



An Intelligent Risk-Adaptive Framework for Secure Examination Paper Distribution, Traceability, and Early Leak Detection

Mr. Swapnil S. Chaudhari¹, Asst. Prof. Sapana A. Fegade²

Research Scholar, Department of Computer Applications, Shramasadhana College of Engineering & Technology
Jalgaon.MH, India¹

Assistant Professor, Department of Computer Applications, Shramasadhana College of Engineering & Technology
Jalgaon.MH, India²

Abstract: Examination-paper leakage threatens the fairness, credibility, and operational continuity of academic and competitive examinations. Conventional protection mechanisms rely heavily on controlled physical handling, restricted access, and manual supervision; digital distribution reduces some transportation risks but introduces new requirements for encryption, authentication, key management, traceability, and monitoring. This research presents a risk-adaptive examination-security framework that integrates four complementary capabilities: secure digital distribution, tamper-evident chain-of-custody tracking, semantic leak detection, and risk-based security control selection. The study follows a design-science and applied-experimental approach. The proposed architecture separates a Java/Spring Boot security core from a Python-based natural-language-processing service. Question papers are protected through encryption, controlled release, copy-level fingerprinting, and auditable access events. Suspected public-source leaks are represented as a semantic-search problem in which original questions are compared with noisy, partial, paraphrased, OCR-derived, and multilingual variants using sentence embeddings. A risk engine maps examination characteristics to an interpretable risk score and recommends stronger controls for higher-risk examinations. The empirical evidence retained from the study is a pilot analysis of 66 documented Indian examination-paper-leak incidents from 2014-2024. Logistic Regression and Random Forest were evaluated using five-fold stratified cross-validation against a majority-class baseline. The baseline achieved an F1-score of 0.842, compared with 0.674 for Logistic Regression and 0.739 for Random Forest, indicating that the small incident dataset and limited features are insufficient for reliable prediction of incident-confirmation status. This classification task is treated as a pilot study rather than as evidence that future examination risk can already be predicted reliably. The manuscript therefore defines a reproducible experimental protocol for evaluating semantic leak detection and the complete risk-adaptive system without fabricating unperformed results. A working prototype implementing all four modules - the Java/Spring Boot security core with role-based access control, the SHA-256 hash-linked chain of custody, a Python microservice offering both a TF-IDF baseline and Sentence-BERT similarity, and the risk-scoring engine - was built and verified through fourteen automated checks on the core security logic and an end-to-end run confirming encryption, tamper-evident logging, and tamper detection. The principal contribution is an integrated and testable research architecture, evidenced by a working prototype, that explicitly connects prevention, traceability, detection, and adaptive response, together with an identification of the data and validation requirements needed before such a system can be deployed in a real examination environment.

Keywords: Examination Security; Question-Paper Leakage; Risk-Adaptive Security; Semantic Similarity; Sentence-BERT; Blockchain; Chain of Custody; Digital Watermarking; NLP; Secure Distribution.

INTRODUCTION

Examinations are used to evaluate knowledge, determine academic progression, select candidates for employment, and support admission and scholarship decisions. Because examination outcomes can directly affect educational and professional opportunities, confidentiality and integrity of question papers are fundamental security requirements. A paper leak can provide an unfair advantage, trigger cancellation or postponement, create financial and administrative costs, and reduce public confidence in the examination authority.

The security problem has become more complex as examination content moves through digital systems and is increasingly communicated through online platforms. A compromised paper may appear as an exact document, a screenshot, an image, a partial question, a paraphrase, or OCR-generated text containing recognition errors. It may also



be written in another language or in mixed-language form. Consequently, a detection system based only on exact keywords or string matching can miss semantically equivalent leaked content.

Secure digital distribution addresses one part of the problem. Encryption protects content during storage and transmission, while authentication and authorization restrict who can access it. However, security does not end when a legitimate user opens the paper. Authorities also need an auditable record of access and transfer events so that a suspected leak can be investigated. Copy-level watermarking or fingerprinting can provide additional evidence about the likely source of a leaked copy.

A second challenge is response speed. Traditional investigation often begins after a suspected leak becomes visible. Semantic similarity provides a complementary early-warning mechanism: rather than searching only for exact copies, a system can represent questions as embeddings and compare their meaning with newly observed public content. Sentence-BERT (SBERT) was designed for efficient sentence-level semantic comparison using vector embeddings and cosine similarity, making this approach technically suitable for large-scale similarity search.

The proposed research therefore treats examination security as a layered system rather than a single technology. Secure distribution reduces the opportunity for unauthorized access; chain-of-custody records improve accountability; watermarking supports source tracing; semantic similarity supports early warning; and a risk engine determines when stronger controls should be applied.

1.1 Problem Statement

Existing examination-security approaches often focus on one stage of the security lifecycle - secure sharing, access control, blockchain-based accountability, watermarking, or content similarity. A practical examination environment requires these capabilities to operate together. The central problem addressed by this research is therefore: how a risk-adaptive, multi-layer security framework can protect examination papers during digital distribution, preserve an auditable chain of custody, identify semantically similar leaked content at an early stage, and recommend stronger controls for higher-risk examinations.

