

Insider Threats and Countermeasures Based on AI Lie Detection

Konstantinos Kalodanis¹^a, Panagiotis Rizomiliotis¹^b, Charalampos Papapavlou²^c,
Apostolos Skrekas³^d, Stavros Papadimas³^e and Dimosthenis Anagnostopoulos¹^f

¹Department of Informatics & Telematics, Harokopio University of Athens, Kallithea, Athens, Greece

²Department of Electrical & Computer Engineering, University of Patras, Rio, Patras, Greece

³Department of Management Science & Technology, Athens University of Economics & Business, Athens, Greece


Keywords: AI, Lie Detection, Insider Threat, EU AI Act.


Abstract: Insider threats continue to pose some of the most significant security risks within organizations, as malicious insiders have privileged access to sensitive or even classified data and systems. This paper explores an emerging approach that applies Artificial Intelligence (AI)-based lie detection techniques to mitigate insider threats. We investigate state-of-the-art AI methods adapted from Natural Language Processing (NLP), physiological signal analysis, and behavioral analytics to detect deceptive behavior. Our findings suggest that the fusion of multiple data streams, combined with advanced AI classifiers such as transformer-based models and Graph Neural Networks (GNN), leads to enhanced lie detection accuracy. Such systems must be designed in accordance with EU AI Act, which imposes requirements on transparency, risk management, and compliance for high-risk AI systems. Experimental evaluations on both synthesized and real-world insider threat datasets indicate that the proposed methodology achieves a performance improvement of up to 15–20% over conventional rule-based solutions. The paper concludes by exploring deployment strategies, limitations, and future research directions to ensure that AI-based lie detection can effectively and ethically bolster insider threat defences.


1 INTRODUCTION


Insider threats perpetrated by individuals who hold legitimate access to organizational data or systems have become an escalating risk in modern cybersecurity environments (Sarkar and Pereira, 2022). Regardless of whether the motives for their action are driven by financial gain, deep-seated ideological beliefs, or even if it were cases of coercion, insiders are effectively in a position to take advantage of privileged access in ways that can culminate in severe repercussions for an organization's critical data. In fact, these actions have critical implications on the basic principles of data confidentiality, integrity, and availability (Bruno et al., 2021). Traditional approaches to Intrusion


Detection Systems (IDS) normally depend on predefined rules or models of anomaly detection that may not be adaptive enough to identify subtle and evolving tactics employed by insiders. Recent advances in AI have opened new avenues to uncover deceptive behaviors (Chittaranjan and Saxena, 2023). AI technologies, such as Machine Learning (ML) and NLP, offer the potential to dynamically analyse large volumes of data and detect patterns that would otherwise go unnoticed (Zhou et al., 2022). Specifically, AI-based lie-detection methodologies use information extracted from various sources, such as textual communications, physiological responses, or social network interactions, to identify patterns that are typical of deception (Kalodanis et al., 2025). While these methods have shown promise in some applications of security screening and law


^a <https://orcid.org/0000-0003-2456-9261>


^b <https://orcid.org/0000-0001-6809-9981>

^c <https://orcid.org/0000-0002-9756-3912>

^d <https://orcid.org/0009-0001-6782-8449>

^e <https://orcid.org/0009-0001-1191-7282>

^f <https://orcid.org/0000-0003-0747-4252>

^f <https://orcid.org/0000-0003-0747-4252>

enforcement (Kim et al., 2021), their potential for insider threat detection has not been thoroughly explored. In this paper, we propose a framework that integrates AI-based lie detection algorithms into organizational security monitoring, specifically detecting and deterring insider threats. Our proposal extends the existing detection systems while simultaneously enhancing their functionality by integrating an additional layer of behavioral analysis to represent a wider view of potential threats. Our primary contributions can be summed up as the following:

- 1. Novel Framework:** The present paper proposes a holistic artificial intelligence-powered lie detection architecture, that integrates text-based analytics with behavioral assessments in real time to discern questionable behavior, achieving a high level of accuracy in addressing evolving threats.
- 2. Experimental Evidence:** The proposed solution is evaluated on synthetic and real insider threat datasets, showing an important improvement in detection accuracy compared to the rule-based solutions.
- 3. Regulatory Considerations:** We discuss about how AI-based insider threat detection systems must be aligned with the requirements of the EU AI Act and address potential ethical issues to ensure that our framework is responsibly implemented.

