture. Thus, temporal modelling appears important, but the attention mechanism should not be presented as providing a statistically demonstrated predictive advantage. Its stronger contribution is interpretability.

The secondary unsupervised-pretraining experiment also produced informative results. Transferring an autoencoder-pretrained CNN encoder slightly increased accuracy, recall, F1-score, MCC, ROC-AUC, and PR-AUC relative to the directly supervised primary model, but precision and specificity decreased. More importantly, the two systems agreed on 98.23% of test predictions and the paired McNemar comparison

was non-significant at p = 1.000. This finding is consistent with the motivation underlying unsupervised insider-threat modelling discussed by Le and Zincir-Heywood (2021), where learning behavioral structure without relying completely on malicious labels can be valuable when annotated threat data are limited. Nevertheless, the present results indicate that unsupervised pretraining should not automatically be assumed to improve downstream classification. In this dataset, its benefit was descriptive rather than statistically established.

The explainability analysis further strengthens the practical interpretation of the model. Grad-CAM identified the strongest global response was on cyclical temporal variables, then structural metadata, then safe numerical variables and then categorical behavioral indicators. There was also a difference in temporal attention depending on the prediction outcome: True Positive shows more of the attention mass towards later phases of the sequence, while True Negative shows more of the attention mass towards earlier phases of the sequence. The results align with the general concern Yuan and Wu (2021) raised, that interpretability in addition to predictive performance is a critical issue to tackle in insider-threat deep learning. In the case of security operations those explanations might help analysts investigate the effect of particular temporal positions or behavioral-image regions on an alert. But grad-CAM and attention values are not necessarily indicative of malicious behavior but of model-localization and weighting behavior.

Several limitations remain. The study uses a single annotated cybersecurity dataset and evaluates binary normal-versus-anomalous behavior; therefore, the results cannot establish generalization across organizations, threat taxonomies, or previously unseen operational environments. The data does not reliably label the psychological motive and does not contain detailed information on attack phases, so that the threat interpretation is not possible for the stages. The evaluation is carried out offline and the two singular anomalous test groups are not identified, which suggests that it is sensitive to very sparse scenarios. Future research should therefore examine cross-dataset and cross-organizational validation, scenario-specific calibration, richer multi-stage insider-threat labels, and prospective deployment studies. Further research may explore self-supervised sequence pretraining, adaptive baseline per user, uncertainty-aware alert prioritization, and mechanisms for incorporating the feedback of analysts, all while remaining at a high level of group leakage control and ensuring interpretable decision support.

Conclusion

This study developed and evaluated DeepVisionGuard, a leakage-aware insider-threat detection framework that transforms temporally structured cybersecurity events into behavioral images and analyzes them using a CNN–BiLSTM–Attention architecture. The primary model achieved 96.18% accuracy and an F1-score of 0.9451; 14 false positives and 14 false negatives in 733 test chunks showed excellent discrimination between normal and anomalous behavior; the MCC score was 0.9158, and the ROC-AUC was 0.9912; the PR-AUC was 0.9872. The ablation results also demonstrated that the CNN-only model achieved the worst result, while the attention mechanism did not further improve the model's predictive performance over CNN–BiLSTM, indicating that its main role is to interpret temporal aspects rather than improve the accuracy of the model. The secondary unsupervised-pretraining experiment resulted in small descriptive improvements in several metrics, but the paired comparison was not statistically significant and thus the directly supervised model is the main DeepVisionGuard variant. Results show that the co-occurrence of behavioral-image representation and temporal modelling can be helpful for accurate and interpretable insider-threat monitoring, but the performance of these systems was dependent on behavioral groups so they should be used in conjunction with human security analysis rather than in lieu of it. The framework needs to be tested across multiple organizations and data sets, with more detailed labels on the threat stages, better coverage for sparse anomaly scenarios, explore user adaptive learning and self-supervised learning, and investigate real-time deployment with mechanisms of analyst feedback.