1.2 Research Objectives

The specific objectives of this research are to design a secure digital distribution mechanism using encryption, authentication, controlled release, and copy-level fingerprinting; to design a tamper-evident chain-of-custody mechanism for recording examination-paper access and transfer events; to define and evaluate a semantic leak-detection pipeline capable of handling exact, partial, paraphrased, OCR-noisy, spelling-noisy, and multilingual content; to develop an interpretable risk-scoring model that maps examination characteristics to low, medium, high, or critical security levels; to define a human-in-the-loop alert process so that automated similarity detection produces potential-leak alerts rather than automatic accusations; and to conduct a pilot empirical analysis of publicly documented Indian examination-paper-leak incidents and identify the limitations of incident-level prediction from a small dataset.

1.3 Research Questions

The study addresses five research questions.

RQ1: How can examination papers be securely distributed while maintaining controlled access and auditability?

RQ2: How robust is semantic similarity for identifying potential leaks when content is partial, paraphrased, OCR-noisy, or multilingual?

RQ3: Which examination-level characteristics can be used to estimate relative security risk? RQ4: How can risk estimates be translated into differentiated security controls?

RQ5: How can chain-of-custody evidence and human verification improve the investigation of suspected leaks?

1.4 Research Contributions

This research makes the following contributions: a unified architecture connecting secure distribution, traceability, semantic leak detection, and risk-adaptive response; an explicit distinction between incident-confirmation classification and forward-looking examination-risk prediction; a reproducible evaluation protocol for comparing lexical similarity with sentence-embedding similarity under realistic leak conditions; an interpretable risk-control mapping that links examination risk levels to progressively stronger security measures; and a pilot analysis of 66 publicly documented Indian incidents, reported transparently as a small-data classification study rather than overstated as a reliable future-risk predictor.



1.5 Scope of the Study

The study focuses on digital examination-paper distribution and publicly accessible online content relevant to scheduled examinations. It covers encryption, controlled access, audit logging, copy-level fingerprinting, semantic similarity, risk scoring, and human verification. It does not claim access to private messaging groups, private accounts, restricted platforms, law-enforcement systems, or confidential institutional databases. It also does not claim that software alone can prevent every physical or insider-assisted leak.

LITERATURE REVIEW

The literature reviewed for this research indicates that examination security has been addressed through several complementary technologies. The most relevant streams are secure question-paper sharing, blockchain-based accountability, time-controlled access, document watermarking, and semantic similarity. The research opportunity lies in connecting these capabilities and testing them under realistic leak conditions.

2.1 Secure Distribution

Islam, Kader, and Shin [2] proposed BSSQS, a blockchain-based question-sharing scheme that uses two-phase encryption, timestamps, hashing, and smart contracts. The published work describes a time-based lock intended to prevent decryption before the permitted time and reports implementation and security analysis. More recent work has also explored blockchain-based question-paper distribution with role-based permissions and multi-signature authentication; Yadav et al. [8] describe a blockchain-powered question-paper leakage prevention system in which PDF-based distribution is mediated through smart contracts and access controls.

2.2 Blockchain and Traceability

Blockchain can be useful for examination security when the objective is to make security events tamper-evident and independently verifiable. However, storing the complete examination paper on a blockchain is unnecessary and may create confidentiality, scalability, and privacy problems. In the proposed framework, the blockchain or append-only ledger is therefore treated as an integrity layer for event metadata and document hashes rather than as the primary storage location for plaintext examination content. Sattar et al. [6] present a blockchain framework for online examination data management and secure questionnaire delivery, while Abdelsalam et al. [1] study blockchain for improving the reliability of online examination results. Anuradha et al. [10] similarly report on the use of blockchain technology to strengthen examination security, reinforcing its value as one layer within a broader defense-in-depth strategy rather than as a stand-alone solution.

2.3 Time-Controlled Access

Time-lock encryption provides a formal foundation for the idea that information should remain inaccessible until a specified condition or deadline. Liu et al. [4] describe time-lock encryption as a mechanism through which encrypted information becomes decryptable only after a deadline, although practical implementations involve important cryptographic assumptions and design trade-offs. For the proposed examination system, the practical implementation is framed more conservatively as controlled key release and server-side authorization at the scheduled release time, rather than claiming that a theoretical time-lock construction is itself implemented.

2.4 Watermarking

Watermarking can complement access control because a legitimate copy can be assigned a unique fingerprint before distribution. Li et al. [3] proposed an invisible PDF watermarking technique designed to remain robust against print-scan, JPEG-compression, and noise attacks; the reported study demonstrated strong robustness under the tested conditions. Watermarking should nevertheless be treated as a traceability mechanism, not as a replacement for leak detection: a watermark can help answer which authorized copy may have been leaked, while semantic similarity can help answer whether public content resembles the examination content.

2.5 Semantic Similarity

Reimers and Gurevych [5] introduced Sentence-BERT, which generates sentence embeddings that can be compared efficiently using cosine similarity; the approach was developed to make semantic similarity search substantially more efficient than repeated pairwise BERT inference. A 2025 study by Shakya, Shrestha, Subedi, Swarnakar and colleagues [7] investigated semantic similarity analysis for exam questions using sentence-transformer models and reported a Spearman correlation of 0.863 on its benchmark. The study demonstrates the feasibility of SBERT for question similarity but is not equivalent to an examination-leak monitoring experiment.