2 PROPOSED FRAMEWORK AND METHODOLOGY

2.1 System Architecture

Our approach incorporates artificial intelligence-based lie detection modules into a more comprehensive security monitoring system. Moreover, the architecture employs a microservices-based framework that allows each component to scale independently, reducing bottlenecks in high-volume data processing scenarios (Randall, 2023). The architecture consists of the components listed below:

- 1. Data Ingestion Layer:** It is responsible for gathering artifacts from corporate networks, including logs, emails, chat transcripts, and other data. In order to ensure compliance with privacy rules and business policy, access restrictions are implemented. It utilizes secure channels and encryption protocols to maintain confidentiality and mitigate the risk of unauthorized access.

- 2. Data Preprocessing Module:** To align with EU AI Act (Kalodanis et al., 2024) provisions on data protection and transparency, this module ensures that personal data is treated with the utmost care. This module maintains compliance with the General Data Protection Regulation (GDPR) by cleaning and normalizing data and deleting Personally Identifying Information (PII) wherever it is possible to do so. Textual data is subjected to tokenization, part-of-speech tagging, and various other NLP preprocessing techniques. Moreover, advanced pseudonymization and differential privacy methods are employed.

- 3. Feature Extraction and Fusion:** The data are analyzed in order to extract linguistic, behavioral, and physiological characteristics. A few examples of important linguistic indicators are complexity measures and variations in general sentiment. Physiological signals could include the variability of the heart rate or micro-expressions captured from video data, provided that could be considered ethically acceptable. This typically means updating internal policies in order to enlighten employees about the intended use of video analysis, having strict data retention periods limits, and pseudonymising or encrypting any personal or biometric data at the earliest possible point.

- 4. AI-based Lie Detection Model:** Makes use of a neural network with multiple branches, each of which specializes in a distinct type of data (textual, behavioral, or physiological). In order to arrive at a single categorization result, attention layers combine the outputs collected from various branches. Furthermore, explainability modules are integrated into each branch, providing interpretable insights into which features contributed most to the classification outcome (Johnson et al., 2022).

- 5. Anomaly Scoring and Reporting:** Produces a score showing the degree of deceit or abnormality that occurred throughout each user session. Alerts are generated for high-risk incidents, which may be further investigated by security analysts or human resources personnel, depending on the organizational protocols. In addition, the system logs all anomaly events in a centralized dashboard, enabling long-term trend analysis and historical audit trails for compliance and forensic purposes (Nakamura, 2023).

The following Table 1 summarizes the primary function, key processes, and outputs of each component, highlighting how data flows seamlessly from ingestion to the final anomaly scoring and alerting process.

Table 1: Overview of the AI-Based Lie Detection Framework.

| Component | Primary Function | Key Processes | Output/ Result |
|------------------------------|--|--|---|
| Data Ingestion Layer | Collects raw data from corporate networks | Gathers logs, emails, chat transcripts, network artifacts - Ensures secure access controls - Applies encryption and network segmentation | Raw data (text, behavioral, physiological artifacts) |
| Data Preprocessing Module | Cleans and normalizes data, ensuring privacy compliance | Removes PII - Tokenization, part-of-speech tagging - Normalization - Employs pseudonymization and differential privacy | Preprocessed, privacy-compliant data for feature extraction |
| Feature Extraction & Fusion | Captures relevant linguistic, behavioral and physiological signals | Identifies sentiment shifts, usage patterns, micro-expressions - Combines multiple data modalities - Synchronizes multi-modal features | Combined feature vectors ready for classification |
| AI-Based Lie Detection Model | Classifies potential deception using a multi-branch neural network | Transformer network for text-Graph neural network for behavior- Fused outputs via attention layers - Integrates explainability modules | Deception probability or classification outcome |
| Anomaly Scoring & Reporting | Generates real-time risk scores and alerts | Produces anomaly/deception scores - Triggers alerts for security analysts or HR- Logs events for trend analysis | Real-time alerts, dashboards, and investigative insights |

2.2 AI Model Design

In our attempt to create a high-level AI-driven framework for the detection of deceptive behaviors, we emphasize the following core principles: modularity, scalability, and interpretability. In doing so, we ensure that every single sub-model of this system can be implemented, revised, or replaced with another without causing any disruption in the operation of the high-level pipeline connecting these components. Furthermore, we incorporate domain-adaptive training strategies to accommodate diverse organizational environments and support a wide range of feature modalities, including textual, behavioral, and physiological data.

