



AI-Based Intelligent Exam Paper Leak Detection and Prevention Framework Using Machine Learning and Cybersecurity Techniques: A Focused Review and Conceptual Framework

Tanvi Mahesh Waghchaure, Prof. Bangar Abhishek Popat, Prof. Bhosale S.B.

Dept. of Computer Engineering

Jaihind College of engineering, Pune, India

Abstract. Exam paper leaks damage the fairness and credibility of examinations, and most leaks involve people who legitimately hold access to the paper, such as setters, moderators, printers, couriers and administrators. This makes the problem closely related to malicious insider threat detection and to data exfiltration control. This paper presents a focused narrative review of verified sources on machine learning (ML) and cybersecurity techniques that could support exam paper leak detection and prevention. It examines three connected themes: insider threat categories and detection approaches, ML and deep learning models for intrusion and anomaly detection, and preventive cybersecurity controls such as encryption and threshold secret sharing. The literature shows that behaviour-based, anomaly and hybrid ML methods are widely used for insider threat detection, that a hybrid explainable deep learning model reached 97.11% accuracy over a 37-class combined test set of 1,957,876 samples, and that an ensemble of five classical ML models reached accuracy, precision, recall and F1 score of 99.4%, 99.6%, 99.7% and 99.7% on a cleaned KDD Cup 1999 dataset. However, none of the reviewed sources studies exam paper leaks, the datasets are network or simulated insider data, rare classes remain hard to detect, and monitoring staff raises privacy and legal questions. The review contributes a comparative analysis of sources, datasets and reported results, a proposed conceptual seven-layer framework linking secure custody, key release control, activity logging, behaviour analysis, explainable alerts, response and audit, and a prioritised list of research gaps.

Keywords: Exam Paper Leak, Insider Threat Detection, Machine Learning, Intrusion Detection, Anomaly Detection, Cyber Security, Secret Sharing

I. Introduction

Examinations depend on the confidentiality of question papers until the moment of the test. A leak undermines fairness for all candidates, and in India the Public Examinations (Prevention of Unfair Means) Act, 2024 treats paper leaks and other unfair means in specified public examinations as offences [1]. Leaks usually do not require an outside attacker to break in: the paper passes through many people who are authorised to see it, so the core risk is misuse of legitimate access.

That risk matches the definition of an insider threat, which a recent survey describes as a risk from individuals with authorised access to an organisation's systems, data or facilities, including data theft, fraud, sabotage and espionage [2]. The same survey



reports, from a 2023 breach investigations report, that insiders were responsible for 35% of data breaches, and states that malicious insiders cause far greater damage than external actors [2]. Detection systems based on ML are now widely proposed for this problem [2]–[4]. This review therefore asks which ML and cybersecurity techniques are mature enough to support exam paper leak detection and prevention, and proposes a conceptual framework that combines them. Treating exam leaks as an insider-threat and exfiltration problem is an inference of this review, not a finding of the sources.

II. Review Methodology

This is a focused narrative review, not a systematic one. Three research sources were provided: a survey of malicious insider threat detection using ML [2], a hybrid explainable deep learning intrusion detection framework for IoT and industrial systems [3], and a hybrid intrusion detection and ensemble learning system for DevOps continuous monitoring [4]. None of them studies exam paper leaks. Three standard sources were added for the preventive side: Shamir's secret sharing scheme [5], the NIST Advanced Encryption Standard [6], and the Indian examinations law [1]. Descriptions and numbers from [2]–[4] are taken as reported and were not reproduced. No separate database search was run and no study counts are claimed. Evidence specific to exam leaks is therefore absent, which is stated as a limitation in Section VIII.

III. Exam Paper Leaks as an Insider Threat Problem

1. Categories of insider threats

The survey categorises insider threats by intent (malicious, accidental or fraudulent), by access level (privileged or non-privileged) and by duration (isolated or persistent) [2]. A paper setter who deliberately shares questions is a malicious privileged insider; a staff member who emails the paper to the wrong address is an accidental one; and a leak sustained over several exam cycles is a persistent one. Misconfiguration and lack of awareness are named as typical causes of accidental incidents [2].

2. Challenges that apply to exam custody

Several challenges described in [2] map to exam paper handling. Credential misuse is hard to detect because insiders who hold valid credentials blend in with routine tasks. Insider collaboration, where insiders coordinate with outsiders or share credentials, is not reliably caught by existing anomaly detection. Encrypted channels hide the content of transfers, so detecting exfiltration depends on traffic patterns and endpoint visibility. Sudden behaviour changes are difficult to separate from legitimate ones, and delay in detection prolongs exposure [2]. For exam papers the time window is short, so detection latency matters more than in many other settings.