2.6 Critical Comparison

Study / Approach	Main Technology	Primary Purpose	Strength	Remaining Limitation
Islam et al. (2019) [2]	Blockchain + encryption + time control	Secure question sharing	Controlled release and auditability	Does not address semantic monitoring of public leaks
Liu et al. (2018) [4]	Time-lock encryption	Deadline-based confidentiality	Formal cryptographic foundation	Not a complete examination workflow
Li et al. (2023) [3]	Robust PDF watermarking	Source tracing	Designed for print-scan robustness	Primarily post-leak tracing
Reimers & Gurevych (2019) [5]	SBERT	Semantic similarity	Efficient sentence embeddings	General-purpose NLP, not exam-leak specific
Shakya et al. (2025) [7]	SBERT / vector similarity	Exam-question similarity	Exam-specific similarity evidence	Does not establish end-to-end leak detection
Yadav et al. (2025) [8]	Blockchain + smart contracts	Question-paper leakage prevention	Role-based and auditable distribution	No semantic leak monitoring
Proposed framework	Integrated multi-layer design	Prevention + traceability + detection + adaptive response	Connects security lifecycle stages	Requires full prototype and larger validation

Table 1: Critical Comparison of Related Work

2.7 Research Gap

Existing studies on secure examination systems commonly focus on individual security capabilities rather than examining the complete examination-security lifecycle as an integrated framework [1]-[8], [10]. A significant detection robustness gap also exists, as there is limited examination-specific evidence on the effectiveness of semantic leak detection under challenging conditions such as partial matches, paraphrased content, OCR-induced noise, spelling variations, and multilingual text. Furthermore, secure examination systems rarely provide an interpretable adaptive-response mechanism capable of translating identified examination risks into a dynamically selected set of appropriate security controls. Finally, the combined effectiveness of prevention, traceability, semantic detection, and adaptive response requires an integrated evaluation, as evidence obtained from separate studies cannot automatically be assumed to transfer to a complete examination-security environment.

Accordingly, the novelty claim of this research is deliberately limited. The study does not claim that encryption, blockchain, watermarking, or SBERT is individually new [1]-[8], [10]. Its contribution is the design and evaluation framework that connects these established techniques around the specific operational problem of examination-paper leakage.

METHODOLOGY

The research follows a design-science and applied-experimental methodology. The design-science component specifies the security architecture and its interfaces. The experimental component evaluates the evidence that can be produced from available data. Because the empirical dataset contains only 66 incidents, the machine-learning analysis is explicitly treated as a pilot study. The proposed semantic leak-detection experiment is specified but is not reported as completed unless original experimental results are available.



3.1 Research Design

The study uses three linked activities: (1) system architecture and security-control design, (2) pilot analysis of documented examination-leak incidents, and (3) a reproducible protocol for future semantic leak-detection evaluation. This avoids the methodological problem of describing all four modules as implemented when only the incident dataset was experimentally evaluated.

3.2 Data Sources

Data Source	Purpose	Status in Study
66 documented Indian examination-paper-leak incidents, 2014-2024	Pilot classification analysis	Empirically analyzed
Exam-question similarity pairs	Semantic leak-detection evaluation	Evaluation protocol defined; results not fabricated
Simulated access logs	Chain-of-custody workflow testing	Design specification
Security-control metadata	Risk scoring	Proposed feature set

Table 2: Data Sources and Study Status

3.3 Pilot ML Problem Definition

The available 66-incident dataset does not directly represent future examination risk. Its target variable is whether a reported incident was officially confirmed or remained an accusation. Therefore, the empirical model is defined as an incident-confirmation classification task, with structural features of a reported incident as input, confirmed leak versus unconfirmed accusation as the target, Logistic Regression and Random Forest as models, a majority-class classifier as baseline, five-fold stratified cross-validation, and Accuracy, Precision, Recall, and F1-score as evaluation metrics. This pilot should not be interpreted as a validated forward-looking risk predictor. A true risk-prediction model requires examination-level observations collected before or during an examination and a target representing a future security outcome.

3.4 Proposed Risk-Adaptive Model

The forward-looking component is specified as an examination-level risk score. Candidate features include number of paper handlers, number of distribution points, physical versus digital distribution, previous incidents associated with the institution or region, number of privileged accounts, use of multi-factor authentication, release-window duration, number of copies, and whether individual-copy fingerprinting is enabled. A simple interpretable formulation is given below.

$$R = \sum(w_i \times x_i), \text{ where } 0 \leq x_i \leq 1 \text{ and } \sum w_i = 1 \quad (1)$$

where w_i denotes the weight assigned to risk feature i and x_i denotes the normalized value of that feature. Risk levels are defined as Low (0-30), Medium (31-60), High (61-80), and Critical (81-100). The weights must be learned or justified using a sufficiently large dataset; the numerical thresholds above are design defaults rather than experimentally validated thresholds, and threshold calibration should be performed on held-out validation data during implementation.