2.2.1 Transformer-Based NLP Sub-Model

The textual sub-model leverages a transformer architecture (e.g., BERT) pre-trained on large language corpora, then fine-tuned on domain-specific insider threat data. In this phase, we integrate a domain adaptation component that uses a curated vocabulary and specialized embeddings for industry-relevant jargon, internal acronyms, and context-dependent phrases. By doing so, the model can better capture organization-specific nuances, which are often overlooked by generic language models (Cai et al., 2023). We propose a multi-task learning setting to predict both honesty vs. deception labels and user sentiment. This multi-task framework not only

enhances generalization but also provides insights into the user's emotional state, which gives extra context for deception detection. Especially, sentiment trajectories—polarity shifts either to positivity or negativity over time—may act as warning signals that suggest increased cognitive load or stress. Such an approach ensures the extraction of subtle linguistic features that may indicate deceptive intent.

2.2.2 Behavioral Graph Neural Network

Because insider threats often involve collusive or structured groups within an organization, we use a graph neural network to model social interactions and user-access patterns. Nodes represent individuals or systems, and edges represent communication frequency or system usage similarity. Also, we document temporal characteristics, including the duration of contacts and their frequency, to describe how the population of users dynamically changes over time and capture any abnormal connectivity spikes. The time component is helpful to detect short but massive communication bursts, which might reveal clandestine planning. GNN embeddings capture topological relationships that might reflect potential collusion or unusual activity. We further incorporate hierarchical pooling techniques that allow the model to learn from both local cliques (e.g., small collaborating groups) and global structures (e.g., department-wide interaction patterns), enhancing our ability to spot more subtle threats.

These advanced pooling methods help avoid the pitfalls of information dilution in large graphs (Moradi et al., 2023). Moreover, we employ explainable GNN mechanisms that highlight critical subgraphs and edge connections driving the model's decision, aiding cybersecurity analysts in root-cause analysis. This interpretability is especially important in organizational settings where audits and compliance checks require clear justification of any flagged behavior.

2.2.3 Fusion and Attention Layers

The outputs from both sub-networks, along with any available physiological signals, flow into a fusion layer. In this fusion process, we apply modality-specific gating functions to regulate how much information from each source contributes to the combined embedding. This approach helps manage imbalances in data quality and quantity across textual, behavioral, and physiological streams. Attention mechanisms prioritize more salient features, enabling the combined model to make a final deception probability estimate. These attention layers go beyond the standard additive approach by employing multi-head attention, which captures different facets of the data in parallel—e.g., comparing linguistic cues to user graph connectivity, or correlating physiological spikes with real-time communication anomalies. Such multifaceted attention reduces the risk of missing essential signals masked by noise in any single modality (Akhter and Machado, 2022). A threshold-based approach flags potentially deceptive sessions. For high-risk cases, the system triggers an alert and logs a comprehensive summary of which fused features most influenced the decision, thereby facilitating swift human-led investigation. Additionally, automated feedback loops allow security specialists to refine these threshold values over time, tuning the model to each unique organizational environment.

2.3 Implementation Details

Our proof-of-concept was developed using Python (v. 3.11) and the PyTorch library for the deep learning modules as well as the PyTorch Geometric package for GNN-based modules. System microservices were Dockerized for portable deployment on many server environments. The textual sub-model was fine-tuned from a GPT-like pretrained language model. Hyperparameter optimization—learning rate, batch size, and embedding dimensionality—used grid search and five-fold cross-validation.

3 EXPERIMENTAL SETUP AND RESULTS

This section details our experimental methodology, including the datasets used, the performance metrics selected, and the procedure followed to train and evaluate our proposed AI-based lie detection framework. Also, we compare our approach against established baselines and highlight scenarios in which our fusion model excels (Zhang and Wu, 2022).