3. Preventive controls

Prevention complements detection. The Advanced Encryption Standard specifies a symmetric block cipher that can protect papers at rest and in transit [6]. Shamir's secret sharing divides a secret into shares so that a chosen threshold of shares reconstructs it while fewer reveal nothing about it [5]. Used for exam papers, the decryption key could be released only when a threshold of custodians combine their shares at the scheduled time, so that no single person can open the paper early. This is a design proposal and is not tested in the reviewed sources.

IV. Machine Learning for Detection

1. Detection approaches and learning methods

The survey groups detection approaches into behaviour-based, rule-based, anomaly, signature-based and heuristic methods [2]. Behaviour-based methods build models of normal user behaviour and flag deviations, but they can produce false positives when legitimate behaviour changes. Rule-based and signature-based methods are effective against known patterns but struggle



with novel threats [2]. Supervised learners such as logistic regression, decision trees, random forests and SVM need labelled incidents, which are scarce; unsupervised methods such as k-means, DBSCAN and isolation forests find new patterns but often give more false positives; deep models such as RNN, LSTM, CNN and graph networks capture complex and temporal patterns but are hard to interpret [2]. Fusion of behaviour, system-log and network data is reported as a trend [2].

2. Data sources

Useful data include network logs, user behaviour data such as login histories and file access logs, and textual sources such as emails [2]. Public datasets described in [2] include the CERT insider threat database, the TWOS dataset from a gamified competition, and the RUU masquerade dataset (Table II). One study cited in [2] describes the CERT data as synthetic.

3. Hybrid and explainable models

NOVA-XIDS combines autoencoder embeddings from three benchmark datasets in a shared 64-dimensional latent space with a residual CNN using CBAM attention, an LSTM for temporal modelling and self-attention, and provides SHAP-based and attention-based explanations and an ONNX-accelerated real-time pipeline with a Streamlit dashboard [3]. HIDEELS evaluates five classical models (logistic regression, k-nearest neighbours, decision tree, naive Bayes and SVM) on KDD Cup 1999 and combines them in an ensemble, using SelectFromModel feature selection that was more than 93% faster than recursive feature elimination and kept 34 features [4]. Explainability is also named in [2] as a requirement for trust in security systems.

V. Reported Evidence

NOVA-XIDS was evaluated on a combined test set of 1,957,876 samples across 37 classes, after training on more than 15 million samples with a 70/15/15 split; it reached 97.11% accuracy, 96.14% weighted F1 score, a macro ROC-AUC of 0.9965 and top-3 and top-5 accuracy of 99.92% and 99.96% [3]. The authors report F1 scores above 0.99 for high-volume attacks, but low recall for rare classes such as SQL injection, MITM ARP spoofing and recon ping sweep, which they attribute to extreme class imbalance [3].

HIDEELS used KDD Cup 1999, whose 42 features and nearly half a million instances were cleaned to 145,586 instances, with samples of 1k, 5k, 10k and 50k and the full set; the task was reduced to normal versus malicious traffic [4]. The decision tree was the best single model, with accuracy, precision, recall and F1 score of about 99.5%, 99.5%, 99.7% and 99.6%, and the ensemble reached 99.4%, 99.6%, 99.7% and 99.7% [4]. The differences are small, and the ensemble's accuracy is slightly below the decision tree's. In [4], normal traffic forms more than half of the data and one attack type, Neptune, more than a third, so the data are imbalanced.

VI. Comparative Analysis

Table I compares the sources, Table II the datasets and Table III the reported results. The results come from different datasets and tasks (37-class classification in [3], binary classification in [4]) and are not directly comparable.