3.5 Proposed Semantic Leak-Detection Experiment

The semantic component should be evaluated against a controlled set of question pairs. Each original question should have several variants: exact copy, paraphrase, partial extraction, OCR corruption, spelling noise, translated version, and mixed-language version, with a negative class containing unrelated questions from the same examination domain. Two approaches should be compared: TF-IDF with cosine similarity as a lexical baseline, and SBERT embeddings with cosine similarity as the semantic method, reported using Precision, Recall, F1-score, ROC-AUC where appropriate, false-positive rate, false-negative rate, and inference time. Threshold selection should be performed on a validation split rather than chosen after observing the test results. The evaluation procedure is summarized below as a formal algorithm.

Algorithm 1: Semantic Leak-Detection Procedure

Input: Original question q , candidate content c , validation-selected threshold T

Output: Potential-leak alert with source, timestamp, similarity, and document hash (if triggered)

Step 1: Preprocess q and candidate content c .

Step 2: Generate embedding $E(q)$ and $E(c)$ using SBERT.



Step 3: Compute cosine similarity $S(q,c)$.

Step 4: Compare S with the validation-selected threshold T .

Step 5: If $S \geq T$, create a potential-leak alert.

Step 6: Attach source, timestamp, similarity, and document hash to the alert.

Step 7: Require human verification before escalation.

This protocol is intentionally presented as an evaluation plan because original SBERT experimental results are not yet available; reporting invented scores would weaken rather than strengthen the paper.

3.6 Research Tools and Technologies

Layer	Technology	Purpose	Status
Backend	Java 21, Spring Boot, Spring Security, REST APIs	Authentication, authorization, paper workflow and APIs	Implemented
Database	H2 (embedded, prototype); MySQL/PostgreSQL (deployment profile)	Metadata, users, exam records and audit references	Implemented (H2)
NLP / ML	Python, scikit-learn (TF-IDF), Sentence-Transformers (SBERT, all-MiniLM-L6-v2)	Similarity analysis and risk modelling	Implemented
Security	AES-256-GCM encryption, SHA-256 hashing, key-release controls	Confidentiality and secure communication	Implemented
Traceability	SHA-256 hash-linked append-only ledger	Tamper-evident event history	Implemented
Frontend	React	Authority dashboard and alert review	Implemented
Deployment	Git, Docker	Versioning and reproducible deployment	Configured, not load-tested

Table 3: Research Tools and Technologies

3.7 Ethical and Legal Considerations

Monitoring activities should be restricted to publicly accessible content unless explicit institutional authorization has been obtained. Automated detection mechanisms should generate potential-leak alerts rather than automatically accusing or identifying an individual or organization as responsible. Personal data must be minimized, appropriately access-controlled, and retained only for the period necessary to support legitimate operational requirements. Question-paper content should not be stored directly in public blockchain transactions; instead, cryptographic hashes and relevant event metadata should be recorded to provide integrity and traceability while protecting sensitive examination content. All automated alerts should undergo human review before being escalated to an examination authority or any external organization.

3.8 Threats to Validity

Threat	Effect	Mitigation
Small pilot dataset	Unstable ML estimates and weak generalization	Treat as pilot; collect larger datasets
News/reporting bias	High-profile incidents may be overrepresented	Use multiple source types and documented selection rules
Target mismatch	Confirmation status is not identical to future security risk	Separate pilot classification from risk prediction
Synthetic leak variants	May not capture real-world attacker behaviour	Validate with authorized historical or institutional examples



Threat	Effect	Mitigation
Threshold sensitivity	Similarity alerts can generate false positives/negatives	Tune threshold on validation data
Language variation	Indian multilingual/code-mixed content is diverse	Evaluate multiple languages and code-mixed samples

Table 4: Threats to Validity

PROPOSED SYSTEM DESIGN

The proposed system is organized around a Java-based security core and a Python-based analytical service. The security core controls paper lifecycle events, access rights, release times, and audit records. The analytical service performs semantic similarity and risk calculations. This separation keeps sensitive document-control functions independent from computational NLP components.