3.1 Datasets

We conducted experiments on two datasets with different characteristics and complexities:

1. Synthetic Insider Dataset (SID): Created by simulating insider threat scenarios in a controlled test environment. It contains 10,000 user sessions with labelled deceptive or benign actions. Each user session includes mock communications, system logs, and pre-defined user roles to mimic actual organizational structures, ensuring that both collusive and single-actor deception attempts are accurately represented. The deception labels in SID were assigned through a scripted storyline, with multiple reviewers verifying scenario consistency before final labelling (Williams et al., 2023). It's developed in English.

2. Real-world Insider Threat Dataset (RITD): Aggregated from an organizational email and chat system. Anonymized for privacy, the dataset includes textual, behavioral, and partial physiological signals (where legally permitted). Over 25,000 user sessions were manually annotated or semi-automatically labelled via a rule-based approach. This dataset captures authentic interactions among employees, encompassing formal communications (e.g., work-related emails) and informal dialogues (e.g., instant messages). Physiological signals were only included for specific job roles and countries where consent and ethical clearance were obtained, thus reflecting realistic enterprise data constraints. Unlike SID, RITD presented a more challenging evaluation setting due to its unstructured nature and potential noise in annotations. Employees engaged in natural communication patterns, meaning deceptive actions were interspersed with non-malicious behavior, requiring the model to differentiate between genuine and deceptive anomalies. The RITD dataset was obtained through a collaboration with a university that consented to share anonymized logs (emails, chat transcripts) and limited physiological data under strict

legal and ethical agreements. Employee consent procedures and secure data-transfer protocols were enforced from the outset. All personal identifiers were hashed or pseudonymized before data reached our research environment ensuring compliance with privacy regulations and internal corporate policies governing sensitive data sharing. Labels for RITD sessions were determined via a hybrid approach. First, automated anomaly detection heuristics flagged potentially suspicious communications or access patterns. Next, a panel of security professionals reviewed these flagged instances while adhering to privacy-by-design principles—ensuring that only the minimal necessary information was accessed to confirm or dismiss an insider-threat label. This dual-layered process not only improved labelling accuracy but also maintained compliance with corporate policies and GDPR-like regulations.

3. Availability of Datasets: We provide detailed methodological descriptions, configuration parameters, and performance metrics to facilitate replicability of our approach.

The Table 2 below offers an overview of the two datasets along that used in the experiments.

3.2 Performance Metrics

We evaluated our model using accuracy, precision, recall, and F1-score. Accuracy provides an overall measure of correct classifications, while precision and recall ensure we capture the model’s ability to correctly identify deceptive instances without generating excessive false alarms. Additional metrics included the Area Under the Receiver Operating Characteristic Curve (AUROC) to measure discriminative power. AUROC is particularly relevant in imbalanced settings, where focusing solely on accuracy could be misleading.

3.3 Experimental Procedure

1. Data Preprocessing: Text was tokenized with a BERT-compatible tokenizer, while behavioral logs were converted into graphs. To ensure privacy compliance, user identifiers were replaced with hashed tokens, and sensitive text was masked where appropriate. The behavioral graph construction incorporated weighted edges based on communication frequency, allowing for nuanced detection of abrupt changes in interaction patterns.

2. Training and Validation: We used a stratified 80/10/10 split for training, validation, and testing. This split was carefully chosen to preserve the ratio of deceptive vs. benign user sessions in each subset, minimizing sampling bias. We conducted a grid search over hyperparameters such as learning rate, batch size, and attention heads in the transformer model. The GNN module was similarly optimized for the number of graph layers and node embedding dimensions (Alhassan and Frolov, 2023).

3. Baseline Models: We compared our approach against: A rule-based system using keyword matching and threshold-based anomalies. Keywords included terms commonly associated with malicious intent, while threshold logic flagged unusual login times and file access counts. A standard LSTM-based deception detection model without multimodal fusion. This was trained solely on textual data, providing a benchmark to assess the added value of the GNN and fusion components (Kim et al., 2022).