Table 1: Comparison of Sources Relevant to Exam Paper Leak Detection and Prevention

Source	Type	Focus	Key content	Limitation
Alzaabi and Mehmood, 2024 [2]	Survey	Malicious insider threat detection with ML	Threat categories; behaviour, rule, anomaly, signature and heuristic approaches; ML/DL methods; datasets; metrics; challenges and trends	Broad survey; no exam-specific evidence



Gupta et al., 2026 [3]	Research article	Explainable IDS for IoT and industrial networks	Latent autoencoder embeddings, CBAM residual CNN, LSTM, self-attention; 37 classes; accuracy 97.11%	Network traffic, not user activity; rare classes weak
Ababneh et al., 2026 [4]	Research article	Ensemble IDS for DevOps monitoring	Five ML models and an ensemble on KDD Cup 1999; decision tree accuracy about 99.5%	Older benchmark; small differences between models
Shamir, 1979 [5]	Foundational paper	Secret sharing	Secret divided into shares; a threshold of shares reconstructs it	Not an ML method; design use is proposed here
NIST FIPS 197 [6]	Standard	Encryption	Specifies the Advanced Encryption Standard	Protects data; does not detect misuse
Public Examinations Act, 2024 [1]	Law (India)	Unfair means in public exams	Treats paper leaks and unfair means in specified public examinations as offences	Scope limited to specified examinations

No source in this set studies exam paper leaks; see Section VIII.

Table 2: Datasets Described in the Reviewed Sources

Dataset	Source	Content	Scale / details	Note
CERT insider threat	[2]	Logs and data on insider threat incidents, from the CERT Insider Threat Center	Versions such as r4.2, r5.2 and r6.2 used in studies in [2]	Described as synthetic in one cited study
TWOS	[2]	Masquerader and traitor behaviour from a gamified competition	Not given in [2]	Includes defensive and offensive strategies
RUU	[2]	Host-based events with masquerade sessions	34 regular users; 14 masqueraders	Host-level, not exam-specific
KDD Cup 1999	[4]	Network traffic with 22 attack types and normal traffic	42 features; nearly half a million instances; 145,586 after cleaning	Normal over 50%; Neptune over one third
CSE-CIC-IDS2018, CIC-BCCC-NRC-IoMT-2024, CIC-IIoT-2025	[3]	Three benchmark IDS datasets merged into a unified set	37 classes; training over 15 million; test 1,957,876	Network flows; heavy long-tail classes

Details as reported in [2]–[4]; not independently verified.

Table 3: Reported Performance of Detection Models

Model (task)	Accuracy	Precision	Recall	F1 score / other
NOVA-XIDS [3] (37-class combined test)	97.11%	Not stated	Not stated	Weighted F1 96.14%; macro ROC-AUC 0.9965
HIDELS decision tree [4] (normal vs malicious)	about 99.5%	about 99.5%	about 99.7%	F1 about 99.6%
HIDELS ensemble [4] (normal vs malicious)	about 99.4%	about 99.6%	about 99.7%	F1 about 99.7%

VII. Proposed Conceptual Framework

Fig. 1 presents a proposed conceptual framework derived from the reviewed sources; it is conceptual and has not been experimentally validated. Papers are kept encrypted with limited copies (layer 1) [6], and the decryption key is released only when a threshold of custodians combine their shares (layer 2) [5]. Access, device, login, email and network activity around the paper is logged (layer 3), which matches the data sources named in [2]. Behaviour and anomaly models, for example sequence models and autoencoders on user and entity behaviour, analyse these logs (layer 4) [2], [3]. Each alert carries ranked reasons, as with the SHAP-based attribution in [3], so that investigators can judge it (layer 5). The response layer suspends access, rotates keys or switches to a backup paper set (layer 6), which is important because of the short time window. Finally, a human review with records and legal and privacy checks closes each case (layer 7) [1], [2], and confirmed cases are used to retrain the models, since labelled incidents are scarce [2].

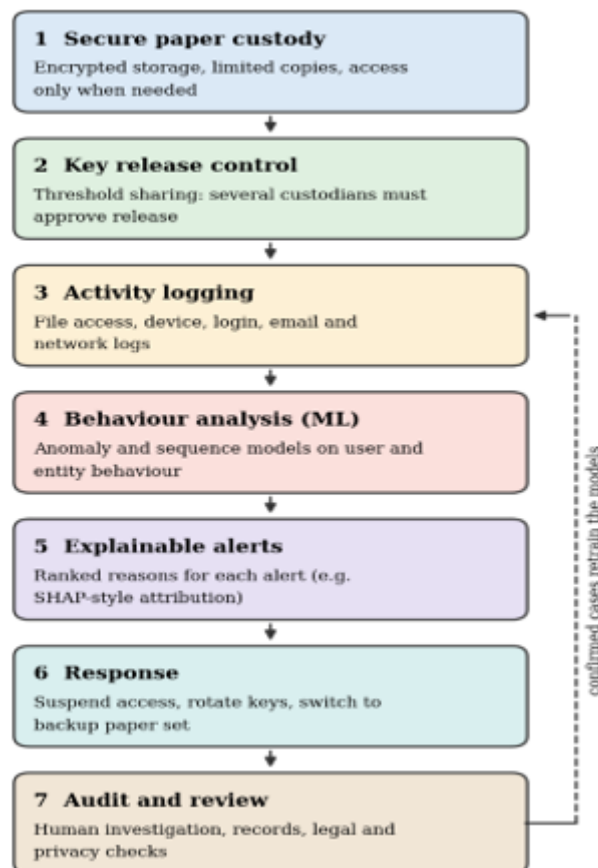


Fig. 1. Proposed conceptual framework for exam paper leak detection and prevention (conceptual, not experimentally validated).