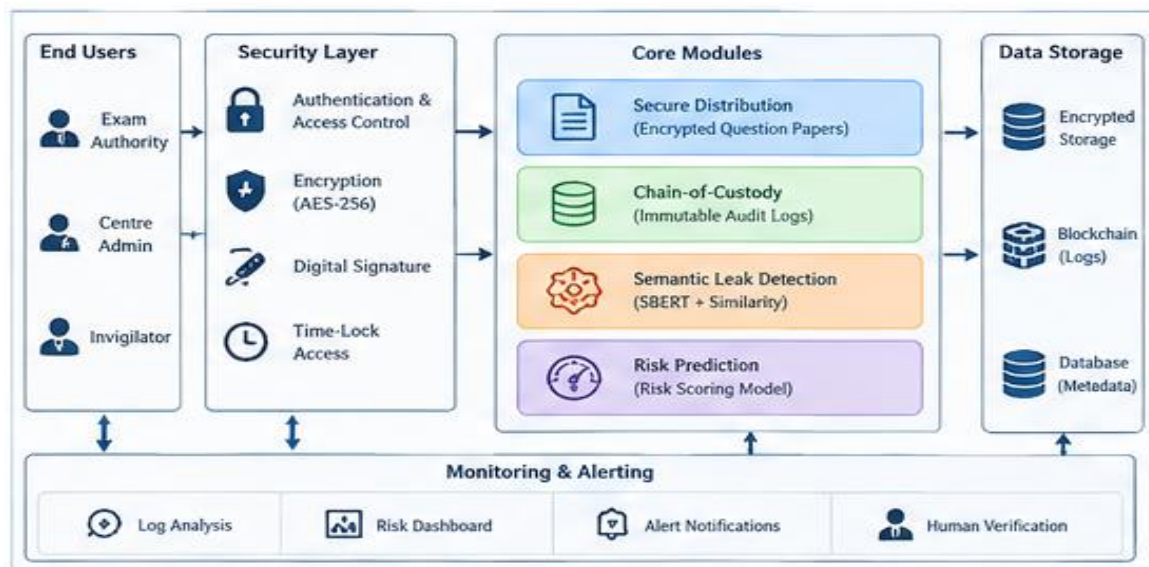


Figure 1: Proposed risk-adaptive examination-security architecture.

4.1 Secure Digital Distribution

The distribution workflow proceeds as follows. The examination authority uploads the paper through an authenticated administrative interface. The system computes a cryptographic hash and encrypts the document before storage or distribution. Each authorized recipient receives a controlled entitlement rather than a reusable shared password. Release is permitted only within the configured examination window, and every access, download, print, or release event is recorded.

4.2 Chain of Custody

The chain-of-custody record is designed as an append-only sequence of security events. A practical event record can contain the fields shown below.

Field	Example
Event ID	EVT-00041
Exam ID	EXAM-2026-001
Document Hash	SHA-256 value
User / Role	Officer-17 / Controller
Action	VIEW / DOWNLOAD / RELEASE
Timestamp	2026-09-09 08:30:00



Field	Example
Copy Fingerprint	COPY-CENTRE-04-USER-17
Previous Event Hash	Hash of prior event
Current Event Hash	Hash of current event

Table 5: Chain-of-Custody Event Fields

A hash-linked event structure makes unauthorized modification easier to detect. If a blockchain or permissioned ledger is used, only non-sensitive metadata and hashes should be recorded; the encrypted paper remains in controlled storage.

4.3 Copy-Level Fingerprinting

Before distribution, each authorized copy can be associated with a unique fingerprint containing the examination identifier, recipient or centre identifier, issuance time, and a random value. The fingerprint may be embedded as an invisible watermark or represented through controlled document metadata. If a leaked copy is recovered, the fingerprint can support source tracing.

4.4 Semantic Leak Detection

The leak-detection service receives candidate public content after appropriate collection and preprocessing. Text is extracted from web pages or OCR-processed images, normalized, and compared against question-level representations. Exact matching can be performed first as a low-cost filter, and SBERT similarity can then be applied to shortlisted candidates. This two-stage design reduces computational cost and makes the alert pipeline easier to audit.

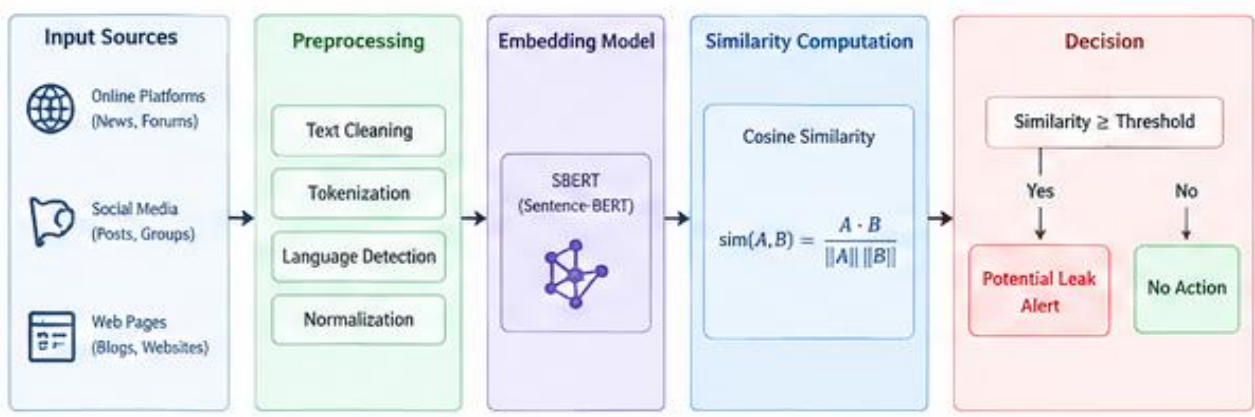


Figure 2: Semantic leak-detection pipeline.

4.5 Risk-Adaptive Decision Engine

The risk engine should not merely display a risk number; it should translate risk into concrete security controls. The following matrix is a design proposal and must be calibrated during future evaluation.