4. Deployment and Interpretability: For real-time applications, we deployed the best-performing model as a microservice accessible via REST APIs, enabling seamless integration within the existing security infrastructure (Lieberman and Tsung, 2023). Additionally, we incorporated explainability modules that generate feature-attribution heatmaps and subgraph importance summaries, assisting security analysts in rapid root-cause analysis of flagged sessions.

Table 2: Overview of the two datasets used in experiments.

| Dataset | Number of User Sessions | Data Types | Annotation Method | Key Characteristics |
|---------|-------------------------|---|---|---|
| SID | 10,000 | Textual logs simulated behavior patterns | Script-based labeling, reviewed by multiple experts | Controlled environment; covers staged collusion, single-actor deception; thorough storyline validation |
| RITD | 25,000+ | Textual (email, chat), partial physiological signals, behavioral logs | Manual annotation + semi-automatic labeling via rule-based approach | Real-world organizational data; anonymized for privacy; includes diverse roles, legal constraints, and country-specific regulations |

Table 3: Performance comparison across different models.

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) | AUROC |
|----------------------------|--------------|---------------|------------|--------------|-------|
| Rule-based System | 78.2 | 76.9 | 71.5 | 74.1 | 0.78 |
| LSTM Deception Model | 83 | 82.4 | 80.1 | 81.2 | 0.85 |
| Proposed Transformer + GNN | 94.8 | 92.3 | 94.0 | 93.1 | 0.97 |

3.4 Simulation Results

Table 3 summarizes the comparison of performance metrics such as percentage (%) of accuracy, precision, recall, F1-score and finally AUROC on the test sets. Our fusion-based approach outperformed all baseline models and showed a strong capacity in handling such subtleties of linguistic cues and complex behavioral interactions. This gap in performance was more pronounced, especially in scenarios of multi-user collusion or disguising behavior patterns by using multiple communication channels (Swenson and Guerrero, 2022).

As shown, the proposed system achieved a 15–20% improvement over conventional solutions in F1-score and AUROC, demonstrating robust detection of deceptive behavior. The enhanced performance was especially apparent in instances of subtle deceit, including indirect communication via corporate chat programs, email exchanges containing false assertions, or intentionally timed actions intended to replicate regular user behavior. Involving both textual and behavior information played an important role in enhancing the effectiveness of the model in detecting previously elusive deceitful actions, and in proving the value of using deep architectures specifically designed for individual types of information.

In addition, graph-based techniques effectively uncovered hidden relations between entities, which, in return, amplified collusion and deviation over a period. As a result, they proved to be more efficient than traditional rule-based approaches, which often failed to detect anomalous behavior out of predefined borders and focused predominantly on keyword-based heuristics. In addition, minimizing false positive rates helped security operations prioritize high-confidence alerts, and therefore enhanced overall effectiveness in preventing insider attack. Moreover, our ablation studies revealed that removing either the transformer-based NLP sub-model or the GNN-based behavioral model

significantly degraded performance, confirming the essential role of multimodal fusion.

4 DISCUSSION

The following section explores the wider ramifications of our findings, analyzing the technological insights obtained from our experiments as well as the ethical and regulatory aspects essential for implementing AI lie detection in practical business environments. We also highlight key limitations and outline future work directions, focusing on generalizability, privacy-preserving mechanisms, and more inclusive organizational contexts.

4.1 Technical Insights

Our experimental results underline the critical need to integrate diverse data sources in order to cover the whole insider threat behavior space. Integration of textual, behavioral, and constrained physiological indicators significantly raises the robustness of the model, where each data modality helps to compensate for deficiencies or perturbations of the other ones. For instance, while textual features provide fine-grained insights into deceptive language patterns, the behavioral (graph-based) features reveal larger-scale patterns of collusion or unusual user interactions.

Additionally, the multi-task learning approach in the transformer-based NLP sub-model improved performance by leveraging auxiliary sentiment signals. By jointly predicting sentiment and deception, our model learns to detect subtle linguistic cues, such as sudden shifts in tone or emotion-laden terms, which might correlate with deceptive intent. Our ablation study indicated that removing the sentiment prediction branch resulted in a notable drop in recall, suggesting that emotional context often complements deception indicators. Furthermore,

explainable AI (XAI) techniques implemented in the transformer sub-model made it possible to pinpoint specific phrases and words that influenced the deception classification. This interpretability is vital for trust-building and ensures that security teams can follow the rationale behind flagged communications (Tani et al., 2023).