References

1. Al-Shehari, T., Kadrie, M., Al-Mhiqani, M. N., Alfakih, T., Alsalman, H., Uddin, M., ... & Dandoush, A. (2024). Comparative evaluation of data imbalance addressing techniques for CNN-based insider threat detection. Scientific reports, 14(1), 24715.
2. AlSlaiman, M., Salman, M. I., Saleh, M. M., & Wang, B. (2023). Enhancing false negative and positive rates for efficient insider threat detection. Computers & Security, 126, 103066.
3. Álvarez, D., Pérez, L., Mateo, A., Larriva-Novo, X., Álvarez-Campana, M., & Villagra, V. A. (2025). System for prediction and early detection of insider attacks (SPEDIA) dataset (Version 2) [Data set]. Zenodo.
4. Bin Sarhan, B., & Altwaijry, N. (2022). Insider threat detection using machine learning

approach. Applied Sciences, 13(1), 259.
5. Budžys, A., Kurasova, O., & Medvedev, V. (2024). Deep learning-based authentication for insider threat detection in critical infrastructure: A. Budžys et al. Artificial Intelligence Review, 57(10), 272.
6. Cai, X., Wang, Y., Xu, S., Li, H., Zhang, Y., Liu, Z., & Yuan, X. (2024). LAN: Learning adaptive neighbors for real-time insider threat detection. IEEE Transactions on Information Forensics and Security, 19, 10157-10172.
7. Castillo Camacho, I., & Wang, K. (2021). A comprehensive review of deep-learning-based methods for image forensics. Journal of imaging, 7(4), 69.
8. Duan, S. M., Yuan, J. T., & Wang, B. (2024). Contextual feature representation for image-based insider threat classification. Computers & Security, 140, 103779.
9. Erola, A., Agrafiotis, I., Goldsmith, M., & Creese, S. (2022). Insider-threat detection: Lessons from deploying the CITD tool in three multinational organisations. Journal of Information Security and Applications, 67, 103167.
10. Gayathri, R. G., Sajjanhar, A., & Xiang, Y. (2020). Image-based feature representation for insider threat classification. Applied Sciences, 10(14), 4945.
11. Gayathri, R. G., Sajjanhar, A., & Xiang, Y. (2024). Hybrid deep learning model using SPCAGAN augmentation for insider threat analysis. Expert Systems with Applications, 249, 123533.
12. Le, D. C., & Zincir-Heywood, N. (2021). Anomaly detection for insider threats using unsupervised ensembles. IEEE Transactions on Network and Service Management, 18(2), 1152-1164.
13. Le, D. C., Zincir-Heywood, N., & Heywood, M. I. (2020). Analyzing data granularity levels for insider threat detection using machine learning. IEEE Transactions on Network and Service Management, 17(1), 30-44.
14. Li, D., Yang, L., Zhang, H., Wang, X., & Ma, L. (2022). Memory-Augmented Insider Threat Detection with Temporal-Spatial Fusion. Security and Communication Networks, 2022(1), 6418420.
15. Li, D., Yang, L., Zhang, H., Wang, X., Ma, L., & Xiao, J. (2021). Image-Based Insider Threat Detection via Geometric Transformation. Security and Communication Networks, 2021(1), 1777536.
16. Mehnaz, S., & Bertino, E. (2019). A fine-grained approach for anomaly detection in file system accesses with enhanced temporal user profiles. IEEE Transactions on Dependable and Secure Computing, 18(6), 2535-2550.
17. Pal, P., Chattopadhyay, P., & Swarnkar, M. (2023). Temporal feature aggregation with attention for insider threat detection from activity logs. Expert Systems with Applications, 224, 119925.
18. Racherache, B., Shirani, P., Soeanu, A., & Debbabi, M. (2023). CPID: Insider threat detection using profiling and cyber-persona identification. Computers & Security, 132, 103350.
19. Randive, K., Mohan, R., & Sivakrishna, A. M. (2023). An efficient pattern-based approach for insider threat classification using the image-based feature representation. Journal of Information Security and Applications, 73, 103434.
20. Song, S., Gao, N., Zhang, Y., & Ma, C. (2024). BRITD: behavior rhythm insider threat detection with time awareness and user adaptation. Cybersecurity, 7(1), 2.
21. Tian, T., Zhang, C., Jiang, B., Feng, H., & Lu, Z. (2025). Insider threat detection for specific threat scenarios. Cybersecurity, 8(1), 17.
22. Wang, Z. Q., & El Saddik, A. (2023). Dtitd: An intelligent insider threat detection framework based on digital twin and self-attention based deep learning models. IEEE access, 11, 114013-114030.
23. Xiao, H., Zhu, Y., Zhang, B., Lu, Z., Du, D., & Liu, Y. (2024). Unveiling shadows: A comprehensive framework for insider threat detection based on statistical and sequential analysis. Computers & Security, 138, 103665.
24. Yuan, S., & Wu, X. (2021). Deep learning for insider threat detection: Review, challenges and opportunities. Computers & Security, 104, 102221.