Risk Level	Indicative Controls
Low	Standard encryption, authenticated access, normal audit logging
Medium	MFA, shorter release window, enhanced audit review, copy fingerprinting
High	MFA + restricted devices, dual authorization, stronger monitoring, frequent audit checks
Critical	Dual-control release, isolated environment, short release window, enhanced monitoring, immediate human review of alerts

Table 6: Risk Level and Indicative Security Controls



4.6 Human Verification

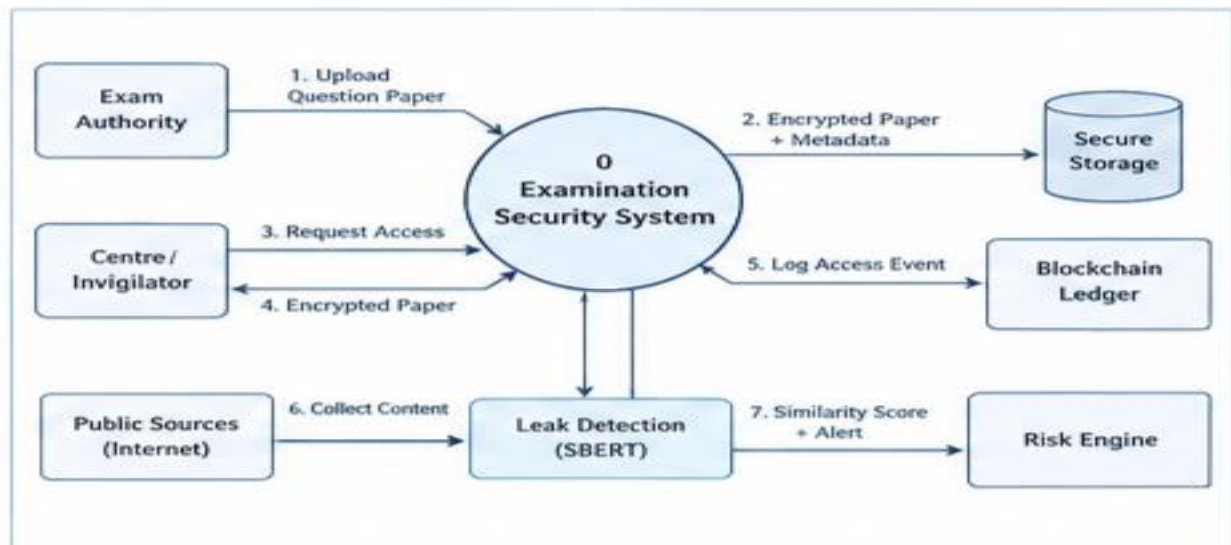


Figure 3: End-to-end security and response workflow.

A semantic similarity score is evidence for investigation, not proof of a leak. Human verification should consider the question content, source credibility, publication time, watermark evidence, chain-of-custody records, and other contextual information before an incident is classified.

4.7 Prototype Implementation and Verification

A working prototype of the architecture in Figure 1 was implemented using the stack summarized in Table 3. The security core was built in Java 21 with Spring Boot and Spring Security, exposing the paper-release, access, ledger, and risk-scoring functions as REST endpoints and enforcing role-based access control across three roles (ADMIN, CONTROLLER, INVIGILATOR) so that only authorized roles can release a paper. Secure distribution was implemented with AES-256-GCM authenticated encryption and a SHA-256 content hash computed at release time. The chain of custody was implemented as an append-only, hash-linked event log, matching the event schema in Table 5, and persisted through Spring Data JPA to an embedded H2 database for local development, with a MySQL profile available for deployment. Semantic leak detection was implemented as a separate Python microservice that computes cosine similarity between an original question and candidate content, using Sentence-BERT (the all-MiniLM-L6-v2 model) when it can be loaded and automatically falling back to a TF-IDF baseline otherwise; the service reports which method actually produced each score rather than allowing a fallback to be mistaken for the full model. A React frontend was built against the REST API to exercise all three modules interactively.

Verification proceeded in two layers. First, the core security logic - encryption and decryption, hash-chain construction, and the risk-scoring formula in Equation (1) - was extracted into framework-independent Java classes and exercised with fourteen automated checks (AES-GCM round-trip recovery, rejection of a wrong decryption key, hash determinism, correct detection and localization of a tampered ledger event, and correctly ordered risk scores across a low-risk and a high-risk configuration), all of which passed. The Python similarity service was exercised with automated tests confirming that a paraphrased question pair was correctly flagged above threshold and an unrelated pair was correctly not flagged. Second, the complete system was run end-to-end on a local machine: a paper was encrypted and released, subsequent VIEW and DOWNLOAD events were logged and hash-linked, chain verification confirmed the ledger was intact, and deliberately altering a stored event's detail field was correctly detected and localized by the verification endpoint. Both the TF-IDF and the Sentence-BERT similarity paths were exercised through the running service and returned distinguishable results for similar versus unrelated text.

This end-to-end run demonstrates that the architecture's core mechanisms operate correctly together; it is not the systematic SBERT benchmark specified in Section 3.5, which still requires the full exact/partial/paraphrased/OCR-noisy/multilingual variant set and remains future work. The prototype was also run on a single local machine without concurrent users, a production key-management service, or load testing, and the demonstration credentials are held in memory rather than issued through an institutional identity provider. These gaps are the natural next steps identified in Section VIII rather than limitations that were overlooked.



PILOT EXPERIMENTAL RESULTS

This section reports only the experiment supported by the available data: classification of documented incidents. The results are not relabeled as evidence of a validated forward-looking risk predictor. The proposed semantic leak-detection experiment remains an evaluation protocol until original results are produced.