4.2 Ethical and Regulatory Considerations

While the accuracy gains are promising, deploying AI lie detection in the workplace raises salient questions about privacy, consent, and potential biases. In addition to these concerns, the EU AI Act introduces specific obligations regarding transparency, data governance, and human oversight for high-risk AI applications, which may include workplace surveillance and insider threat detection. General Data Protection Regulation (GDPR) imposes stringent guidelines on the processing and storing of personal data, mandating transparent data flows and lawful bases for data collection. Moreover, employee consent must be obtained in many jurisdictions, and workers' councils or unions frequently request documentation detailing how AI-driven surveillance affects labor rights (Russo and Forti, 2022). Our design incorporated privacy-preserving techniques such as data minimization and user pseudonymization. Specifically, we hashed direct user identifiers and redacted sensitive textual content not necessary for deception detection in accordance with the GDPR's data minimization principle. We also only allowed access through role-based access controls, granting the possibility to trace anomalies back to a specific subject only to the extent that users had a legitimate interest in accessing such raw data.

It is important to note that pseudonymization alone can be insufficient to protect employee identities in certain scenarios. To counter this threat, we advocate a multi-layered privacy strategy going beyond simple pseudonymization. Departmental names, for instance, can be substituted by more abstract-coded categories or by random sets in attempts to protect individual departments from being correlated to concrete behavior patterns. Furthermore, the use of techniques like k-anonymity or differential privacy can subsequently also reduce the threat of retracing pseudonymized information back to individuals while retaining aggregate trends required for insider threat detection. By combining these stronger anonymization methods organizations can minimize re-identification risks and maintain a balanced approach between robust security

monitoring and the fundamental privacy rights of employees.

Additionally, a formalized ethics review process—potentially involving third-party auditors—is recommended before large-scale deployment to ensure compliance with emerging legal frameworks and ethical standards (Baker and McFadyen, 2023). Such audits should also evaluate the system's alignment with EU AI Act requirements, particularly in terms of its risk classification, record-keeping practices, and the clarity of its decision-making processes, thus offering greater transparency and user trust.

4.3 Limitations and Future Work

This paper of its current approach does not deal with the cross-linguistic and cross-cultural dimensions of deception which could vary considerably across different geographical locations. Deception signals and dynamics of trust among colleagues could be influenced by local linguistic rules, cultural norms, and regional labor laws, necessitating the development of globally adaptable lie detection systems. To achieve this, future models should be able to handle multilingual and code-switching contexts, incorporating regional colloquialisms from different geographic regions through comprehensive lexical databases. The paper also highlights ethical concerns surrounding the use of physiological data, emphasizing privacy implications and psychological effects on employees under surveillance in their workplace and furnish biometric information. Despite the potential of multimodal deception analytics, significant legal and ethical scrutiny is required.

Future research efforts will be devoted to extending the proposed framework by incorporating multilingual domain adaptation, next-generation explainable AI modules for greater transparency, and conducting live experiments to assess practical visibility. Last but not least, real corporate setting pilots, after careful ethical and legal screening, will provide more insight into user adoption, model drift, and the short-term effectiveness of our methodology to guide further improvements (Ren et al., 2023).

5 CONCLUSIONS

This paper presented an AI-based lie detection framework for insider threat detection, emphasizing a multimodal fusion of textual, behavioral, and physiological data. Our solution leverages transformer architectures and graph neural networks

to uncover deception in real time, outperforming baseline methods on both synthetic and real-world datasets. While the results are encouraging, careful consideration of privacy, ethics, and regulatory compliance is imperative. AI-based lie detection can serve as a powerful complement to human analysts—provided it is designed and deployed responsibly.

A key conclusion is the adaptability and expandability of the proposed framework. This research supports AI-based lie detection as a viable strategy for handling insider threats. Nonetheless, the benefits of such technology can only become a reality through constant technological improvements, compliance with protective legal frameworks, and continued workers' trust. AI-powered lie detection cannot be considered an autonomous, standalone remedy but works as a tool that, when utilized wisely, can contribute immensely towards security and stability in an institution.