VIII. Research Gaps and Future Directions

Gaps supported by the reviewed evidence. (i) No exam-specific evidence: none of the sources studies exam paper leaks, and the link to insider threat is an inference of this review. (ii) Data mismatch: the datasets are network traffic or simulated insider data, and one cited study describes the CERT data as synthetic [2]; KDD Cup 1999 is an older benchmark [4]. (iii) Class imbalance: rare attack classes show low recall in [3], and the data in [4] are dominated by normal traffic and one attack type. (iv) Small differences: the decision tree and ensemble results in [4] differ by about 0.1 to 0.2 points, and the ensemble's accuracy is slightly lower.

(v) Reporting: in [3] one overall performance plot is described in the text as a placeholder. (vi) Privacy and legal limits: monitoring staff raises privacy and compliance issues [2]. (vii) Limited coverage of this review: it relies on three research sources.

Directions requiring investigation. Priorities include ethically approved simulation datasets of exam paper custody with labelled leak scenarios; evaluation against credential sharing and insider collaboration with outsiders, which are named as hard cases in [2]; multimodal fusion of access, device and network logs [2]; federated or privacy-preserving training so that sensitive logs need not be centralised [2]; measurement of detection latency, which [3] addresses with an ONNX real-time pipeline; pilots with examination bodies; and testing threshold key release together with detection.

IX. Conclusion

Exam paper leaks can be treated as a problem of insider misuse and data exfiltration, for which ML-based behaviour analysis, hybrid and explainable models, encryption and threshold key release are all available building blocks. Reported detection results are high on benchmark datasets, but those datasets are network or simulated data, rare classes are weak, differences between models are small, and no reviewed source studies exam papers. A layered framework that combines secure custody, key release control, logging, behaviour analysis, explainable alerts, response and audit, as outlined in the proposed framework, is a reasonable design direction but needs empirical validation with realistic exam-custody data and with privacy and legal safeguards.

References

1. Government of India, "The Public Examinations (Prevention of Unfair Means) Act, 2024."
2. F. R. Alzaabi and A. Mehmood, "A review of recent advances, challenges, and opportunities in malicious insider threat detection using machine learning methods," *IEEE Access*, vol. 12, pp. 30907–30927, 2024, doi: 10.1109/ACCESS.2024.3369906.
3. S. Gupta, D. K. Verma, S. K. Srivastava, and C. Joshi, "An explainable latent-attention hybrid deep learning framework for high-fidelity intrusion detection in IoT and industrial cyber systems," *IEEE Access*, vol. 14, pp. 90608–90629, 2026, doi: 10.1109/ACCESS.2026.3701636.
4. J. Ababneh et al., "Enhancing DevOps continuous monitoring phase: Hybrid intrusion detection and ensemble learning system (HIDEELS)," *IEEE Access*, vol. 14, pp. 4732–4755, 2026, doi: 10.1109/ACCESS.2026.3650793.
5. A. Shamir, "How to share a secret," *Commun. ACM*, vol. 22, no. 11, pp. 612–613, Nov. 1979.
6. National Institute of Standards and Technology, "Advanced encryption standard (AES)," Federal Information Processing Standards Publication 197, Nov. 2001. [7] D. E. Denning, "An intrusion-detection model," *IEEE Transactions on Software Engineering*, vol. SE-13, no. 2, pp. 222–232, Feb. 1987, doi: 10.1109/TSE.1987.232894.
7. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proceedings of the Eighth IEEE International Conference on Data Mining*, 2008, pp. 413–422, doi: 10.1109/ICDM.2008.17.



8. National Institute of Standards and Technology, "Security and Privacy Controls for Information Systems and Organizations," NIST Special Publication 800-53, Rev. 5, Sep. 2020. [Online]. Available: <https://doi.org/10.6028/NIST.SP.800-53r5>.