5.1 Dataset

Attribute	Description
Dataset name	India Exam Paper Leak Analysis
Period	2014-2024
Total records	66
Confirmed leaks	48
Unconfirmed	18
Source	Publicly documented reports
Target variable	Confirmation status

Table 7: Dataset Description

The public repository describes 66 recorded Indian examination-paper-leak incidents from 2014-2024 and provides a data-analysis project built around the dataset.

5.2 Evaluation Setup

Logistic Regression and Random Forest classifiers were evaluated using five-fold stratified cross-validation. A majority-class baseline was included because the confirmed class represents 72.7% of the records. The baseline is essential: without it, a trained model could appear useful even when it performs no better than simply predicting the dominant class.

5.3 Results

Model	Accuracy	Precision	Recall	F1-Score
Majority Baseline	0.727	0.727	1.000	0.842
Logistic Regression	0.590	0.818	0.589	0.674
Random Forest	0.652	0.808	0.691	0.739

Table 8: Pilot Classification Results (5-fold Stratified Cross-Validation)

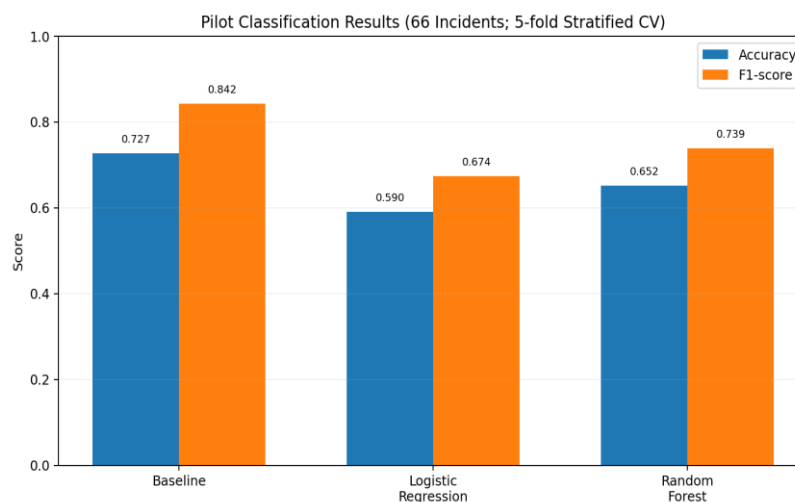


Figure 4: Accuracy and F1-score comparison for the 66-incident pilot study.



5.4 Interpretation

Neither trained classifier outperformed the majority-class baseline on Accuracy or F1-score. Logistic Regression achieved 0.818 Precision but only 0.589 Recall, while Random Forest achieved 0.808 Precision and 0.691 Recall. The baseline achieved the highest F1-score of 0.842 because the dataset is small and imbalanced toward confirmed incidents. The appropriate conclusion is therefore not that machine learning is ineffective for examination security; the supported conclusion is narrower: the available 66-record dataset and the selected structural features are insufficient to reliably classify confirmation status. The experiment provides evidence for the need to collect richer, examination-level data before attempting a production risk-prediction model.

5.5 Evidence Matrix

Framework Component	Evidence in This Study	Evidence Status
Secure distribution	AES-256-GCM encryption implemented and unit-tested (round-trip and wrong-key rejection)	Prototype implemented and tested
Chain of custody	SHA-256 hash-linked ledger implemented; tamper detection confirmed end-to-end	Prototype implemented and tested
Watermarking	Literature-supported source-tracing mechanism	Literature evidence; not implemented
Semantic leak detection	TF-IDF and SBERT (all-MiniLM-L6-v2) both implemented and exercised end-to-end	Prototype implemented and tested; systematic benchmark (Section 3.5) still pending
Risk prediction (Equation 1)	Risk engine implemented in Java and unit-tested; 66-incident pilot classification	Engine implemented and tested; incident-level prediction remains a pilot
Integrated framework	Java/Spring Boot core, Python NLP microservice, and React frontend run together end-to-end on a local machine	Prototype integrated; not load-tested or deployed

Table 9: Evidence Matrix

DISCUSSION

The pilot result changes how the framework should be interpreted. The risk engine cannot currently be presented as a validated predictor of future examination leaks because the available dataset describes reported incidents rather than pre-examination security conditions. This limitation is important and has been made explicit in the research design.

The stronger contribution of the proposed framework is therefore architectural and methodological. It defines how multiple security layers can exchange evidence: an access event can be linked to a document hash, a document can be linked to a copy fingerprint, a public candidate can be linked to a similarity score, and a potential-leak alert can be linked back to the relevant examination and access history. This creates a traceable investigation path rather than a collection of unrelated security tools.

The semantic component is particularly important because examination leaks are not necessarily exact copies. A detection system that only compares strings may miss paraphrases and OCR variants. SBERT provides an established mechanism for converting sentences into embeddings suitable for cosine-based semantic comparison, but examination-specific thresholds and multilingual evaluation remain necessary before operational deployment.

Blockchain should also be used selectively. It is valuable for tamper-evident event records and shared verification, but it should not be treated as a universal security solution. Encryption, key management, authentication, endpoint security, and operational procedures remain necessary even when a ledger is used.