REFERENCES

- Akhter, R., & Machado, C. (2022). Multi-head attention for cross-modal alignment in deception detection. *Neurocomputing*, 508, 364–377.
- Alhassan, K., & Frolov, S. (2023). Auto-tuning hyperparameters in GNN-based insider threat detection. *IEEE Transactions on Network and Service Management*, 20(1), 14–25.
- Baker, L. J. & McFadyen, S. (2023). Bioethical perspectives on AI-based lie detection: Implications for labor rights. *AI and Society*, 38, 1023–1041.
- Bruno, A., Rossi, F., & D'Angelo, S. (2021). Challenges in insider threat detection: The hardest nuts to crack. *Computers & Security*, 117, 102345.
- Bruno, A., Rossi, F., & D'Angelo, S. (2021). Strengthening data confidentiality, integrity, and availability in enterprise networks: A systematic review. *Computers & Security*, 106, 102265.
- Cai, Y., Tang, H., & Zhao, F. (2023). Domain adaptation techniques for transformer-based language models in secure communication. *Expert Systems with Applications*, 223, 119007.
- Chittaranjan, G., & Saxena, S. (2023). Uncovering deceptive behaviors in insider threat detection: A context-aware AI approach. *Computers & Security*, 129, 104118.
- Johnson, T., Russel, A., & Kwok, A. B. (2022). Interpretable neural networks for deception detection in textual data. *Neural Computing and Applications*, 34, 4567–4578.
- Kalodanis, K., Rizomiliotis, P., Feretzakis, G., Papapavlou, C., & Anagnostopoulos, D. (2025). High-Risk AI Systems—Lie Detection Application, *Future Internet*, 17(1), 26.
- Kalodanis, K., Rizomiliotis, P., & Anagnostopoulos, D. (2024). European Artificial Intelligence Act: an AI security approach, *Information and Computer Security: Volume 32 Issue 3*.
- Kim, D., Yoon, S., & Park, T. (2022). Text-only vs. multi-modal approaches in deception detection: A comparative study. *Neural Computing and Applications*, 34, 14457–14468.
- Lieberman, T., & Tsung, W. (2023). Design and deployment of microservices for real-time threat monitoring and mitigation. *IEEE Internet of Things Journal*, 10(5), 4250–4261.
- Moradi, F., & Huang, Y. (2023). Differential privacy techniques in next-generation corporate monitoring systems. *Computers & Security*, 123, 102010.
- Nakamura, K. (2023). Comprehensive anomaly detection logging for security analytics. *Expert Systems with Applications*, 213, 118966.
- Randall, E. S. (2023). AI frameworks for multi-modal data ingestion in corporate security. *Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security*.
- Ren, X., Blum, C., & Park, D. (2023). Live deployment of multimodal deception detection systems: A case study in a multinational corporation. *Proceedings of the 22nd IEEE International Conference on Trust, Security and Privacy in Computing and Communications*.
- Russo, A., & Forti, V. (2022). Bridging compliance and innovation: GDPR challenges in AI-driven workplace solutions. *Computer Law & Security Review*, 45, 109827.
- Sarkar, K., & Pereira, S. (2022). Insider threat detection in modern cybersecurity environments: A risk-based approach. *Computers & Security*, 125, 102016.
- Swenson, K., & Guerrero, I. (2022). Multi-channel data fusion for detecting collusive threats in enterprise networks. *Computers & Security*, 123, 102002.
- Tani, T., Moreira, D., & Khoueiry, R. (2023). Explainable AI in high-stakes security applications: Visualizing deception cues. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 13(2), 19.
- Williams, D., Smith, O., & Dominguez, M. (2023). Synthetic data generation for insider threat modeling. *Proceedings of the 16th ACM Conference on Data and Application Security and Privacy*.
- Zhang, K., & Wu, E. (2022). Comprehensive validation strategies for insider threat detection frameworks. *IEEE Access*, 10, 73200–73215.
- Zhou, C., Wang, L., & Cohen, E. (2022). Expanding deception detection through multimodal physiological signal analysis. *Computers & Security*, 120, 102823.