LIMITATIONS

This study has several limitations. The pilot dataset contains only 66 documented incidents and therefore provides limited statistical power, and the target of the pilot model is incident-confirmation status rather than future examination-leakage risk. The prototype described in Section 4.7 implements and unit-tests the core logic of all four modules, but it has not been load-tested with concurrent users, penetration-tested, or evaluated against the systematic SBERT benchmark.



specified in Section 3.5; its demonstration credentials and key handling are not production-grade. Public-source data can contain selection and reporting bias, and simulated or constructed leak variants may not represent every real-world attack pattern. Multilingual and code-mixed examination content requires broader language-specific evaluation, and operational deployment would require institutional authorization, security testing, privacy review, and integration with existing examination procedures.

FUTURE WORK

Future work will proceed along several directions. Building on the working prototype described in Section 4.7, an authorized examination-question benchmark will be constructed with exact, partial, paraphrased, OCR-noisy, spelling-noisy, multilingual, and code-mixed variants, and TF-IDF, SBERT, and potentially multilingual sentence-embedding models will be compared systematically using held-out evaluation data - extending the illustrative end-to-end run already performed into the full benchmark specified in Section 3.5. Examination-level security features and incident outcomes will be collected to build a genuine forward-looking risk model, and risk thresholds will be calibrated using validation data rather than arbitrary fixed thresholds. False-positive and false-negative costs will be evaluated, since a false alarm and a missed leak have different operational consequences. The prototype's demo authentication will be replaced with a production key-management service and an institution-approved identity provider, and the system will be load-tested with concurrent users. Penetration testing and threat modelling of authentication, key management, document storage, APIs, and audit records will be conducted, and watermark recovery will be evaluated after printing, scanning, compression, cropping, and image capture. Finally, a controlled pilot will be conducted with an examination authority before any real operational deployment.

CONCLUSION

This research presents a structured approach to examination-paper security based on four complementary capabilities: secure digital distribution, chain-of-custody traceability, semantic leak detection, and risk-adaptive security controls. The central design principle is that no single technology is sufficient: encryption protects confidentiality, audit mechanisms improve accountability, watermarking supports source tracing, semantic similarity supports early warning, and risk scoring provides a mechanism for selecting stronger controls when an examination requires them.

The empirical analysis is deliberately conservative. The available 66-incident dataset supports a pilot classification experiment, not a validated future-risk predictor. Logistic Regression and Random Forest did not outperform the majority-class baseline on Accuracy or F1-score, demonstrating that richer data and a better-defined target are required. This negative result is useful because it prevents overclaiming and identifies a concrete direction for future research.

The principal research contribution is therefore a testable, risk-adaptive architecture, evidenced by a working prototype covering all four modules and verified through automated unit tests and an end-to-end run, together with a clear evaluation methodology. Future implementation and controlled experimentation - especially a systematic SBERT benchmark under partial, paraphrased, OCR-noisy, and multilingual conditions, and multi-user load testing of the integrated system - can determine whether the proposed integration delivers measurable benefits beyond the capabilities of individual security technologies.

REFERENCES

- [1]. Abdelsalam, M., Shokry, M., & Idrees, A. M. (2024). A proposed model for improving the reliability of online exam results using blockchain. *IEEE Access*, 12, 7719-7733.
- [2]. Islam, A., Kader, M. F., & Shin, S. Y. (2019). BSSSQS: A blockchain-based smart and secured scheme for question sharing in the smart education system. *Journal of Information and Communication Convergence Engineering*, 17(3), 174-184.
- [3]. Li, L., Zhang, H.-J., Meng, J.-L., & Lu, Z.-M. (2023). Robust PDF watermarking against print-scan attack. *Sensors*, 23(17), 7365.
- [4]. Liu, J., Jager, T., Kakvi, S. A., & Warinschi, B. (2018). How to build time-lock encryption. *Designs, Codes and Cryptography*, 86, 2549-2586.
- [5]. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3982-3992.



- [6]. Sattar, M. R. I., Efty, M. T. B. H., Rafa, T. S., Das, T., Samad, M. S., Pathak, A., Khandaker, M. U., & Ullah, M. H. (2023). An advanced and secure framework for conducting online examination using blockchain method. *Cyber Security and Applications*, 1, 100005.
- [7]. Shakya, N., Shrestha, M., Subedi, R., Swarnakar, N., et al. (2025). Semantic similarity analysis for exam questions using Sentence Transformer Model. *International Journal on Engineering Technology*, 2(2), 98-103.
- [8]. Yadav, H. K. Y. G., Adithya, K., Ehshanulla, T., Mokshith Srinivas Yadav, M., & Gowrishankar, Y. (2025). Blockchain-powered question paper leakage prevention system. *SCITEPRESS, ICRDICCT 2025*.
- [9]. Choudhary, K. (2025). India Exam Paper Leak Analysis: Exam paper leak incidents in India, 2014-2024 [Data set]. *GitHub*.
- [10]. Anuradha, K., Jahnvi, K., Jahanavi, P., Bhargavi, M., et al. (2025). Enhancing exam security with blockchain technology. *International Journal for Research in Applied Science and Engineering Technology*, 13(3), 1967-1971.